Parallel genome universes

Tom Misteli

National Cancer Institute, US National Institutes of Health, Bethesda,Maryland, USA., mistelit@mail.nih.gov

A new computational approach gives us the best chance at understanding how genomes are arranged in three-dimensional space and what that may mean for their function.

It is remarkable to realize that despite all the talk—and work—on genomes and their sequences, nobody knows how they are actually organized inside the cell[1]. The recent development of biochemical methods to probe global genome organization has given a glimmer of hope that systematic mapping of genomes in space will soon be possible. In this issue, Kalhor *et al.*[2] report an improved experimental protocol and a breakthrough in how to analyze data from these approaches. The computational strategy provides a more realistic picture of the genome landscape and will help elucidate the links between genome organization and function.

The method of choice for mapping the topology of genomes in cells is a group of techniques referred to as chromosome conformation capture[3]. In these approaches, chromatin in intact cells is chemically crosslinked to capture the physical associations between genome regions. The DNA is then chopped up with a frequently cutting restriction enzyme, and the resulting fragments are re-ligated at very high dilution such that the joining of crosslinked fragments is favored over the joining of noncrosslinked DNA. Finally, the identity of the crosslinked regions, which reflect the original spatial genome associations, is determined by PCR or by sequencing, depending on the specific variation of the chromosome conformation capture method. The development of these methods recently culminated in an approach called Hi-C, which allows unbiased probing of all genome interactions in a cell population using high-throughput sequencing, providing a tool to comprehensively map genome interactions[4].

Kalhor *et al.*[2] have now refined chromosome conformation capture methods in two ways. The first advance improves the quality of the primary data. Rather than doing the ligation step in solution, they tether the DNA fragments to beads for ligation. This simple step substantially improves the signal-to-noise ratio, resulting in a cleaner, more reliable list of genome interactions. This is a nice refinement of Hi-C, but the true value of Kalhor *et al.*'s[2] work lies in the second improvement, which relates to data analysis.

Chromosome conformation capture methods generate long lists of inter action frequencies between genome regions. Using computational approaches, one can then derive a three-dimensional map of the genome by determining the neighbors of any given point (Fig. 1). A fundamental problem with this approach is that the data used in generating the spatial maps represent the inter actions that occur in the entire population of the millions of cells used in the experiment. Therefore, the resulting genome landscape depicts the average pattern of organization, assuming that all cells in the population are the same or at least similar, which we know is generally not true. Which genome associations occur in which individual cells or subpopulations of cells cannot be determined from this approach. For example, chromosome conformation capture may detect a high frequency of interaction between a chromosome region and two other chromosomes, but it cannot determine whether these two associations occur simultaneously in the same cell or separately in two populations of cells.

Kalhor *et al.*[2] now begin to address this very important but thorny issue. To computationally analyze their data, they create a large collection of models representing a spectrum of possible configurations of the chromosomes in space, rather than searching for one model that can accommodate all interaction data as has been done previously (Fig. 1). This population-based modeling approach does not attempt to squeeze all of the experimental data into a single best-fit model, but instead searches for an ensemble of models to accurately represent the primary data. This set of models is selected computationally by using a scoring function and an optimization algorithm that matches the variability of genome configurations in the various models with the statistical variability present in the primary interaction data.

The results from this population-based analysis suggest that the approach works. First of all, the authors find remarkable diversity among models that can account for the experimental data. Few genome interactions are found in all models, and as little as ~20% of genome contacts are shared between any two models. This observation is in line with what is seen when chromosome conformation capture methods are validated by single-cell techniques such as fluorescence *in situ* hybridization, which allows mapping of genome interactions in individual cells using microscopy. When comparing data from chromosome conformation capture methods with fluorescence *in situ* hybridization data, one typically finds that a positive interaction often occurs in as little as 5% of cells in a population, clearly suggesting heterogeneity in the population[5].

Nevertheless, this does not mean that there are no patterns in spatial genome organization—quite the opposite. Kalhor *et al.*[2] identify two of the best-characterized cell biological properties of genomes as common features in their models. They find nonrandom positioning of chromosomes in a radial pattern—that is, some chromosomes preferentially localize to the center of the nucleus, others to the periphery. And they find clustering of specific chromosomes, with chromosomes containing an abundance of transcribed regions clustering in the nuclear interior[6]. The identification of these well- established properties is a reassuring validation of the approach.

Kalhor *et al.*'s 'parallel universe' approach has several advantages. It clearly provides a more-nuanced representation of the primary data. The method also yields more realistic

models of the heterogeneity of a population of cells. It is increasingly clear that individual cells in a population, even those grown in the laboratory, do not all behave in the same way. For example, a set of genes may appear to be constitutively active when a cell population is analyzed *en masse* by PCR, but when single cells are probed, one may find that the genes cycle between on and off states[7].

The most important consequence of population-based data analysis, however, is that it is a first step toward testing new questions. How many distinct biological subpopulations exist in a cell population? Does the organization of the genome shift in these populations? And are the population's overall biological changes driven by the ensemble behavior of the population or by a small subpopulation (e.g., stem cells)? Answering these questions will go a long way toward understanding the central question in the field, which is how genome organization is linked to biological function.

It is natural that we seek simplicity when we explore biological processes. But at times, it is important to acknowledge their complexity.We are at such a point in the study of genome organization. Most experimental data point to a high degree of heterogeneity and an element of stochastic organization. The methods we use to analyze genomes must take these realities into account

References

1. Misteli T Cell 128, 787–800 (2007). [PubMed: 17320514]

2. Kalhor R, Tjong H, Jayahilaka N, Alber F & Chen L Nat. Biotechnol 30, 90–98 (2012).

3. Simonis M, Kooren J & de Laat W Nat. Methods 4, 895–901 (2007). [PubMed: 17971780]

4. Lieberman-Aiden E et al. Science 326, 289–293 (2009). [PubMed: 19815776]

5. Hakim O et al. Genome Res. 21, 697–706 (2011). [PubMed: 21471403]

6. Takizawa T, Meaburn KJ & Misteli T Cell 135, 9–13 (2008). [PubMed: 18854147]

7. Levsky JM, Shenoy SM, Pezo RC & Singer RH Science 297, 836–840 (2002). [PubMed: 12161654]
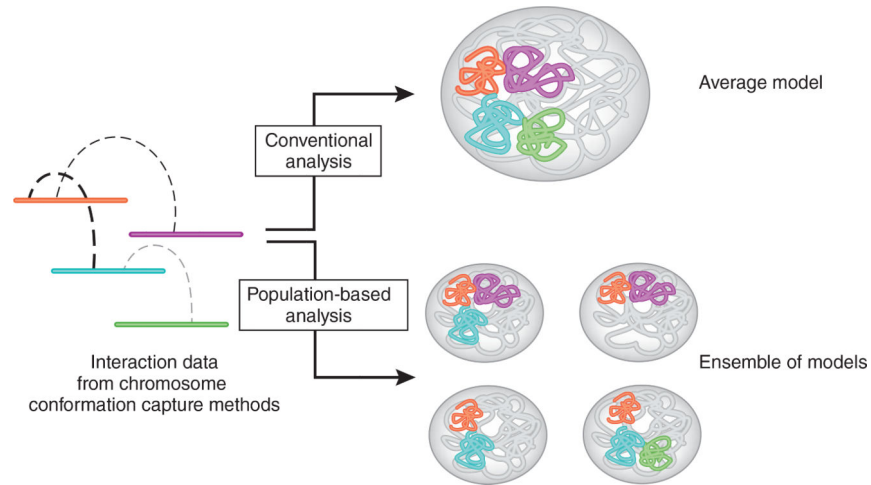
Figure 1.
Population-based modeling of genome organization. Chromosome conformation capture techniques generate information about the frequencies of interactions (dashed arcs; darker indicates greater frequency) between genome locations on various chromosomes (orange, blue, purple and green). Modeling approaches can use this frequency information to determine the spatial arrangement of chromosomes. Conventionally, all interaction data are pooled to generate a single average-based model of genome organization (top). Kalhor *et al.*[2] use an alternative approach, population-based modeling, in which the interaction data are represented in a collection of models whose composition is based on the frequencies of interactions (bottom). For instance, the most frequent interaction, which is between the orange and blue chromosomes, occurs in the most models. This method more accurately represents the heterogeneity of genome organization seen in a cell population.