

Repeat Analysis to Improve Whole Genome Assembly

Manfred Grabherr^{1,2}, Pablo Alvarez¹, Will Brockman¹,
Jonathan Butler¹, CheeWhye Chin¹, Sante Gnerre¹,
Michael Kleber¹, Evan Mauceli¹, Jill Mesirov¹, Bruce Birren¹,
Kerstin Lindblad-Toh¹, Chad Nusbaum¹, Eric S. Lander¹,
David B. Jaffe¹

Keywords: whole genome assembly, repeat regions, kmer

1 Introduction.

Recent improvements in ARACHNE have led to the improved ability to assemble consensus for repetitive regions. However, this poses two problems. First, the placement of reads falling entirely within these regions is inherently ambiguous, which results in misassemblies on a scaffold level. Second, contigs will be misassembled early on, since walking through a sufficiently long repeat results in multiple possibilities for how to continue on the other end. Kmer-based repeat analysis can determine possible break points in the assembly as well as devise strategies to move reads in order to improve overall consistency.

2 Methodology.

A sorted table of all 48-mers in the assembly is generated, which allows for quick lookup of the positions of multiply used kmers. Kmers that are contiguously shared between different stretches of the assembly are collapsed into intervals; these are the repeat sequences for which there are multiple exact copies in the genome. We build a graph in which each node represents a scaffold and the edges are labeled by intervals in which sequence is repeated.

By walking from node to node, the structure of repeats can be identified. If there are many links to different nodes (or links to the same node, but at different locations), this indicates dispersed repeat sequence; in this case, repeat regions can be grouped according to the connectivity of the graph, allowing for categorizing repeat “families”. If, however, there are regions in which a node corresponds to only one other node, and the intervals are distributed consistently throughout the scaffold, then this points to either a segmental duplication, or – if that can be ruled out – an artifact of the assembly process, in which the two haplotypes of a polymorphic genome have been assembled separately.

To assist the assembly process, we can now easily – and quickly – assess whether a read falls entirely into stretches that have a copy somewhere else in the genome. We subsequently avoid using those and their partners when building scaffolds. For reads that are anchored in unique sequence (which can still be made up of repetitive kmers; *unique* now means that there does not exist a complete interval around the read’s location that is duplicated) and whose partner has been

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

² E-mail: grabherr@broad.mit.edu

placed elsewhere due to repetition, the partner can be moved to its proper place to strengthen the existing scaffold structure. Moreover, regions without consistent (long) insert coverage can be tagged as broken if a.) there is repetitive sequence, b.) based on the region surrounding the repeat there should be insert coverage, and c.) moving reads does not provide this insert coverage.

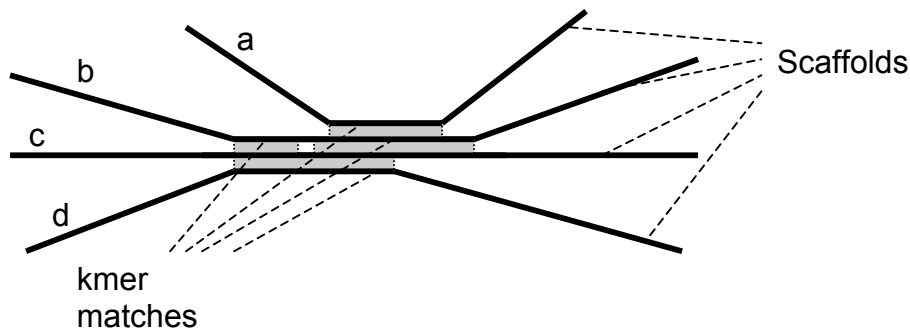


Figure 1: Perfect kmer matches between different scaffolds in the assembly.

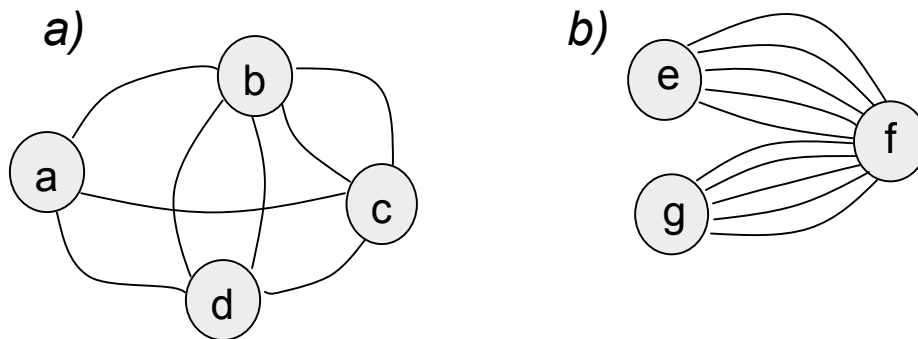


Figure 2: a) graph connecting scaffolds a, b, c and d, where the overlap intervals are stored in the edges according to fig. 1. b) graph indicating that two haplotypes were assembled apart in three scaffolds e, f and g.

3 References and bibliography.

[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.