More to Hi-C than meets the eye

Myong-Hee Sung & Gordon L Hager

Diversification and specialization of high-throughput technologies demand assay-specific treatment of data for reliable interpretation. A new study shows that data generated using the Hi-C approach contain hidden features of interchromosomal DNA interactions, which are revealed through analysis with an integrated probabilistic model that corrects for multiple sources of bias in the data.

Progress in many areas of biology has been fueled by new technology. Sophisticated techniques that enable the examination of various molecular processes at ever-increasing depth and coverage are evolving at a mind-boggling pace. Until recently, high-throughput innovations were limited, with the most notable being gene expression profiling by DNA microarrays. Bioinformaticians tackled the underlying quantitative issues and proposed numerous computational methods to alleviate pitfalls in early applications. Eventually the analysis of expression microarray data matured to a consensus-forming stage[1]. Now, a multitude of innovative experimental approaches driven by breakthroughs in parallel-sequencing technology (such as RNA-seq, ChIP-seq, DNase-seq, 4C, 5C, ChIA-PET, Gro-seq, RIP-seq, MeDIP-seq) have been developed, leading to even more complex analysis issues. When analyzing raw data from high-throughput assays, one often finds that the data does not "speak for itself." Among the many contributing factors is a lack of mathematical and statistical models appropriate for the particular assay, where the complexity of computational problems can baffle the users as well as pioneers of new technology. In this issue of *Nature Genetics*, Yaffe and Tanay take a second look at the recently described Hi-C method[2] and uncover additional insights concerning interchromosomal DNA interaction networks[3].

The challenges of Hi-C

Hi-C was designed to be the ultimate extension of the chromosome conformation capture (3C) assay. In 3C and its variants, the cross-linked genome is enzyme digested and the resulting pieces of genomic DNA are ligated. Linkages along the linear chromosomes are erased, and new connections are generated based on proximity within the nucleus[4,5]. The Hi-C approach was designed to detect all pairwise physical associations of DNA in the genome. This is achieved by incorporating biotin into the ends of the digested DNA before ligation, and then carrying out physical selection of these fragments and paired-end sequencing. This technique enables a comprehensive identification of *cis* and *trans* DNA-DNA interactions, which should provide insight into nonrandom higher-order organizational features of the genome.

This extraordinary advance comes at a high price. Simple calculations illustrate a major limitation: the human genome is cut into ~1.5 million fragments by a commonly used 6-bp–recognizing restriction enzyme. So the number of possible pairwise contacts between the DNA fragments is roughly $(1.5 \text{ million})^2$, or 10^{12}, which is then probed by paired-end sequencing reads whose number typically ranges from 10^7 to 10^8. That means that the average number of reads sampled per possible fragment pair is at most 10^{-4}, effectively producing no useful information at the restriction-fragment level. Instead, the original Hi-C study[2] and the study by Tanay and Yaffe[3] extracted information at a larger scale, with 1-Mb bins along the genome. Hi-C experiments still remain relatively low resolution because the two-way coverage of the genome (all of the genome compared against all of the genome) makes the assay scale N^2 (where N is the total number of potential contacts): for example, a 10-fold improvement in resolution requires 100-fold deeper sequencing, which is prohibitive given current cost and yield.

Proper analysis of Hi-C also seems to be a great challenge. In their report, Yaffe and Tanay carry out a careful computational investigation and demonstrate that the raw data contain systematic biases from multiple steps in the assay protocol. The artifacts are especially detrimental to detection of interchromosomal (*trans*)

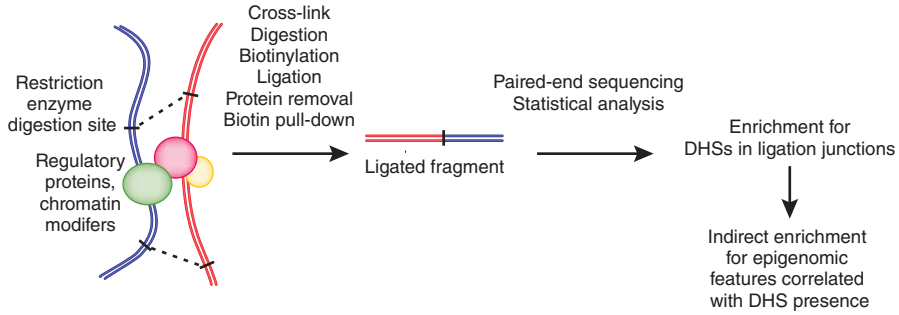

Figure 1 Correcting Hi-C data for systematic biases and subsequent analyses suggest that DNase I hypersensitivity is most directly correlated with physical DNA interactions. Ligation junctions formed by spatial proximity tend to harbor DHSs that correspond to cell type–specific regulatory factor–binding sites. Other features, such as epigenetic marks related to transcriptional activity, are also enriched through their association with DHSs.

Myong-Hee Sung and Gordon L. Hager are at the Laboratory of Receptor Biology and Gene Expression, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA. e-mail: hagerg@exchange.nih.gov

interactions whose signal is subtle (the probability of *trans* contact is overall much lower than that of intrachromosomal (*cis*) contact[6]). In fact, the authors noticed that Hi-C data sets derived from two different enzymes produced *trans* contact maps that were not correlated, even though, in theory, the choice of restriction enzyme should not change the biological conclusions. They identified four different sources of bias, among which the local GC content in ligated DNA was the dominant factor behind the enzyme-specific profiles. Here, a simple successive correction of biases would not work due to the intertwined error structures. Rather, they used a probabilistic model to account for existing biases by a maximum-likelihood method. The corrected *trans*-contact map is now concordant between the two replicate data sets from different restriction enzymes. From the *trans*-contact map, three clusters of genomic regions emerge: active loci, inactive and centromere-proximal loci, and inactive and centromere-distal loci. This refines the two spatial compartments reported in the original Hi-C study[2]. As an added assurance, the contiguity along the linear chromosomes was recovered solely from the *trans*-interaction patterns. However, it is unclear whether the centromere-based segregation of repressed chromatin reflects an interphase organization or an artifact of actively dividing blasts.

Interchromosomal interactions

Historically, interchromosomal interactions have been more elusive than *cis* contacts[7,8]. For one thing, the signal strength is modest regardless of the experimental assay, whether in a single-cell analysis or a 3C variant[3,8]. Furthermore, it has been difficult to determine the relevant epigenetic features that influence *trans* interactions, as multiple marks are apparently associated with contact regions. Discovery of the most directly associated genomic feature would undoubtedly shed light on the mechanisms driving nonrandom organizations of the nucleus. To address this, Yaffe and Tanay looked for global correlates of organization by relating the corrected Hi-C data to available epigenomic profiles from the ENCODE human lymphoblast database, including DNase I hypersensitivity, histone marks, RNA polymerase II occupancy and CTCF binding. Extensive enrichment analysis and comparisons singled out the presence of DNase I hypersensitive sites (DHSs) as an important hallmark of *trans* interactions (**Fig. 1**). The DHS-associated effect on contact probability persists largely independently of other markers, whereas the other features' effects depend on DHS presence. Because DHSs often represent regulatory sites that are created and/or maintained by transcription factors and other chromatin-binding proteins[9–11], the nuclear architecture may be shaped collectively by the cell type–specific ensemble of chromatin-modifying proteins.

Our knowledge is still based on a small number of studies. However, with a well-formulated methodology now in place, the genomics community is primed to extract robust nuclear architectural information from Hi-C experiments on more cell types and states.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Irizarry, R.A., Wu, Z. & Jaffee, H.A. *Bioinformatics.* **22**, 789–794 (2006).
2. Lieberman-Aiden, E. *et al. Science* **326**, 289–293 (2009).
3. Yaffe, E. & Tanay, A. *Nat. Genet.* **43**, 1059–1065 (2011).
4. Dekker, J. *Nat. Methods* **3**, 17–21 (2006).
5. Simonis, M., Kooren, J. & de Laat, W. *Nat. Methods* **4**, 895–901 (2007).
6. Hakim, O., Sung, M.H. & Hager, G.L. *Curr. Opin. Cell Biol.* **22**, 305–313 (2010).
7. Kocanova, S. *et al. PLoS Genet.* **6**, e1000922 (2010).
8. Hakim, O. *et al. Genome Res.* **21**, 697–706 (2011).
9. Hesselberth, J.R. *et al. Nat. Methods* **6**, 283–289 (2009).
10. John, S. *et al. Nat. Genet.* **43**, 264–268 (2011).
11. Biddie, S.C. *et al. Mol. Cell* **43**, 145–155 (2011).

A transposon in *tb1* drove maize domestication

Miltos Tsiantis

A new study shows an inserted retroelement in the regulatory sequences of the maize *tb1* gene, which controls shoot branching, was the target of human selection during the domestication of maize from its wild relative teosinte. The insertion allele was already present at low frequency in teosinte populations before selection, highlighting the significance of standing genetic variation in the evolution of morphological diversity.

A major challenge in biology is to understand the origins of morphological diversity—how form changes through evolution. As Darwin noted, domestication offers valuable insight into this problem by providing a direct path between ancestral and descendant species[1]. For example, maize was domesticated from its wild ancestor teosinte, and this process resulted in reduced branching (**Fig. 1**). Twenty years ago, in a search for the genes underpinning maize evolution, John Doebley and colleagues conducted a simple, yet powerful, experiment that exploited the interfertility of maize and teosinte. They crossed the two species and found that allelic variation at five loci accounted for their major differences in form[2]. One of these genes was *teosinte branched 1* (*tb1*), which encodes a transcriptional regulator involved in growth repression. In maize, *tb1* expression is elevated relative to teosinte, correlating with repressed branch outgrowth[3]. A signature of selection was also identified at the *tb1* locus in the form of a selective sweep, suggesting that *tb1* was targeted by human selection[4]. That is, the *tb1* gene showed much less nucleotide diversity in maize relative to teosinte compared to genes that have not been targeted by human selection. Reduced levels of polymorphism were found in upstream regions but not in the coding sequence of *tb1*, suggesting that regulatory changes contributed to the differences in form between maize and teosinte. In this issue of *Nature Genetics*, Doebley and colleagues make a key advance in understanding the precise genetic basis for the origin of maize by showing that a transposon insertion 60 kb upstream of the *tb1* ORF underlies species-specific *tb1* expression and consequent diversification of morphology[5].

Miltos Tsiantis is in the Department of Plant Sciences at the University of Oxford, Oxford, UK.
e-mail: miltos.tsiantis@plants.ox.ac.uk

Hunting causal variants

To delimit the genetic interval that contains the variants responsible for morphological diversity, the authors followed two complementary approaches. First, they mapped recombination breakpoints in maize-teosinte introgression lines to determine the phenotypic consequences of harboring different