Improving Genome Assembly without Sequencing

Michael C. Schatz[*], Arthur L. Delcher[1], Pawel Gajer[1], Jason Miller[1], Martin Shumway[1], Steven L. Salzberg[1]

Keywords: genomic finishing, sequencing gap closure, base-calling

1 Introduction

Assembly of genomes from whole-genome sequencing (WGS) projects is one of the most complex computational problems in genomics. WGS assemblers such as Arachne [1] and Celera Assembler [2] are able to process data from millions of individual sequence "reads" and construct an accurate representation of a genome. These assemblies are in the form of contigs (contiguous stretches of DNA sequence) and scaffolds (contigs linked by information from paired-end sequences).

Our group has developed a new system, AutoJoiner, which can refine an assembly by connecting together contigs within scaffolds. Its primary strategy is to find reads whose 3' end points into a gap, and to extend those reads until they close the gap. Gaps can be closed either by finding reads that span them from one direction or by finding reads on both sides that meet somewhere in the middle. This strategy is effective because the WGS assemblers must trim the reads to a relatively high threshold of accuracy in order to handle the large scale problem of whole-genome assembly. At the level of an individual gap between two contigs, however, these criteria can be relaxed, and we find that gaps up to several hundred base pairs in length can be spanned by extended reads. At the same time, AutoJoiner must take care not to overcollapse tandem repeats in the process of joining adjacent contigs.

We have also developed a second-generation basecaller, AutoEditor, that re-calls bases from the original (raw) chromatogram data based on the multiple alignment produced by an assembler. AutoEditor, which has been published previously, corrects approximately 80% of sequencing errors with a false-correction rate of fewer than 1 in 8800 corrections [3]. AutoEditor was run on all assemblies in order to improve their fidelity before passing them on to AutoJoiner.

2 Results

To evaluate the effectiveness of AutoJoiner, we ran it on 33 genomes and chromosomes (17 bacterial genomes, the 14 separately assembled chromosomes of the eukaryote *Cryptococcus neoformans*, and two Drosophila species, *D. yakuba* and *D. virilis*). All of these assemblies except for the Drosophilas had been completely closed at TIGR, giving us a highly accurate standard for comparison. For the evaluation, we assembled the original shotgun reads using the Celera Assembler with a standard set of parameters.

[*] To whom correspondence should be addressed. The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD, North America. Email: mschatz@tigr.org

[1] The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD, North America

AutoJoiner resolved just over 26% of the intra-scaffold gaps across all projects (395 gaps out of 1490 total). Only 3 of the joins were invalid (0.76%), with no invalid joins for bacterial genomes. The success of the join was determined by aligning the joined region to the finished genome. The average identity to the finished sequence was 99.45%. The range of gap sizes that AutoJoiner closed was from -1069.5 bp to 249 bp, with a mean of -26 bp and a standard deviation of 182 bp. Negative gap sizes indicate that there was an overlap between the contigs prior to any contig extension. For the draft assemblies of D. *yakuba* and D. *virilis*, both produced by Celera Assembler, AutoJoiner closed 537/5507 (9.9%) and 1102/8277 (13.3%) of the gaps respectively.

References

[1] Batzoglou, S., Jaffe, D.B., Stanley, K., et al. 2002. ARACHNE: A whole-genome shotgun assembler. *Genome Research* 12:1:177-189.

[2] Gajer, P., Schatz, M., and Salzberg, S.L. 2004. Automated correction of genome sequence errors. *Nucleic Acids Research* 32:2:562-569.

[3] Myers, E.W., Sutton, G.G., Delcher, A.L., et al. 2000. A whole-genome assembly of Drosophila. *Science* 287: 2196-2204.