May 20th, 12:30 PM

High-throughput Genome Scaffolding from in vivo DNA Interaction Frequency

Noam Kaplan  
*University of Massachusetts Medical School*

Job Dekker  
*University of Massachusetts Medical School*

Follow this and additional works at: https://escholarship.umassmed.edu/cts_retreat

Part of the [Computational Biology Commons](https://escholarship.umassmed.edu/cts_retreat), [Genomics Commons](https://escholarship.umassmed.edu/cts_retreat), [Systems Biology Commons](https://escholarship.umassmed.edu/cts_retreat), and the [Translational Medical Research Commons](https://escholarship.umassmed.edu/cts_retreat)

This work is licensed under a [Creative Commons Attribution-Noncommercial-Share Alike 3.0 License](https://creativecommons.org/licenses/by-nc-sa/3.0/).
High-throughput genome scaffolding from in vivo DNA interaction frequency

Noam Kaplan* and Job Dekker

Program in Systems Biology, Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA, 01605-0103, USA

E-mail: noam.kaplan2@gmail.com

Despite the advancement of DNA-sequencing technology, assembly of complex genomes remains a major challenge. Surprisingly, the quality of published assemblies of complex genomes has in fact decreased, due to the discrepancy between the rate of advancement of short read sequencing and that of scaffolding technology. Short read sequencing typically yields accurate, but disconnected, contigs. However, high-throughput scaffolding of contigs into chromosomes, based on long-insert paired-end read libraries, is a difficult task and yields highly fragmented genomes. Further scaffolding, which is required to improve the degree of completion of genome sequences, typically relies on laborious or low-throughput methods.

We have developed a novel sequencing-based high-throughput approach to genome assembly, based on the notion that loci that are near each other in the genomic sequence have a high probability of interacting with each other. Using probabilistic models, we demonstrate that genome-wide in vivo chromatin interaction frequency measurements, easily measurable with 3C-based experiments, can be used as genomic distance proxies to accurately detect the position of individual contigs over large distances without requiring any sequence overlap. Furthermore, we demonstrate our approach can karyotype and scaffold an entire genome de novo. Applying our approach to incomplete regions of the human genome, we predict the positions of 65 previously unplaced contigs, in agreement with alternative methods in 26/31 cases attempted in common. Our approach can theoretically bridge any gap size, is simple, robust, scalable and applicable to any species.

High-throughput genome scaffolding from in vivo DNA interaction frequency
Nature Biotechnology, 31:1143-1147
http://www.nature.com/nbt/journal/v31/n12/full/nbt.2768.html