Nucleus 4:5, 390–398; September/October 2013; © 2013 Landes Bioscience

The statistical-mechanics of chromosome conformation capture

Justin M O'Sullivan[1,*], Michael D Hendy[2], Tatyana Pichugina[1], Graeme C Wake[3], and Jörg Langowski[4]

[1]Liggins Institute; University of Auckland; Auckland, New Zealand; [2]Mathematics and Statistics; University of Otago; Dunedin, New Zealand; [3]Institute of Natural and Mathematical Sciences; Massey University; Auckland, New Zealand; [4]Deutsches Krebsforschungszentrum; Biophysics of Macromolecules; Heidelberg, Germany

Keywords: chromosome conformation capture, chromosome organization, proximity ligation, genome organization

Abbreviations: 3C, chromosome conformation capture; 3D, three dimensional

Since Jacob and Monod's characterization of the role of DNA elements in gene control, it has been recognized that the linear organization of genome structure is important for the regulation of gene transcription and hence the manifestation of phenotypes. Similarly, it has long been hypothesized that the spatial organization (in three dimensions evolving through time), as part of the epigenome, makes a significant contribution to the genotype-phenotype transition. Proximity ligation assays commonly known as chromosome conformation capture (3C) and 3C based methodologies (e.g., GCC, HiC, and ChIA-Pet) are increasingly being incorporated into empirical studies to investigate the role that three-dimensional genome structure plays in the regulation of phenotype. The apparent simplicity of these methodologies—crosslink chromatin, digest, dilute, ligate, detect interactions—belies the complexity of the data and the considerations that should be taken into account to ensure the generation and accurate interpretation of reliable data. Here we discuss the probabilistic nature of these methodologies and how this contributes to their endogenous limitations.

Introduction

The fuzzy genome: From populations to single cells

The spatial organization of genomes has been investigated using different methodologies since studies into the cell's components began. Techniques using low (e.g., refs. 1 and 2) and high through-put microscopy (e.g., ref. 3) and proximity-based ligation have enabled the generation of large data sets that facilitate the investigation of specific and global chromosome interaction patterns from an ever increasing number of organisms that includes bacteria,[4,5] baker's yeast,[6,7] fission yeast,[8] *Drosophila*,[9] mouse[10] and human.[11] Collectively these studies confirm that the positioning of loci varies within and between cells, even in identical conditions,[11,12] reviewed in references 13 and 14.

Proximity ligation assays are probabilistic in nature and as such rely upon population analyses. The probabilistic nature of these assays is due to both biological and technical aspects of the methods. Biological variation arises from the facts that: (1) DNA is highly flexible with a persistence length of 50 nm (150 base pairs). Therefore, a chromosome or DNA segment of tens or hundreds of megabases may assume many different folding conformations with equal probability and (2) despite the (unintentionally misleading) tendency to present proximity ligation data as unique conformations or summary heat maps, there is no a priori biological, evolutionary or physical reason to assume a unique genome conformation in different individuals, or even within a population of cells in the same environmental condition(s). Technical variation also contributes to the probabilistic nature of the data, in this case variation arises from the cross-linking, digestion and ligation reactions which are inherently probabilistic. Thus, identifying an interaction does not mean the interaction is always present—rather that it is present within an undetermined proportion of the population. Conversely, the failure to detect an interaction does not mean that it never occurs, rather that the method does not detect it under those experimental conditions because the sites may not be cut or ligated.

Results from proximity ligation based experiments confirm contacts between genomic loci separated by large genomic distances (or between different chromosomes).[4-6,8-11,15-20] However, the existence of interactions does not automatically equate to biological significance, and interactions may simply originate from the very act of packing chromosomes into the nucleus or cell. Of course, this argument ignores the possibility that the restraints imposed by nuclear, or cell, shape have been integrated into chromosome organization over the course of evolution. Despite this, it is clear that some chromosomal interactions observed by modern techniques represent specific biologically relevant linkages[4,5,15,21,22].

Could there be a single spatial solution to a genome's structure?

What is the average structure of a sportsperson playing a game? For the purpose of this analogy, any game will do; however we will specify that the game is rugby union. Over the course of a game, any one player's structure changes as the game ebbs and flows toward its ultimate conclusion (**Fig. 1A–E**). The changes that occur to the player's structure are not to the extent that arms or legs are removed—albeit sometimes they are broken—but rather the spatial arrangement differs from minute to minute

*Correspondence to: Justin M O'Sullivan;
Email: justin.osullivan@auckland.ac.nz
Submitted: 07/23/2013; Revised: 09/12/2013; Accepted: 09/17/2013
http://dx.doi.org/10.4161/nucl.26513

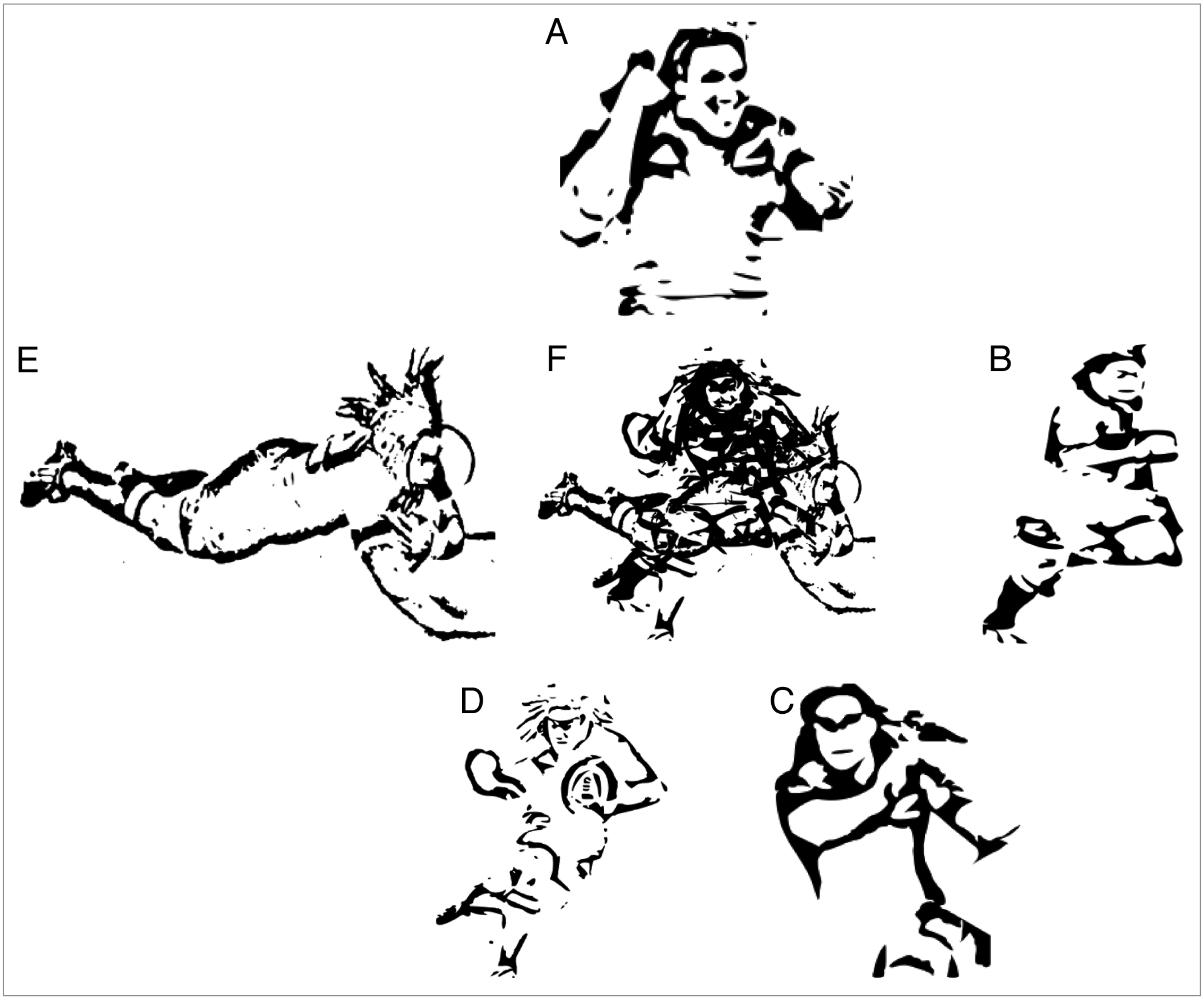

Figure 1. Proximity ligation assays are probabilistic and determine the average structure within a population of cells. (**A–E**) Schematics of the individual conformations that Ma'a Nonu (All Black center) assumes during a game of rugby. Examples of the conformations Ma'a Nonu assumes are illustrated for: (**A**) the traditional All Black haka which occurs at the beginning of the game; (**B**) the act of passing the ball to his left.; (**C**) running; (**D**) a side-step and fend to avoid an opponent who is attempting to tackle him from the right; and (**E**) the act of scoring a try by diving to place the ball on the ground over the try line. (**F**) The average structural conformation obtained from (**A–E**).

according to his or her role within the current play. For example at the moment captured in **Figure 1A** the player's (i.e., Ma'a Nonu) right hand is positioned next to his head as he participates in a ritualized challenge (e.g., haka; http://www.allblacks.com/index.cfm?layout=haka) performed at the beginning of the game. By contrast, at some point later in the game, the player's right hand can be spatially adjacent to the left shoulder (**Fig. 1B**), in front of the right shoulder (**Fig. 1C and D**) or located at some distance (determined by the length of the arm) below the right shoulder (**Fig. 1E**). Could we determine the player's structure by taking low resolution photographs of him or her at multiple time points across the game? Theoretically yes, the body parts would all be in similar relative positions (e.g., the head connected to the neck, arms to shoulders, etc.); however, even allowing for rotations that enable gross characteristics (e.g., the torso) to be aligned the average structure would be a dis-ordered amalgamation, the uniformity of which depends on the number of images which were taken and the body parts that were resolvable in the images.

This analogy represents the situation that would be predicted to occur for maps of genome organization that represent the different cell cycle stages (i.e., are unsynchronized) and even within cells that are synchronized to exactly the same moment of the cell cycle. Moreover, if we extrapolate the analogy to include the other players involved in the game, it demonstrates that averaging results across different cell types is also problematic.[6,16]

Variation is an inherent property of polymer structure

Precise positioning of entire chromosomes, which are in effect polymers consisting of several millions of monomers, would

Table 1. Loci that interact within a genome tend to do so with more than one locus. No. of interactions per fragment, the number of interactions that were empirically determined for fragments that interact with at least one other fragment within the data sets

			No. of interactions per fragment		
Organism	**Condition/cell line**	**Method**	**Mean**	**Max**	**Reference**
Saccharomyces cerevisiae	Glucose	GCC	11[a]	30	17 and 21
	Glycerol lactate	GCC	6[a]	20	17
	Galactose	GCC	3[a]	21	17
Escherichia coli	exponential	GCC	7.83	38	5
	serine hydroxamate	GCC	6.67	25	5
Homo sapiens	GM12878	5C	3.88[b] 3.25[c]	20	15
	K562	5C	4.07[b] 3.84[c]	> 9	15
	Hela-S3	5C	5.35[b] 4.62[c]	> 9	15

[a]Mode number of interacting fragments for restriction fragments that interacted with more than just adjacent fragments within the *S. cerevisiae* genome. [b]Mean number of interacting fragments for transcription start sites (TSS), throughout the ENCODE pilot regions representing 1% of the human genome, with at least one non-adjacent interaction for expressed genes in GM12878. [c]Mean number of interacting fragments for TSS sites for non-expressed genes in GM12878. GCC, genome conformation capture[6]; 5C, 3C carbon-copy.[72]

have too high an entropic cost to make it a feasible mechanism for controlling genomic processes in cells.[23,24] In fact, it would imply holding the interphase chromatin together by some kind of rigid scaffold at support points not farther away from each other than the persistence length of chromatin—which has been estimated between 20 and 200 nm,[25] a sizeable range but clearly much smaller than the extent of the typical eukaryotic nucleus. However, it is clear that the directed positioning of specific chromosomal regions does occur to different degrees during the cell cycle and in different cell types.[26-32]

Current experimental methods (reviewed in refs. 33–35) are limited and are unable to empirically define the precise positions of all loci within the genome of a single cell at high resolution. Therefore, to define features within the chromosome conformation ensemble, we are limited to modeling the ensembles of chromosome conformations using approaches developed for polymer physics coupled with data obtained from proximity ligation.[11,16,21,22,27,36] Despite the fact that there is inter-conformation variability between the structures in the ensembles, recent results have identified conserved patterns and clusters within the genomes that have been modeled to date.[4,21,22]

Analyses of high resolution 3D structures of the *S. cerevisiae* genome that incorporated a proximity-ligation data set confirmed that yeast have preferred positions in the nucleus.[21] Moreover, interaction-dependent clustering of tRNAs, early firing origins of replication and Gal4 upstream activating sequences were identified. Similarly, Ben-Elazar et al.[37] reported the co-localization of co-regulated genes within space following analysis of statistically generated models that incorporated a global analysis of the *S. cerevisise* genome.[7] These findings support the hypothesis that genome structure and function are interlinked.

Simulated three-dimensional (3D) structures of the *Caulobacter crescentus* genome, generated using 5C data and live-cell imaging, have illustrated that the *C. crescentus* genome is ellipsoidal with periodically arranged arms.[4] Further clustering analysis of the individual structures within this ensemble identified four structurally similar configurations.

While the application of high-throughput sequencing to proximity-ligation data has led to significant increases in data generation and a leap in our understanding of genome organization, the reconstruction of 3D genome conformations based on such data is in its infancy. Appropriate methods for analysis are still being developed. The main problems in the linking of data to models of genome organization are that: (1) the ligation probabilities are not exclusively a function of the proximity, but influenced by many other factors such as intrinsic reactivity, local coverage by proteins, crowding, etc.; (2) these methodologies have the intrinsic property that they only provide information on pairwise interactions—simultaneous interactions of several loci in a cluster can only be inferred, not be directly identified; (3) restriction enzyme choice affects the frequency of detection of restriction fragments and ligation products; (4) PCR biases affect the chances of detecting and identifying interactions; and (5) the copy number of genomic regions affects the chances of detecting interactions with and between these loci. Despite these methodological limitations and the structural variability that exists within conformation ensembles, the results of 3D modeling have identified conserved patterns and clusters within those genomes that have been analyzed thus far. These models have the potential to generate new hypotheses about the relationship between nuclear structure and function.

Proximity ligation assays[5,15,17,21] have demonstrated that, in addition to the directly adjacent neighbors, loci interact with up to 38 partners depending on the organism or data set under study (**Table 1**). The variation in observed numbers of interacting partners for loci within an organism reflect the: (1) physical characteristics of the chromosomes; (2) depth of interaction coverage: e.g., the depth of sequencing, the choice of restriction enzyme and resulting fragment lengths affect if an interaction is

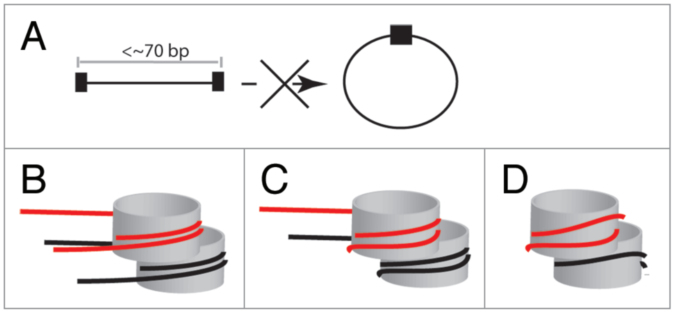

Figure 2. Restriction fragments lengths and protein binding can limit the ability of free ends present on cross-linked fragments to ligate. (**A**) The circularization of linear DNA fragments is length dependent with fragments below ~70 bp classically regarded as not being able to circularize. (**B–D**) Cartoons illustrate how nucleosome (gray disc) binding at or near the end of restriction fragments can prevent the ligation of free ends and limit the potential ligation products from two restriction fragments (red and black lines).

Table 2. The number of ligation configurations that are possible when 2–10 ligation competent fragments are held within one cross-linked complex

Number of interacting fragments	Number of different ligation configurations
2	10
3	76
4	764
5	9496
6	140 152
7	2 390 480
8	46 206 736
9	997 313 824
10	23 758 664 096

detected and hence the coverage; (3) relative complexity of the functional restraints that result in the formation of an interaction; and (4) the environment within which the genome was analyzed. From a functional perspective, it is tempting to speculate that an association with large numbers of partners indicates that loci are clustered at hubs[38,39] for repair,[40] replication or transcription.[41-43] However, the validity of this interpretation is debatable as it ignores biological issues including the fact that cell volumes, and hence global chromatin compaction levels, vary and this variation can alter our ability to detect interactions by unblocking restriction sites. Moreover, the interpretation also ignores the technical issue that proximity ligation assays determine the population average and do not identify the combination of interactions that occur within a single cell.

At present, most of the existing proximity-ligation data are a population average of both large numbers of cells and and cell cycle stages. It could be argued that proximity ligation assays performed on single cells would enable the direct determination and identification of clusters within each cell that was being tested. However, this approach would still fail to clarify the problem completely for the following reasons: (1) restriction digestion is incomplete; (2) ligation is a probabilistic event that covalently joins only two free DNA ends; and (3) there is PCR and copy number bias.

Restriction digestion is a source of variation

The choice of enzyme affects the patterns that are observed and the interpretations that are made (e.g., ref. 5 vs. ref. 44). This is particularly important given the recent trends to map transcriptional activity back onto interaction networks, which requires small fragment sizes to allow the accurate overlapping of interactions with transcriptional or mediating factors (e.g., ref. 5 vs. ref. 44).

Recent experiments have shown that the choice of the restriction enzyme alters the amount of chromatin that is solubilized, thus affecting the frequency with which interactions are detected and refining the interaction pattern.[45] For example, the ability of a fragment to circularize[46] and be solubilized[45,47] is dependent on the length of the DNA fragment[48] and hence the choice of the restriction enzyme. Moreover, restriction enzyme choice affects ligation rates with blunt end ligation is less efficient than sticky end ligation.[49-52]

The specificity of the enzyme and any star-activity needs to be taken into account, particularly given the large quantities of enzyme used in the standard protocols. Non-specific cleavage and subsequent ligation can complicate analyses.[48] Finally, the efficiency of chromatin cleavage is important, as too low an efficiency will result in the detection of only adjacent interactions due to failure to separate the restriction fragments. Significant locus specific differences do exist, for example cleavage efficiencies[52] of 85% and >70% can be achieved with HindIII and MboI within the β-globin locus.[47] Similarly, cleavage efficiencies of between 65–89% are achieved using AseI at the mouse immunoglobulin kappa locus.[53,54]

Variable ligation is a significant cause of methodological variation in proximity ligation assays

Intra-molecular ligation is the critical step in all proximity ligation assays. As such, it is important to incorporate the physico-chemical consequences of this step into the interpretation of results from such assays. The process of ligation is an inherently probabilistic event. The probabilistic nature of this process results from the fact that ligation events mediated by enzymes, for example the T4 ligase, depend upon two compatible free ends simultaneously associating with the active site of the ligase.[50] The association of all three components (i.e., two free ends and a ligase molecule) is the limiting event, hence the improvement in ligation frequencies achieved in reactions containing a 3:1 molar excess of insert to vector DNA during cloning[50,55] and enzymes that can bind and not release the free ends.[56]

In the case of the intra-molecular ligations detected by proximity ligation, the local concentration of ligation competent free ends is artificially increased through the formation of intermolecular cross-links that hold the interacting restriction fragments together. However, the efficiency of the resulting ligation reactions is also affected by the relative flexibility of the DNA fragments. The flexibility of DNA for fragment lengths above 200 bp, is well known and given by the persistence length of 50 nm.[57] However, the fact that the restriction fragments are (1) different

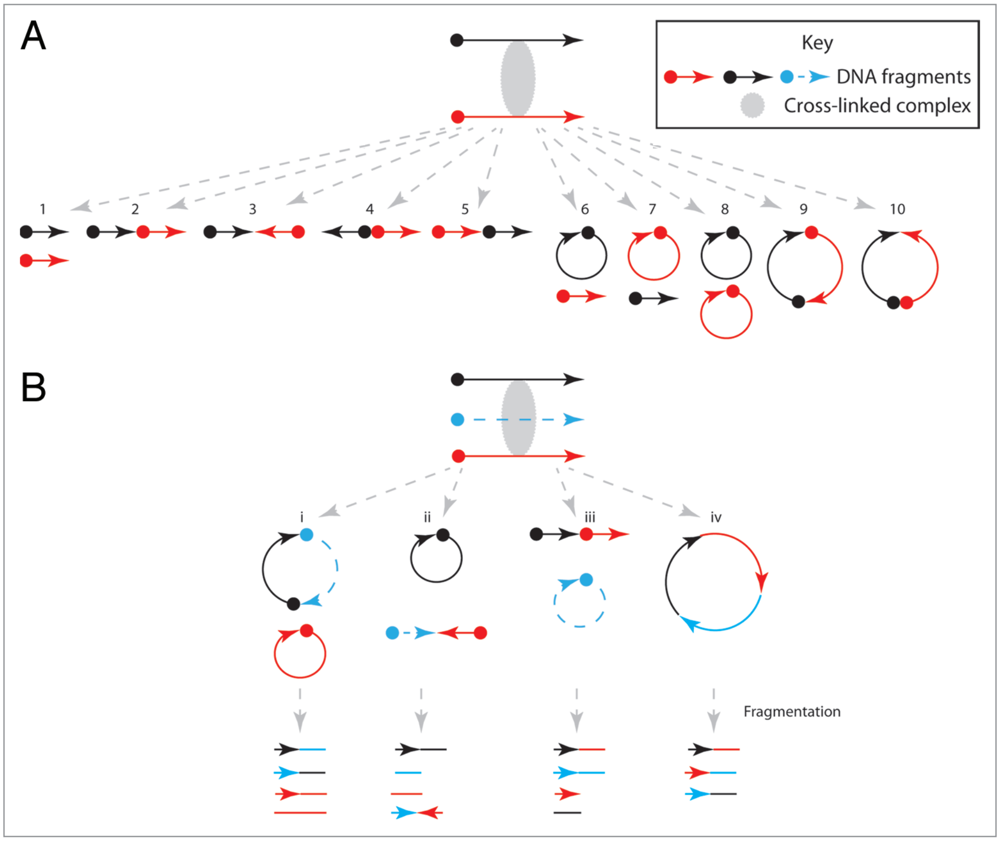

Figure 3. Ligation of restriction fragments within a cross-linked complex results in super exponentially incrementing numbers of ligation products. (**A**) Ligation events between two restriction fragments, held together within a cross-linked complex (gray oval), results in 10 possible combinations of products. (**B**) Ligation events between three or more restriction fragments result in overlapping sequences following fragmentation for sequencing.

lengths (**Fig. 2**), (2) bound by different proteins or protein complexes (e.g., nucleosomes or nucleoid associated proteins) that constrain their flexibility and ability to interact and (3) that the positions of the bound complexes or inter-linking are unlikely to be evenly distributed along or on the restriction fragments meaning that the length of DNA which is free to bend at the end can be limiting (**Fig. 2B–D**). There is ongoing debate as to the minimum length that can circularize.[46,58,59]

Arguably the single most important limitation of the proximity ligation step is that any one restriction fragment can only ligate to two other restriction fragments—one at each free end. The chances of this occurring are very low, for example Sexton et al. only identified 2 products in 167 as consisting of three ligated restriction fragments that definitely originated from a tripartite complex.[9] Moreover, even if only two ligation-competent fragments are cross-linked, ten different products can form during the ligation reaction (**Fig. 3A**). In fact the number of individual products that can form through ligation events within clusters containing incrementing numbers of restriction fragments (n) increases super exponentially such that complexes containing 10 ligation competent partners can produce 23 758 664 096 products (**Table 2**; **Supplemental Material**). Critically, not all of the products that form are able to be distinguished from each other, particularly when they are fragmented for sequencing (**Fig. 3B**).

Recent experiments have shown that those fragments which do interact typically do so with more than just their adjacent fragments and up to ~30 partners have been identified within studies in populations of yeast cells (**Table 1**). As stated earlier, proximity ligation assays do not allow the identification of clusters within a single cell. However, measurements based on transcriptional activity and numbers of active polymerases have led to estimates that transcription factories contain eight polymerases engaged on different loci in Hela nuclei.[42] Moreover, the formation of centromeric and telomere clusters indicates that biologically relevant clusters do exist. This complicates attempts to calculate actual ligation frequencies, and hence inferences about the percentage of the population in which an interaction occurs.[47]

Repetitive elements also pose a potential bias in analyses of proximity ligation data as they are difficult to position on the genome with any certainty.[34] Therefore, it is difficult to determine which particular repeat(s) are involved in a specific interaction. For example, telomeres and sub-telomeric regions are repetitive, have strand specific base compositions and are known to cluster in a wide a variety of organisms and at specific cell cycle stages (reviewed in ref. 60). This raises issues around the modeling of regions that are effectively invisible to the proximity ligation methodology. However, the junctions of specific repetitive regions can be mapped and, assuming they interact with other regions, can be used to localize the beginnings and ends of the repetitive domain. The structure of the remainder of the repetitive domain can be approximated using the biophysical characteristics of the DNA. However, the failure to be able to uniquely position interactions with repetitive regions is a limitation that needs to be borne in mind when analyzing and interpreting proximity ligation data.

Interactions with repetitive elements that are arranged in tandem (e.g., the *Saccharomyces cerevisiae* rDNA repeats) or episomes (e.g., mitochondrial genomes or high copy number plasmids) can be collapsed to one particular region or DNA element. This is a common approach, as current analyses of diploid and replicated genomes are unable to discern which copy of a particular chromosome is involved in an interaction—yet there is no a priori reason to assume that both copies are involved in the same interactions.

Current proximity ligation methods are practically and technically incapable of identifying all of the products that form from the complexes that are present within the individual cells. Therefore, the full complement of interactions that are occurring within a single cell cannot be identified even if all of the cells within a set population are individually assayed because the ligation products that represent some of these interactions cannot form—physically or statistically.

Bias due to PCR amplification

The amplification of sequences prior to sequencing, either as part of the proximity ligation methodology itself or as part of the library preparation for sequencing, can introduce bias into the results. This is particularly true for samples that have high nucleotide base biases (e.g., skewed AT or GC compositions[61,62]) and is dependent upon the enzymes, processes and equipment that are used to prepare the samples for sequencing.[61] For example, bias can be introduced at any step during Illumina sequencing, including: (1) library preparation which can introduce intermolecular ligation events, (2) cluster amplification which can affect GC bias, (3) synthesis, and (4) post-sequencing data processing.

PCR bias affects both the chances of seeing a sequence, the sequencing error probability and thus the chances of mapping it back to the reference genome.[62,63] Protocols are constantly being modified to reduce these biases (e.g., ref. 61 and "PCR-Free" sample preparation kits [TrueSeq DNA PCR-Free Sample preparation kit; FC-121–3001; Illumina]). These alterations to sequencing protocols introduce differences that affect cross-study comparisons between existing and new data sets, or for data that is generated on different platforms at different centers. Such biases are obvious in run-specific frequencies for the identification of intermolecular ligation rates such that different numbers of intermolecular ligation events are observed in different preparations of the same sample.[5]

It could be argued that the AT or GC biases we are discussing are not important when looking at genome spatial organization. However, not only do different genomes have different GC biases,[64] but repetitive elements and regions with skewed AT or GC compositions exist within prokaryotic[64,65] and eukaryotic genomes, are biologically important[61] and may play significant roles in genome organization. Centromeres are a classic example, however long nucleosome-free regions (LNFRs) in resting human T cells have skewed AT (average 73% AT) or GC (average 76% GC) contents, and these regions have been suggested to be involved in global remodeling.[66]

With the exception of the terminal fragments on linear chromosomes, all loci interact with two linearly adjacent restriction fragments when assayed by proximity ligation. These linearly adjacent interactions complicate interpretations as they may represent: (1) real interactions mediated by some protein complex; (2) protection of the restriction site; or (3) incomplete digestion. As such, they are typically removed from subsequent analyses. While effective, the question remains whether the bioinformatic removal of these interactions from data sets overcomes all of the complications associated with their presence. Empirical methods, in particular the use of modified blocker primers,[67] can be implemented to prevent or reduce the amplification of the adjacent restriction fragments in 4C assays and other few against many targeted proximity ligation assays. These methods result in the relative amplification of the signal from the less common long-range or non-adjacent interactions.

Bias due to copy number variation

Copy number of fragments or ploidy within the genome is not just a complicating factor for repetitive regions, rDNA repeats, plasmids or organelle genomes. Rather it is an issue for interactions involving any loci or chromosomes that exhibit copy number variation: (1) as result of genome replication; and (2) due to specific variation in response to environmental cues (reviewed in ref. 68). Variation due to replication can complicate analyses of populations with altered proportions in the different cell cycle stages. Alternatively, replication complicates bacterial analyses due to the fact that these cells show varying states of aneuploidy due to the existence of multiple instances of initiation from their origin of replication (e.g., *Escherichia coli*).[5] As a result of these issues, copy number variation should not simply be considered as an artifact to be removed because it complicates the mapping of the connected loci. Rather, it remains possible that connections between loci that show copy number variation are in fact biologically significant and informative.

The ability to distinguish or not effects that are correlated with alterations to copy number is an acute issue for proximity ligation studies which do not incorporate some aspect of copy number variation analysis. The GCC technique[6] enables copy number across the entire genome to be accurately determined and accounted for in studies of genome organization.[5] While this is useful, the corollary is that the amount of sequencing required is currently prohibitive for studies of large genomes (>300 Mb). It may be possible to account for copy number in other global proximity ligation techniques using external standards that are applied at known concentrations prior to the steps that are used to enrich for ligated fragments prior to sequencing.

It is possible to determine the number of different configurations of interactions using an advanced inductive combinatorial argument (**Supplemental Material**). The outcome is an algorithm which gives the number of configurations a(n) as a function of the number of candidate fragments n, which can join from either end.

Conclusion

Given the intrinsic weaknesses within proximity-ligation assays, it is clear that the amount of information that can be obtained by direct analyses of identified interactions is limited. Moreover, while some of the biases can be compensated or corrected[48] the direct interpretations of the crosslinking data directly will only inform on those elements which are physically connected and thus cross-linkable. It remains debatable how to proceed to overcome this limitation. Is it simply a question of depth of sequencing of the ligation products, or does the answer lie in the incorporation and analysis of in silico reconstructions or forward predictive models. Reconstructions of varying complexity have already been generated,[7,11,16,21,37] and the analyses of these models is complex. Predictive models can be used to ask different questions and could be generated beginning with ensembles of polymer models that recapitulate the random unpacking of DNA within the nucleus.[21,69] The integration of predicted chromosomal contacts generated using transcriptional, replication and repair networks would morph these random models into "accurate" testable representations of in vivo genome organization. Early attempts at this approach have been made.[70,71] However, the combinations of paramaters that need to be used to generate "accurate" predictive models remain unknown.

Ambiguity comes not just from the visual representations of genomic organization; rather, it is a fundamental aspect of the proximity based ligation methodologies that are being used to study structure and the spatial organization of genomes themselves. Without methodological improvements that include the continued integration of alternative data, we will remain unable to accurately identify the pattern of structural spatial associations within single cells. Statistical limitations mean that the required improvements are unlikely to be based on the proximity-based ligation methodologies as they are currently employed. However, it remains undeniable that the potential advances in our understanding of the structure:function relationships within the nucleus are key to understanding not only nuclear processes but also the processes of development and how cells respond to their environment.

Disclosure of Potential Conflicts of Interest

No potential conflict of interest was disclosed.

Supplemental Materials

Supplemental materials may be found here:
www.landesbioscience.com/journals/nucleus/article/26513

Acknowledgments

This work was supported by Gravida: National Centre for Growth and Development (O'Sullivan JM and Wake G); The Marsden fund (O'Sullivan JM); Liggins Postgraduate Scholarship (Pichungina T); and a NZ-Germany cooperation grant of the German Ministry of Education and Research (Langowski J and O'Sullivan JM).

References

1. Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K, Fauth C, Müller S, Eils R, Cremer C, Speicher MR, et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. PLoS Biol 2005; 3:e157; PMID:15839726; http://dx.doi.org/10.1371/journal.pbio.0030157
2. Fisher JK, Bourniquel A, Witz G, Weiner B, Prentiss M, Kleckner N. Four-dimensional imaging of E. coli nucleoid organization and dynamics in living cells. Cell 2013; 153:882-95; PMID:23623305; http://dx.doi.org/10.1016/j.cell.2013.04.006
3. Berger AB, Cabal GG, Fabre E, Duong T, Buc H, Nehrbass U, Olivo-Marin JC, Gadal O, Zimmer C. High-resolution statistical mapping reveals gene territories in live yeast. Nat Methods 2008; 5:1031-7; PMID:18978785; http://dx.doi.org/10.1038/nmeth.1266
4. Umbarger MA, Toro E, Wright MA, Porreca GJ, Baù D, Hong S-H, Fero MJ, Zhu LJ, Marti-Renom MA, McAdams HH, et al. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. Mol Cell 2011; 44:252-64; PMID:22017872; http://dx.doi.org/10.1016/j.molcel.2011.09.010
5. Cagliero C, Grand RS, Jones MB, Jin DJ, O'Sullivan JM. Genome conformation capture reveals that the Escherichia coli chromosome is organized by replication and transcription. Nucleic Acids Res 2013; 41:6058-71; PMID:23632166; http://dx.doi.org/10.1093/nar/gkt325
6. Rodley CD, Bertels F, Jones B, O'Sullivan JM. Global identification of yeast chromosome interactions using Genome conformation capture. Fungal Genet Biol 2009; 46:879-86; PMID:19628047; http://dx.doi.org/10.1016/j.fgb.2009.07.006
7. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. A three-dimensional model of the yeast genome. Nature 2010; 465:363-7; PMID:20436457; http://dx.doi.org/10.1038/nature08973
8. Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M, Fu Z, Noma K. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic Acids Res 2010; 38:8164-77; PMID:21030438; http://dx.doi.org/10.1093/nar/gkq955
9. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 2012; 148:458-72; PMID:22265598; http://dx.doi.org/10.1016/j.cell.2012.01.010
10. Zhang Y, McCord RP, Ho YJ, Lajoie BR, Hildebrand DG, Simon AC, Becker MS, Alt FW, Dekker J. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell 2012; 148:908-21; PMID:22341456; http://dx.doi.org/10.1016/j.cell.2012.02.002
11. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009; 326:289-93; PMID:19815776; http://dx.doi.org/10.1126/science.1181369
12. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 2006; 38:1348-54; PMID:17033623; http://dx.doi.org/10.1038/ng1896
13. Dostie J, Bickmore WA. Chromosome organization in the nucleus - charting new territory across the Hi-Cs. Curr Opin Genet Dev 2012; 22:125-31; PMID:22265226; http://dx.doi.org/10.1016/j.gde.2011.12.006
14. Tanizawa H, Noma K. Unravelling global genome organization by 3C-seq. Semin Cell Dev Biol 2012; 23:213-21; PMID:22120510; http://dx.doi.org/10.1016/j.semcdb.2011.11.003
15. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature 2012; 489:109-13; PMID:22955621; http://dx.doi.org/10.1038/nature11279
16. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science 2002; 295:1306-11; PMID:11847345; http://dx.doi.org/10.1126/science.1067799
17. Rodley CD, Grand RS, Gehlen LR, Greyling G, Jones MB, O'Sullivan JM. Mitochondrial-nuclear DNA interactions contribute to the regulation of nuclear transcript levels as part of the inter-organelle communication system. PLoS One 2012; 7:e30943; PMID:22292080; http://dx.doi.org/10.1371/journal.pone.0030943
18. Rodley CD, Pai DA, Mills TA, Engelke DR, O'Sullivan JM. tRNA gene identity affects nuclear positioning. PLoS One 2011; 6:e29267; PMID:22206006; http://dx.doi.org/10.1371/journal.pone.0029267
19. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012; 485:376-80; PMID:22495300; http://dx.doi.org/10.1038/nature11082
20. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol 2012; 30:90-8; PMID:22198700; http://dx.doi.org/10.1038/nbt.2057

21. Gehlen LR, Gruenert G, Jones MB, Rodley CD, Langowski J, O'Sullivan JM. Chromosome positioning and the clustering of functionally related loci in yeast is driven by chromosomal interactions. Nucleus 2012; 3:370-83; PMID:22688649; http://dx.doi.org/10.4161/nucl.20971

22. Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol 2011; 18:107-14; PMID:21131981; http://dx.doi.org/10.1038/nsmb.1936

23. Rubinstein M, Colby RH. Polymer physics. Oxford: Oxford University Press, 2003.

24. Grosberg AIU, Khokhlov AR. Statistical physics of macromolecules. New York: AIP Press, 1994.

25. Langowski J, Heermann DW. Computational modeling of the chromatin fiber. Semin Cell Dev Biol 2007; 18:659-67; PMID:17936653; http://dx.doi.org/10.1016/j.semcdb.2007.08.011

26. Parada LA, McQueen PG, Misteli T. Tissue-specific spatial organization of genomes. Genome Biol 2004; 5:R44; PMID:15239829; http://dx.doi.org/10.1186/gb-2004-5-7-r44

27. Fraser J, Rousseau M, Shenker S, Ferraiuolo MA, Hayashizaki Y, Blanchette M, Dostie J. Chromatin conformation signatures of cellular differentiation. Genome Biol 2009; 10:R37; PMID:19374771; http://dx.doi.org/10.1186/gb-2009-10-4-r37

28. Zhang H, Jiao W, Sun L, Fan J, Chen M, Wang H, Xu X, Shen A, Li T, Niu B, et al. Intrachromosomal looping is required for activation of endogenous pluripotency genes during reprogramming. Cell Stem Cell 2013; 13:30-5; PMID:23747202; http://dx.doi.org/10.1016/j.stem.2013.05.012

29. Apostolou E, Ferrari F, Walsh RM, Bar-Nur O, Stadtfeld M, Cheloufi S, Stuart HT, Polo JM, Ohsumi TK, Borowsky ML, et al. Genome-wide chromatin interactions of the Nanog locus in pluripotency, differentiation, and reprogramming. Cell Stem Cell 2013; 12:699-712; PMID:23665121; http://dx.doi.org/10.1016/j.stem.2013.04.013

30. Andrey G, Montavon T, Mascrez B, Gonzalez F, Noordermeer D, Leleu M, Trono D, Spitz F, Duboule D. A switch between topological domains underlies HoxD genes collinearity in mouse limbs. Science 2013; 340:1234167; PMID:23744951; http://dx.doi.org/10.1126/science.1234167

31. Marshall WF, Fung JC, Sedat JW. Deconstructing the nucleus: global architecture from local interactions. Curr Opin Genet Dev 1997; 7:259-63; PMID:9115425; http://dx.doi.org/10.1016/S0959-437X(97)80136-0

32. Starr DA. A nuclear-envelope bridge positions nuclei and moves chromosomes. J Cell Sci 2009; 122:577-86; PMID:19225124; http://dx.doi.org/10.1242/jcs.037622

33. de Laat W, Dekker J. 3C-based technologies to study the shape of the genome. Methods 2012; 58:189-91; PMID:23199640; http://dx.doi.org/10.1016/j.ymeth.2012.11.005

34. Grand RS, Gehlen LR, O'Sullivan JM. Methods for the investigation of chromosome organization. In: Urbano KV, ed. Advances in Genetics Research. New York: Nova Publishers, 2011.

35. Sajan SA, Hawkins RD. Methods for identifying higher-order chromatin structure. Annu Rev Genomics Hum Genet 2012; 13:59-82; PMID:22703176; http://dx.doi.org/10.1146/annurev-genom-090711-163818

36. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, et al. Determining the architectures of macromolecular assemblies. Nature 2007; 450:683-94; PMID:18046405; http://dx.doi.org/10.1038/nature06404

37. Ben-Elazar S, Yakhini Z, Yanai I. Spatial localization of co-regulated genes exceeds genomic gene clustering in the Saccharomyces cerevisiae genome. Nucleic Acids Res 2013; 41:2191-201; PMID:23303780; http://dx.doi.org/10.1093/nar/gks1360

38. de Laat W, Grosveld F. Spatial organization of gene expression: the active chromatin hub. Chromosome Res 2003; 11:447-59; PMID:12971721; http://dx.doi.org/10.1023/A:1024922626726

39. Patrinos GP, de Krom M, de Boer E, Langeveld A, Imam AM, Strouboulis J, de Laat W, Grosveld FG. Multiple interactions between regulatory regions are required to stabilize an active chromatin hub. Genes Dev 2004; 18:1495-509; PMID:15198986; http://dx.doi.org/10.1101/gad.289704

40. Jackson DA, Balajee AS, Mullenders L, Cook PR. Sites in human nuclei where DNA damaged by ultraviolet light is repaired: visualization and localization relative to the nucleoskeleton. J Cell Sci 1994; 107:1745-52; PMID:7983144

41. Pombo A, Jackson DA, Hollinshead M, Wang Z, Roeder RG, Cook PR. Regional specialization in human nuclei: visualization of discrete sites of transcription by RNA polymerase III. EMBO J 1999; 18:2241-53; PMID:10205177; http://dx.doi.org/10.1093/emboj/18.8.2241

42. Jackson DA, Iborra FJ, Manders EMM, Cook PR. Numbers and organization of RNA polymerases, nascent transcripts, and transcription units in HeLa nuclei. Mol Biol Cell 1998; 9:1523-36; PMID:9614191; http://dx.doi.org/10.1091/mbc.9.6.1523

43. Iborra FJ, Pombo A, Jackson DA, Cook PR. Active RNA polymerases are localized within discrete transcription "factories' in human nuclei. J Cell Sci 1996; 109:1427-36; PMID:8799830

44. Wang W, Li GW, Chen C, Xie XS, Zhuang X. Chromosome organization by a nucleoid-associated protein in live bacteria. Science 2011; 333:1445-9; PMID:21903814; http://dx.doi.org/10.1126/science.1204697

45. Gavrilov AA, Gushchanskaya ES, Strelkova O, Zhironkina O, Kireev II, Iarovaia OV, Razin SV. Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. Nucleic Acids Res 2013; 41:3563-75; PMID:23396278; http://dx.doi.org/10.1093/nar/gkt067

46. Ringrose L, Chabanis S, Angrand PO, Woodroofe C, Stewart AF. Quantitative comparison of DNA looping in vitro and in vivo: chromatin increases effective DNA flexibility at short distances. EMBO J 1999; 18:6630-41; PMID:10581237; http://dx.doi.org/10.1093/emboj/18.23.6630

47. Gavrilov AA, Golov AK, Razin SV. Actual ligation frequencies in the chromosome conformation capture procedure. PLoS One 2013; 8:e60403; PMID:23555968; http://dx.doi.org/10.1371/journal.pone.0060403

48. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet 2011; 43:1059-65; PMID:22001755; http://dx.doi.org/10.1038/ng.947

49. Pheiffer BH, Zimmerman SB. Polymer-stimulated ligation: enhanced blunt- or cohesive-end ligation of DNA or deoxyribooligonucleotides by T4 DNA ligase in polymer solutions. Nucleic Acids Res 1983; 11:7853-71; PMID:6359064; http://dx.doi.org/10.1093/nar/11.22.7853

50. Sambrook J, Fritsch EF, Maniatis T. Molecular cloning: A laboratory manual. Cold Spring Harbor Laboratory Press, Clod Spring Harbor, New York, 1989.

51. Zyskind JW, Bernstein SI. Recombinant DNA laboratory manual. San Diego: Academic Press, 1992.

52. Naumova N, Smith EM, Zhan Y, Dekker J. Analysis of long-range chromatin interactions using Chromosome Conformation Capture. Methods 2012; 58:192-203; PMID:22903059; http://dx.doi.org/10.1016/j.ymeth.2012.07.022

53. Liu Z, Garrard WT. Long-range interactions between three transcriptional enhancers, active Vkappa gene promoters, and a 3′ boundary sequence spanning 46 kilobases. Mol Cell Biol 2005; 25:3220-31; PMID:15798207; http://dx.doi.org/10.1128/MCB.25.8.3220-3231.2005

54. Simonis M, Kooren J, de Laat W. An evaluation of 3C-based methods to capture DNA interactions. Nat Methods 2007; 4:895-901; PMID:17971780; http://dx.doi.org/10.1038/nmeth1114

55. Frackman S, Kephart D. Rapid Ligation for the pGEM-T adn pGEM-T Easy Vector Systems. Promega Notes 1999; 71:8

56. Wilson RH, Morton SK, Deiderick H, Gerth ML, Paul HA, Gerber I, Patel A, Ellington AD, Hunicke-Smith SP, Patrick WM. Engineered DNA ligases with improved activities in vitro. Protein Eng Des Sel 2013; 26:471-8; http://dx.doi.org/10.1093/protein/gzt024; PMID:23754529

57. Hagerman PJ. Flexibility of DNA. Annu Rev Biophys Biophys Chem 1988; 17:265-86; PMID:3293588; http://dx.doi.org/10.1146/annurev.bb.17.060188.001405

58. Shore D, Langowski J, Baldwin RL. DNA flexibility studied by covalent closure of short fragments into circles. Proc Natl Acad Sci U S A 1981; 78:4833-7; PMID:6272277; http://dx.doi.org/10.1073/pnas.78.8.4833

59. Vafabakhsh R, Ha T. Extreme bendability of DNA less than 100 base pairs long revealed by single-molecule cyclization. Science 2012; 337:1097-101; PMID:22936778; http://dx.doi.org/10.1126/science.1224139

60. Harper L, Golubovskaya I, Cande WZ. A bouquet of chromosomes. J Cell Sci 2004; 117:4025-32; PMID:15316078; http://dx.doi.org/10.1242/jcs.01363

61. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 2011; 12:R18; PMID:21338519; http://dx.doi.org/10.1186/gb-2011-12-2-r18

62. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 2012; 40:e72; PMID:22323520; http://dx.doi.org/10.1093/nar/gks001

63. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 2008; 36:e105; PMID:18660515; http://dx.doi.org/10.1093/nar/gkn425

64. Bohlin J, Snipen L, Hardy SP, Kristoffersen AB, Lagesen K, Dønsvik T, Skjerve E, Ussery DW. Analysis of intra-genomic GC content homogeneity within prokaryotes. BMC Genomics 2010; 11:464; PMID:20691090; http://dx.doi.org/10.1186/1471-2164-11-464

65. Haywood-Farmer E, Otto SP. The evolution of genomic base composition in bacteria. Evolution 2003; 57:1783-92; PMID:14503620

66. Schwarzbauer K, Bodenhofer U, Hochreiter S. Genome-wide chromatin remodeling identified at GC-rich long nucleosome-free regions. PLoS One 2012; 7:e47924; PMID:23144837; http://dx.doi.org/10.1371/journal.pone.0047924

67. Vestheim H, Jarman SN. Blocking primers to enhance PCR amplification of rare sequences in mixed samples - a case study on prey DNA in Antarctic krill stomachs. Front Zool 2008; 5:12; PMID:18638418; http://dx.doi.org/10.1186/1742-9994-5-12

68. Kondrashov FA. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc Biol Sci 2012; 279:5048-57; PMID:22977152; http://dx.doi.org/10.1098/rspb.2012.1108

69. Tjong H, Gong K, Chen L, Alber F. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. Genome Res 2012; 22:1295-305; PMID:22619363; http://dx.doi.org/10.1101/gr.129437.111

70. Li S, Heermann DW. Transcriptional regulatory network shapes the genome structure of Saccharomyces cerevisiae. Nucleus 2013; 4:216-28; PMID:23674068; http://dx.doi.org/10.4161/nucl.24875

71. Fritsche M, Li S, Heermann DW, Wiggins PA. A model for Escherichia coli chromosome packaging supports transcription factor-induced DNA domain formation. Nucleic Acids Res 2012; 40:972-80; PMID:21976727; http://dx.doi.org/10.1093/nar/gkr779

72. Dostie J, Dekker J. Mapping networks of physical interactions between genomic elements using 5C technology. Nat Protoc 2007; 2:988-1002; PMID:17446898; http://dx.doi.org/10.1038/nprot.2007.116