HHS Public Access

Author manuscript

Epigenomics. Author manuscript; available in PMC 2015 April 22.

Published in final edited form as:

Epigenomics. 2014 ; 6(5): 455–457. doi:10.2217/epi.14.47.

Interpreting 4C-Seq data: how far can we go?

Ramya Raviram[1], **Pedro P. Rocha**[1], **Richard Bonneau**[2,3,4], and **Jane A. Skok**[1,*]

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

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

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

[4]Simons Foundation, New York, New York, 10010, USA.

The linear sequence of the genome has been extremely valuable in mapping regulatory elements relative to the genes they control. However, it has become increasingly evident that characterizing the three-dimensional organization of the genome is critical to get a better understanding of long-range regulation. Early studies using fluorescent in-situ hybridization (FISH) revealed that individual chromosomes occupy distinct spaces in the nucleus with minimal intermingling between territories[1]. Recent advances using chromosome conformation capture (3C) techniques have confirmed these findings and further improved the depth at which we can determine the organization of chromosomes and the physical interactions that occur within and between them[2, 3]. Variations of the 3C technique include (i) Hi-C, to capture all pairwise interactions, (ii) 5C, to capture interactions within and between loci of interest and (iii) 4C-Seq, to capture all interactions with a single locus of interest. The choice of technique depends on the biological question being asked and the scale at which this needs to be examined. While Hi-C has been instrumental in characterizing higher-order organization of chromosomes in the nucleus, it lacks the resolution that is required for analysis of specific interactions, such as between enhancers and promoters. This can be achieved with 4C-Seq, which allows interrogation of interactions from a single viewpoint or bait, to the rest of the genome. Several studies have used 4C-Seq to better understand phenomena such as X chromosome inactivation[4], enhancer-promoter interactions[5, 6], organization of antigen receptor loci[7], choice of translocation partners[8, 9] and collinear transcriptional regulation[10]. Here we aim to focus on the current state of the 4C-Seq method and the limitations and challenges of the associated computational analysis.

Analysis of 4C-Seq data can be complicated by several technical biases intrinsic to the method. The first bias to consider is that the majority of 4C-seq signal is found on the bait chromosome with lower coverage in *trans*. This bias does not represent noise but is in agreement with the chromosome territory model, which would predict fewer inter-chromosomal interactions than intra-chromosomal interactions. Second, there is a decrease in signal along the *cis* chromosome as a function of distance from the bait. Third, similar to

*To whom correspondence should be addressed: Tel: 212-263-0504, jane.skok@med.nyu.edu.

other 3C-based techniques, 4C-Seq relies on using restriction enzymes to digest the chromatin, and the frequency of sites in the genome which the enzyme recognizes determines the resolution of the assay. Finally, bias arises from the inverse-PCR amplification step that is required for identification of interacting regions. This can lead to an artificial overrepresentation of regions that amplify with greater efficiency. Thus, when developing methods for analyses of 4C-Seq data it is important to take these issues into account. Current methods provide tools to map 4C-Seq reads, normalize data, identify regions of significant interactions and visualize signal across the genome. These tools have provided a good starting point to characterize chromosomal interactions, however there needs to be improvement in incorporating all of the above inherent biases of 4C.

PCR artifacts or identical reads are typically discarded in other genome wide techniques, however in 4C-Seq a distinction cannot be made between repeated amplification of a single captured interaction versus amplification of multiple interactions. To deal with this issue some of the available pipelines transform the data to a binary signal (a score of zero or one) based on the presence or absence of a read at each restriction enzyme fragment[6, 11, 12]. To investigate the extent of the PCR amplification bias a recent study incorporated random barcodes in their experimental strategy and found no selective bias in their usage[13]. This demonstrates that binary transformation removes information that can be obtained from the number of captured interactions at a given restriction enzyme fragment. Hence, it is important to develop methods of analysis that integrate the number of fragments that have interactions with the total number of reads at each of these locations.

As mentioned above, the resolution of 4C-Seq is determined by the frequency of restriction enzyme sites in the genome. The resolution achieved from a six bp cutter enzyme can be around 3-4Kb and this can be increased to 200-400 bp using a four bp cutter. Six bp cutters have been used to examine interactions[8, 14] in *cis* and in *trans*, while 4bp cutters have only successfully characterized interactions in the region surrounding the bait[5, 6, 10]. The limitations of 4bp cutter derived analysis arise due to low reproducibility of 4C signal between replicates in far-*cis* and *trans*. It is thus clear that the choice of restriction enzyme used in the experimental design can directly limit the type of interactions that can be assayed.

Due to the polymer nature of chromatin, an interaction between two loci can occur over a region of chromatin in a population of cells. To account for this, a genomic window is typically used to analyze interactions as opposed to dealing with individual fragments.[6, 11, 13, 15]. However, it should be noted that the use of arbitrary window sizes may obscure the boundary of interacting domains that can be detected so it is important to first determine the appropriate size. To accomplish this the first step is establishing the resolution at which interactions are reproducible between replicates. This will vary depending on whether regions are (i) proximal to the bait, (ii) in far-*cis*, or (iii) in *trans* (all regions on *trans* chromosomes can be treated the same way since there is an equal probability of their interacting with the bait). For example, when considering regions that are close to the bait, the window size can be smaller because of increased coverage and reproducibility in this location. Thus, the resolution is highest in regions proximal to the bait. Since coverage and

reproducibility depend on linear and spatial separation from the bait, use of the same analytical approach cannot be uniformly applied across the genome.

Most 4C based studies look at interactions that occur within ~500kb of the viewpoint because this is where the signal is highest and most reproducible. The combined use of genetics and DNA FISH based approaches have nonetheless revealed that regulatory elements can act over a linear distance of more than 50Mb[16] on the same chromosome and between genes on different chromosomes[17] by being brought into close contact at high frequency in the nucleus. Although DNA FISH is the gold standard when it comes to measuring these types of interactions, it has a low resolution and there is a limit to the number of loci that can be simultaneously analyzed. Furthermore, FISH cannot be used to identify associations in an unbiased manner in the way that 4C-Seq can. But is it possible to reliably identify this type of longer-range interaction by 4C?

A number of labs have identified regions with significant interactions in far-*cis* and *trans* and ranked the intensity of signal within a sample[8, 12]. In contrast, few studies have quantitatively examined differences in signal of longer-range interactions between different conditions[14]. Because 4C-Seq pipelines used for this purpose have not extensively analyzed reproducibility it is not clear whether it is feasible to reliably perform quantitative analyses of longer-range interactions. Obviously, for analysis of longer-range interactions it is of paramount importance to assess reproducibility as well as to validate associations with DNA FISH using the appropriate controls. Importantly, genetic approaches are additionally required to determine functional relevance. Furthermore, to improve our understanding of the role of chromosomal interactions in regulatory processes it is essential to develop methods that integrate 4C-Seq with other genome-wide data sets. Clearly the field is developing rapidly but it is important that new and improved tools of analysis are developed in tandem.

References

1. Cremer T, Cremer C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nature reviews. Genetics. 2001; 2(4):292–301.
2. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science. 2002; 295(5558):1306–1311. [PubMed: 11847345]
3. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature reviews. Genetics. 2013; 14(6):390–403.
4. Splinter E, De Wit E, Nora EP, et al. The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. Genes & development. 2011; 25(13):1371–1383. [PubMed: 21690198]
5. Ghavi-Helm Y, Klein FA, Pakozdi T, et al. Enhancer loops appear stable during development and are associated with paused polymerase. Nature. 2014
6. Van De Werken HJ, Landan G, Holwerda SJ, et al. Robust 4C-seq data analysis to screen for regulatory DNA interactions. Nat Methods. 2012; 9(10):969–972. [PubMed: 22961246]
7. Medvedovic J, Ebert A, Tagoh H, et al. Flexible Long-Range Loops in the VH Gene Region of the Igh Locus Facilitate the Generation of a Diverse Antibody Repertoire. Immunity. 2013; 39(2):229–244. [PubMed: 23973221]
8. Rocha PP, Micsinai M, Kim JR, et al. Close proximity to Igh is a contributing factor to AID-mediated translocations. Molecular cell. 2012; 47(6):873–885. [PubMed: 22864115]

9. Simonis M, Klous P, Homminga I, et al. High-resolution identification of balanced and complex chromosomal rearrangements by 4C technology. Nature methods. 2009; 6(11):837–842. [PubMed: 19820713]
10. Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D. The dynamic architecture of Hox gene clusters. Science. 2011; 334(6053):222–225. [PubMed: 21998387]
11. Van De Werken HJ, De Vree PJ, Splinter E, et al. 4C technology: protocols and data analysis. Methods in enzymology. 2012; 513:89–112. [PubMed: 22929766]
12. Denholtz M, Bonora G, Chronis C, 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):602–616. [PubMed: 24035354]
13. Williams RL Jr. Starmer J, Mugford JW, et al. fourSig: a method for determining chromosomal interactions in 4C-Seq data. Nucleic acids research. 2014; 42(8):e68. [PubMed: 24561615]
14. De Wit E, Bouwman BA, Zhu Y, et al. The pluripotent genome in three dimensions is shaped around pluripotency factors. Nature. 2013; 501(7466):227–231. [PubMed: 23883933]
15. 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 research. 2013; 41(13):e132. [PubMed: 23671339]
16. Collins A, Hewitt SL, Chaumeil J, et al. RUNX transcription factor-mediated association of Cd4 and Cd8 enables coordinate gene regulation. Immunity. 2011; 34(3):303–314. [PubMed: 21435585]
17. Hewitt SL, Farmer D, Marszalek K, 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. Nature immunology. 2008; 9(4):396–404. [PubMed: 18297074]