ASSEMBLING THE 3-D GENOME:
A Puzzle with Many Solutions

By Katharine Miller

As a result of experimental techniques developed about a decade ago, researchers now have data that can be used to reconstruct how the genome is arranged inside the nucleus. This 3-D structure likely plays a role in determining cellular function by affecting cells' ability to access, read and interpret genetic information.

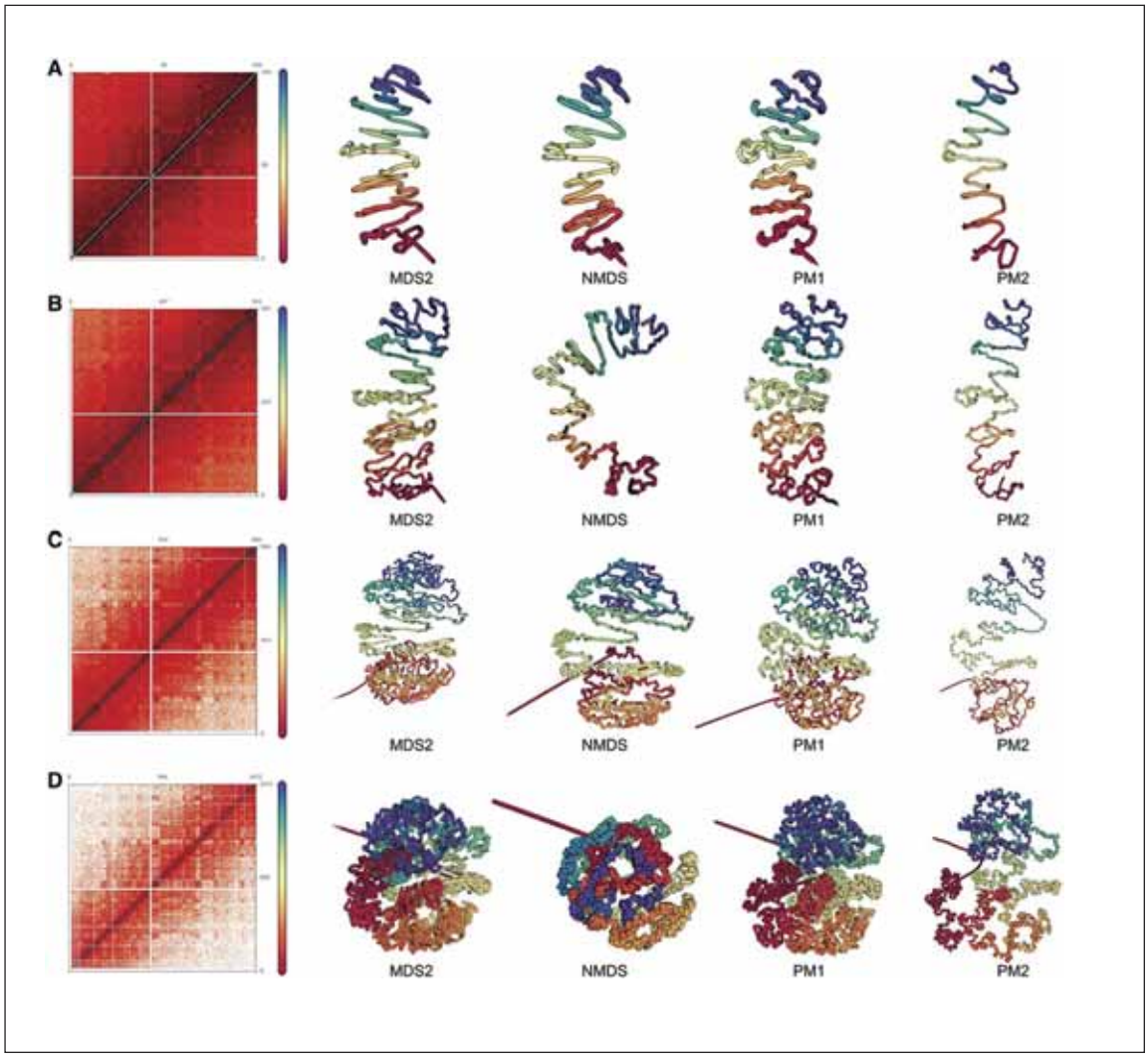
"We want to use 3-D genome reconstruction to understand the guiding principles of genome organization," says **Frank Alber, PhD**, associate professor of molecular and computational biology at the University of Southern California. "There is a lot to be learned. We are just at the beginning."

Experiments called chromosome conformation capture—of which there are now multiple types, including 3C, 4C, 5C and Hi-C—allow scientists to determine the frequency with which loci on the genome are in contact with one another—considering all possible interactions. These contact frequencies are derived from experiments that are done on 10 to 20 million cells at a time, and therefore do not represent the 3-D structure of any one cell. Using computational approaches, however, researchers have developed ways to assemble plausible 3-D structures.

One approach is to convert contact frequencies to Euclidean distances—using one of several mathematical or probabilistic options—and then optimize the distances to generate a consensus structure. Other researchers try to infer which contacts co-occur and then generate an ensemble of possible structures. Both methods—consensus and ensemble—generate structures that are essentially fictional: There is currently no way to know whether a structure generated by these computational methods actually occurs in nature. Yet insights can be gained from these methods nonetheless.

Converting Frequencies to Distance

Researchers have tried a variety of mathematical approaches to convert contact frequencies into distance. Initial efforts assumed that the frequency of intrachromosomal contacts could be directly mapped to Euclidean distance in 3-D, says **William Noble, PhD**, professor of genome sciences at the University of Washington. They plotted genomic distance as a function of contact count and then swapped genomic distance for a distance in 3-D (Euclidean distance) that was calibrated using imaging.



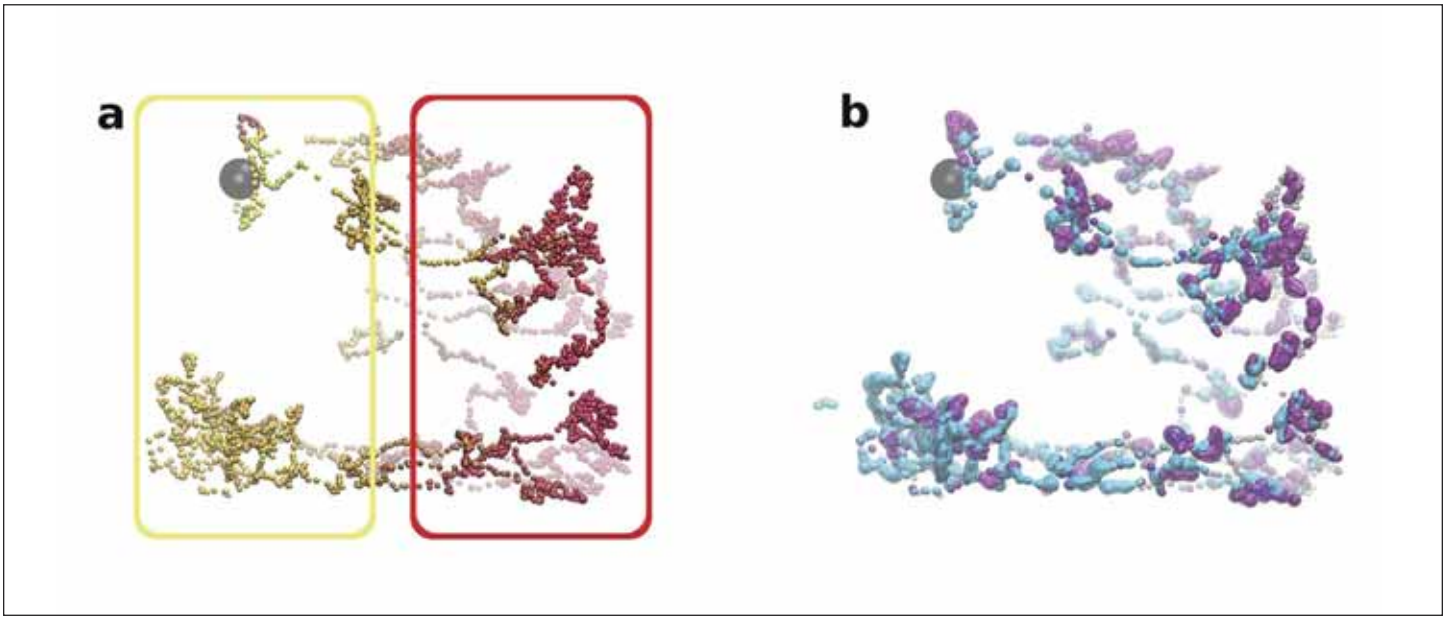
Noble and his colleagues compared two different variants of his Poisson method of predicting the 3-D structure of chromosome 1 with several different multidimensional scaling algorithms (MDS) at different resolutions: 1 Mb (A), 500 kb (B), 200 kb (C) and 100 kb (D). The second Poisson method was more stable in response to resolution changes than were the other methods. Reprinted from N Varoquaux, F Ay, WS Noble, and JP Vert, A statistical approach for inferring the 3D structure of the genome,* Bioinformatics *(2014) 30 (12): i26-i33.

"It's basically a ruler," Noble says. Other methods have determined the ruler differently. For example, a tool called ChromSDE fits a parametric curve to the data. And recently Noble used a model that assumes the contacts are generated according to a Poisson process where events occur randomly over a given time interval with a particular (Poisson) form of distribution.

In each case, researchers optimize the distances to converge on a single, consensus structure that is essentially an average of all possible structures that could exist in the millions of cells sampled. Noble says the consensus structure is useful for visualization and hypothesis generation, but he cautions: "It's risky to conclude anything from these models alone, and validations [using fluorescence *in situ* hybridization or FISH] are expensive and therefore sparse."

A recent paper published in *Nature Methods* by **Julien Mozziconacci, PhD**, Lecturer in Physics and Biology at Pierre & Marie Curie University, Paris, France, and his colleagues took a graph theoretic approach to the same problem. They graphed the contact frequencies from a set of Hi-C data as a single structure where the weights on the graph are the inverse of the contact frequency. "The higher the frequency, the closer the distance," Mozziconnaci says. For unconnected nodes, the graph assigned the shortest possible distance in order to fulfill the triangular inequality, i.e., given three points in space—A, B, and C—the sum of the distance between A and B and the distance between B and C is always greater than or equal to the distance between A and C. "If you don't have this property then you are not talking about distances," Mozziconnaci says.

Genome sequence data and related information, such as function and epigenomic data, are essentially one-dimensional. Annotating the 3-D structure with this information and viewing it in a 3-D genome browser allows novel observations, Mozziconacci says.



Unlike the other approaches described above, Mozziconacci's approach, called ShRec3D, does not include an iterative optimization step. "The matrix analysis directly gives the structure," he says. The triangular inequality is satisfied on the graph, but not necessarily in Euclidean 3-D space. "In the end, the structure is a view of the mind. There is no such structure in 3-D space."

Despite their fictional nature, one advantage of the various consensus approaches, Mozziconacci says, is that they can be integrated with a 3-D genome browser. Genome sequence data and related information, such as function and epigenomic data, are essentially one-dimensional. Annotating the 3-D structure with this information and viewing

ShRec3D allows researchers to superimpose existing information onto reconstructed 3-D chromosome structures. For example, chromatin might be partitioned into compartments as shown in (a), where yellow indicates gene-rich, GC-rich regions, on the left and red indicates gene-poor, AT-rich regions, on the right. Or researchers might display, on the 3-D structure, linear information such as shown in (b), where cyan regions harbor a high level of acetylation; pink regions harbor a high level of tri-methylation; and purple regions harbor both modifications. Reprinted with permission from A Lesne, J Riposo, P Roger, A Cournac, J Mozziconacci, 3-D genome reconstruction from chromosomal contacts,* Nature Methods *(2014) doi:10.1038/nmeth.3104.

it in a 3-D genome browser allows novel observations, Mozziconacci says. For example, a researcher might display where a particular transcription factor lies in 3-D space relative to other loci with which it is known to interact. Mozziconacci looks forward to a time when different techniques, such as sequencing and microscopy, are brought together in a unified model. "People get very excited about getting the crystal-like structure of the genome, but we need to assess the structure-function relationship," Mozziconacci says. "I don't think we've seen many insights on the function side yet. That's still to be discovered."

Ensemble Methods

If structural heterogeneity in the genome reflects functional variations among cells, consensus approaches might not provide the full picture, Alber says. "It's unlikely that the genome falls into a single optimum structure."

So he and his colleagues use large Hi-C datasets to generate a range of possible 3-D genome structures. "We deconvolute the Hi-C data into a population of individual structures that, as a whole, are statistically consistent with the data." The aim is to figure out which contacts are most likely to co-occur. Simply embedding the data in 3-D limits the interaction among two regions. Alber also considers the cooperativity principle—if two regions are interacting then perhaps neighboring interactions are more likely.

The frequency of each contact is then accurately reproduced in an ensemble. So if we infer from Hi-C experiments that A contacts B in 15 percent of cells, A will contact B in 15 percent of the ensemble. "This is an approximation of the true population," Alber says. "We don't know what the true population is, and the data are incomplete, but we integrate additional information to get a better approximation."

Once there's an ensemble of tens of thousands of structures, there remains the question of what biology you learn from it. There's a need for new structural biology tools that can mine the structures in the population to find patterns of co-occurrence (when A contacts B does it also tend to contact F?) and relate them to function, Alber says.

From Fiction to Reality: Single-cell Hi-C

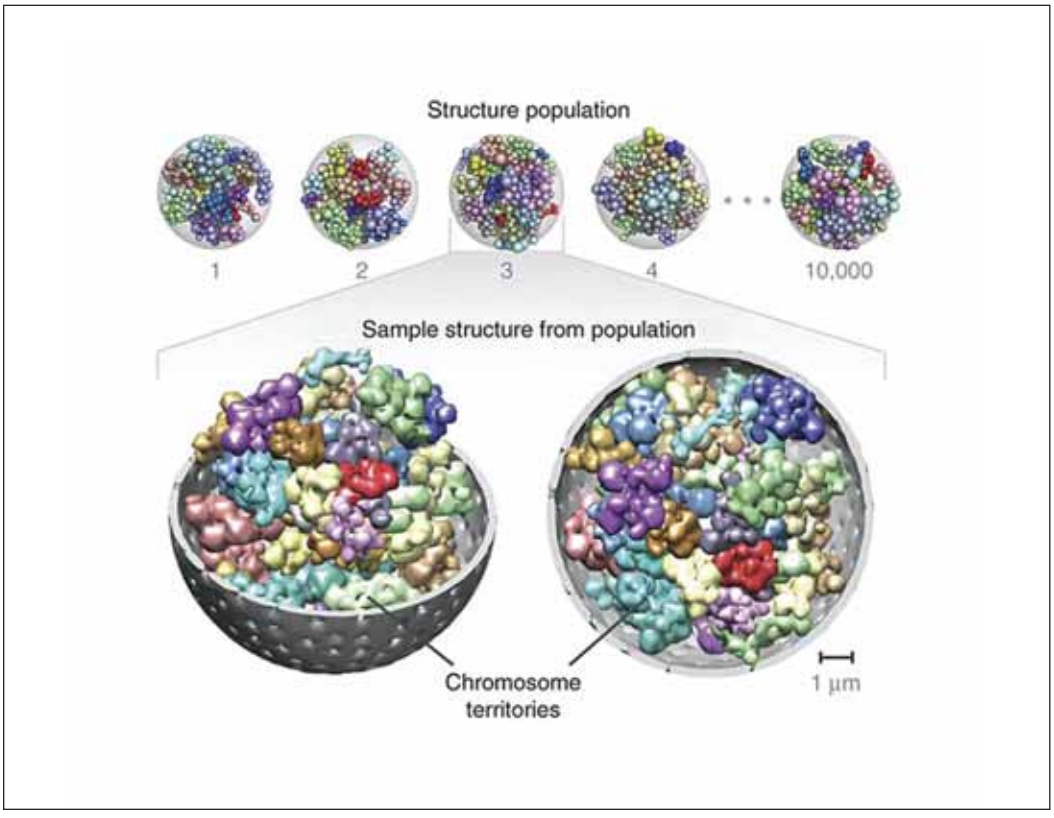
The consensus approach tends to average out the real differences among heterogeneous structures, Alber says. Take the situation where A contacts B half of the time and A contacts C half of the time. In fact, in half the structures these pairs are actually interacting, but the consensus will put these both at a specific distance from each other based on frequency.

But the ensemble approach may suffer from a different kind of unreliability, says **Mark Segal, PhD**, professor of biostatistics at the University of California, San Francisco. The ensemble may not correspond to actual variation in a cell population, he says. "It might, but it's unfounded. It could all be algorithmic as opposed to correlating with anything biological."

One thing all agree on: Single-cell assays have the potential to be more informative. "That's what's coming next," Segal says.

One thing all agree on: Single-cell assays have the potential to be more informative. "That's what's coming next," Segal says.

Single-cell Hi-C is done, as the name suggests, on a single cell. It still suffers from the same low efficiency that troubles Hi-C generally—it may only identify 1000 loci in any one cell. But if done on hundreds of thousands of cells, it could produce an ensemble that could then be linked computationally to current ensembles—lending them a basis in reality. □



Alber's ensemble approach produces a population of more than 10,000 genome structures. A schematic view of the calculated structure population is shown on top. A randomly selected sample from the population is magnified at the bottom. All 46 chromosome territories are shown. Homologous pairs share the same color. The nuclear envelope is displayed in gray. For visualization purposes, the spheres are blurred in the magnified structure because the use of 2 × 428 spheres to represent the genome makes the territories appear more discrete than they actually are. Reprinted with permission from R Kalhor, H Tjong, N Jayathilaka, F Alber, and L Chen, Genome architectures revealed by tethered chromosome conformation capture and population-based modeling,* Nature Biotechnology *30(90–98) (2012).