

Synteny-Assisted Assembly of Genomes at Low Coverage

Sante Gnerre,^{1,2} Pablo Alvarez,¹ Will Brockman,¹
Jonathan Butler,¹ CheeWhye Chin,¹ Manfred Grabherr,¹
Michael Kleber,¹ Evan Mauceli,¹ Jean Chang,¹ Michele Clamp,¹
Jill Mesirov,¹ Kerstin Lindblad-Toh,¹ Eric S. Lander,¹
David B. Jaffe¹

Keywords: assisted assembly, low-coverage, synteny

Introduction

We present a new methodology for the assembly of genomes at low coverage that uses syntenic alignments to augment an existing assembly.

This approach was motivated by the fact that, as part of an NHGRI initiative to annotate the human genome, we are currently sequencing mammals at $2\times$ coverage. All mammalian genomes will be aligned to the human genome and used to identify conserved features such as genes and regulatory elements. To date we have sequenced the African savannah elephant (*Loxodonta Africana*), the Nine-banded armadillo (*Dasyopus novemcinctus*), and an inbred New Zealand White Rabbit (*Oryctolagus cuniculus*), all females, using a mix of 4 kb plasmid and 40 kb Fosmid libraries.

To facilitate the alignment we are assembling each genome individually prior to alignment. As the new methodology augments this initial assembly, its quality significantly impacts the success of this strategy.

Methodology

In order to disambiguate repetitive and polymorphic regions, traditional assembly methodologies must distinguish between genuine and artifactual read-read alignments and linking information. At low coverage, the ability to make this distinction is greatly compromised, resulting in a highly fragmented assembly which places only a fraction of the reads.

By examining the alignments of the reads of the low coverage genome to high quality assemblies of phylogenetically comparable genomes, we can confirm links (as in Figure 1) and read-read alignments (as in Figure 2) that would otherwise be unsubstantiated and therefore unusable.

¹Broad Institute at MIT and Harvard, 320 Charles Street, Cambridge, MA, 02141.

²E-mail: sante@broad.mit.edu

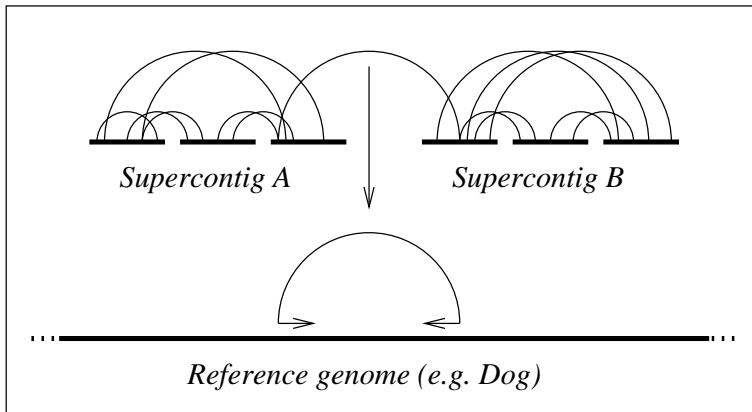


Figure 1: Confirm a link by alignment onto a related genome.

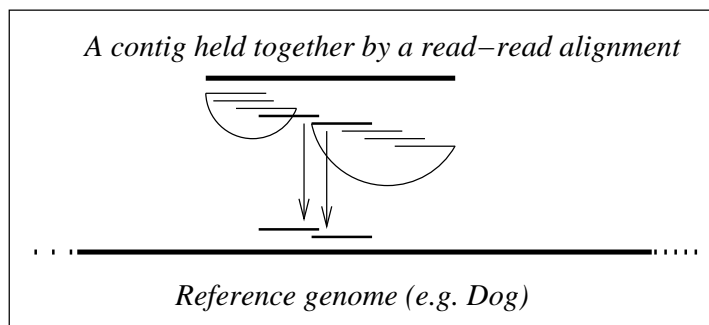


Figure 2: Confirm a read-read alignment by alignment onto a related genome.

This confirmed information is then integrated into an assembly generated with ARACHNE [1, 2], improving the quality and accuracy of its consensus, its coverage of the genome, and its overall connectivity. In the case of elephant at $2\times$, for example, we found that 48% of the reads could be uniquely placed on the human genome, 45% could be uniquely placed on the dog genome, and in total 57% could be uniquely placed on at least one of the two genomes. After the integration, read usage went from 70.5% to 81.3%; total contig length went from 2.0 Gb to 2.3 Gb; and N50 (ungapped) supercontig length went from 15.5 Kb to 25 Kb.

References

- [1] Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E. S. 2002. Arachne: A whole-genome shotgun assembler. *Genome Res.* **12**: 177–189.
- [2] Jaffe, D. B., Butler, J., Gnerre, S., Mauceli, E., Lindbad-Toh, K., Mesirov, J. P., Zody, M., and Lander, E. S. 2003. Whole-Genome Sequence Assembly for Mammalian Genomes: Arachne 2. *Genome Res.* **13**: 91–96.