GENOMICS

Genomes in 3D improve one-dimensional assemblies

Chromosome conformation capture data provide scaffolds for *de novo* genome assemblies.

It is the story of the Ugly Duckling for scientists; data that initially had been discarded turn out to be very useful.

When Job Dekker of the University of Massachusetts in Worcester and his team first developed 'Hi-C' in 2009, a method to probe all genome-wide interactions in three dimensions (3D), they came across a phenomenon that at the time seemed annoying. "When two loci are close to each other in the linear sequence, they contact each other more frequently," explains Dekker. "The signal is very, very strong, and you have to normalize it out of the data to find interesting interactions." But computational biologist Noam Kaplan, upon joining the laboratory as a postdoc, saw the discarded Hi-C data from a different perspective. "If we see things that are interacting frequently in 3D, we know that they must be close in the one-dimensional sequence," says Kaplan. And he realized that this knowledge could be a boon for genome assemblies that are still very fragmented when derived only from high-throughput sequencing data.

Independently, Jay Shendure of the University of Washington in Seattle, together with his graduate student Joshua Burton, also discussed ways to use Hi-C data for better genome assemblies. Shendure's group is part of an effort to develop a $1,000 genome, and their focus was on increasing contiguity, the length of assemblies without gaps. "We can easily generate 100 times as much sequencing data as the entire Human Genome Project," says Shendure, "but the best assemblers in the world can't get anything close to the quality of the original assembly." Top computational tools can assemble short reads into 40-kilobase 'contigs' but cannot bridge larger gaps to place those contigs with respect to one another on chromosomes. The Human Genome Projects had physical and genetic maps that helped place sequence, but these maps are labor-intensive to make and their production is not scalable.

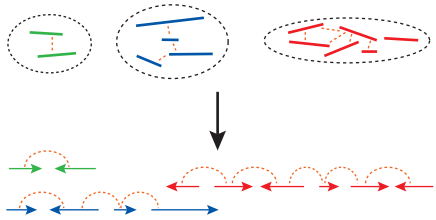
Both groups realized that Hi-C provided the data to put contigs in the right order and construct scaffolds. "The idea has been percolating for a while," says Shendure, "but the challenge is in the algorithm."

Researchers in the two labs tackled this challenge differently. Kaplan and Dekker employed a two-tier approach, first using the higher interaction frequency between loci on the same chromosome to place contigs on chromosomes and then using a probabilistic model to predict the genomic locus along the chromosome based on interaction frequency and genomic distance. This worked well for *de novo* assembly of the human genome, and the researchers also adapted it to predict the locations of previously unplaced fragments of the human genome. The approach only required a library of paired-end short inserts and Hi-C data.

Shendure and his team, on the other hand, used libraries of paired-end short reads, a library of 3-kilobase mate pairs and Hi-C data for their algorithm, named Lachesis, in reference to one of the Greek Fates. In a tiered approach, they created high-quality *de novo* assemblies of the human, mouse and fly genomes. Unlike the algorithm by the Dekker group, Lachesis cannot infer the chromosome numbers of an organism, but it can orient contigs the right way after placement.

Looking forward, both Dekker and Shendure see the need for integrated rather than step-wise data analysis. "An approach that simultaneously takes into account all data types in a single step is likely to do better," says Shendure.

Additional improvements will also come from the experimental side. Current Hi-C data show cell type–specific interactions of genomic loci, which Hi-C originally had been designed to discover, that can mask the signal used for scaffolding. Dekker and his team recently solved the structure of the metaphase chromosome, and Kaplan suggests that such metaphase Hi-C data will get rid of cell type–specific interactions.



Hi-C data help find the right genomic position of short sequence reads. Adapted from *Nature Biotechnology* (Burton et al., 2013).

Researchers in Shendure's lab will focus on making the Hi-C protocol robust for tissues from diverse organisms, as current data are all derived from cell lines.

Other recent work has shown that Hi-C data can be used to reconstruct haplotypes; Kaplan sees the combination of genome assembly and haplotype phasing as an exciting possibility.

Neither the Dekker nor the Shendure teams have a track record in genome assembly, so both groups are eager for assembly experts to try out their algorithms and suggest further improvements. Recent community-driven comparisons have made clear that there is not one program that outperforms all others, but the algorithms from the Dekker and Shendure labs will provide an important starting point for bridging large gaps in the assemblies.

Nicole Rusk

RESEARCH PAPERS

Burton, J.N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).

Kaplan, N. & Dekker, J. High-throughput genome scaffolding from *in vivo* DNA interaction frequency. *Nat. Biotechnol.* **31**, 1143–1147 (2013).