RAILS and Cobbler: Scaffolding and automated finishing of draft genomes using long DNA sequences

Rene L Warren[1]

[1]BC Cancer Agency, Genome Sciences Centre, Vancouver, BC, Canada

10 November 2016

Paper DOI: http://dx.doi.org/10.21105/joss.00116
Software Repository: https://github.com/warrenlr/RAILS
Software Archive: https://dx.doi.org/10.5281/zenodo.167092

Summary

Despite major advances in DNA sequencing technologies we do not yet have complete genome sequences. Producing high-quality, contiguous, draft assemblies *de novo* is of paramount importance as it informs on genetic content and organization of the genome (Pagani et al. 2012). The past decade has seen improvements in sequence throughput, a substantially lower DNA sequencing cost and increased read lengths. Whereas the base accuracy of short (currently ~250 bp) read lengths such as those from Illumina have improved (>99%), the base accuracy of long sequence read platforms (Pacific Biosciences, Oxford Nanopore) remains low for generating reference-grade genome assemblies without read error correction. Gap-filling tools designed to help finish draft genomes in an automated fashion, which includes our own (Paulino et al. 2015), have been recently developed (Tsai, Otto, and Berriman 2010, Boetzer and Pirovano (2012)). They are typically designed to work with short sequencing reads, not high-quality long sequences from other draft assemblies. In many such projects that employ short sequence reads for *de novo* assembly, a k-mer graph assembly approach is often favored, as it effectively discards errors and spurious sequences, albeit at the cost of long-range information loss and limited ability to resolve long repeats. However, researchers routinely produce various assembly drafts varying the parameter k length in search of the most contiguous assembly. This multitude of assembly drafts is comprised of sequences with untapped potential, representing a wealth of information for gap-filling and scaffolding. Here, I make available two bioinformatics software tools, Cobbler and RAILS (Rene L Warren 2016) to exploit this information for automated finishing and scaffolding with long DNA sequences, respectively. They can be used to scaffold & finish high-quality draft genome assemblies with any long, preferably high-quality, sequences such as scaftigs/contigs from another genome draft. They both rely on accurate, long DNA sequences to patch gaps in existing genome assembly drafts. More specifically, Cobbler is a utility to automatically patch gaps (ambiguous regions in a draft assembly, represented by N's). It does so by first aligning the long sequences to the assembly, tallying the alignments and replacing N's with the sequences from these long DNA sequences. RAILS is an all-in-one scaffolder and gap-filler. Its process is similar to that of Cobbler. It scaffolds a given genome draft with the help of long DNA sequences (contig sequences are ordered/oriented using alignment information) using the scaffolding engine I originally developed for SSAKE (René L. Warren et al. 2007) and LINKS (Warren et al. 2015). The newly created gaps are automatically filled with the DNA string of the provided long DNA sequences. In a simulated long sequences experiment (1, 2.5, 5, 15 kbp sequences) designed from the human genome reference, Cobbler closed >65% of gaps in a human genome assembly draft (Table 1; test provided with the distribution, correlation of close gaps with length estimates from draft assembly R=0.8253). Using the same sequence data, RAILS further scaffolded that same baseline assembly from (N50 length) 5.6 to 7.3 Mbp, representing a 30% increase in contiguity (Table 2). RAILS and Cobbler are implemented in PERL and run on any systems where PERL is installed.

Table 1. Patching gaps in a genome assembly draft with Cobbler, using simulated 1, 2.5, 5 and 15 kbp simulated long sequences from human genome reference GRCh38.

Metric	Value
Total gaps	148,091
Number of gaps patched	95,523
Proportion of gaps patched	65.1%
Average length (bp)	343.39
Length st.dev +/-	931.12
Total bases added	32,801,755
Largest gap resolved (bp)	13,662
Shortest gap resolved (bp)	1

Table 2. Assembly statistics on human genome scaffolding and finishing post Cobbler and RAILS (reporting sequences 500 bp and larger).

Stage	n:500	n:N50	n:NG50	NG50 (bp)	N50 (bp)	max (bp)	sum (bp)
Baseline	65,905	145	164	5,144,025	5,597,244	26.41e6	2.794e9
Cobbler	65,905	145	161	5,312,196	5,658,133	26.66e6	2.827e9
RAILS	64,210	113	125	6,935,685	7,266,542	32.14e6	2.836e9

References

Boetzer, Marten, and Walter Pirovano. 2012. “Toward Almost Closed Genomes with Gapfiller.” *Genome Biology* 13 (6): R56. doi:10.1186/gb-2012-13-6-r56.

Pagani, I., K. Liolios, J. Jansson, I. -. M. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides. 2012. “The Genomes Online Database (Gold) V. 4: Status of Genomic and Metagenomic Projects and Their Associated Metadata.” *Nucleic Acids Res* 40. doi:10.1093/nar/gkr1100.

Paulino, Daniel, René L. Warren, Benjamin P. Vandervalk, Anthony Raymond, Shaun D. Jackman, and Inanç Birol. 2015. “Sealer: A Scalable Gap-Closing Application for Finishing Draft Genomes.” *BMC Bioinformatics* 16 (1): 230. doi:10.1186/s12859-015-0663-4.

Tsai, Isheng J., Thomas D. Otto, and Matthew Berriman. 2010. “Improving Draft Assemblies by Iterative Mapping and Assembly of Short Reads to Eliminate Gaps.” *Genome Biology* 11 (4): R41. doi:10.1186/gb-2010-11-4-r41.

Warren, Rene L. 2016. “RAILS and Cobbler: Scaffolding and Automated Finishing of Draft Genomes Using Long Sequences.” https://github.com/warrenlr/RAILS.

Warren, René L., Granger G. Sutton, Steven J. M. Jones, and Robert A. Holt. 2007. “Assembling Millions of Short Dna Sequences Using Ssake.” *Bioinformatics* 23 (4): 500–501. doi:10.1093/bioinformatics/btl629.

Warren, René L., Chen Yang, Benjamin P. Vandervalk, Bahar Behsaz, Albert Lagman, Steven J. M. Jones, and Inanç Birol. 2015. “LINKS: Scalable, Alignment-Free Scaffolding of Draft Genomes with Long Reads.” *GigaScience* 4 (1): 35. doi:10.1186/s13742-015-0076-3.