A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data

David Coil, Guillaume Jospin, and Aaron E. Darling

January 22, 2014

Abstract

1 Motivation:
Open-source bacterial genome assembly remains inaccessible to many biologists due to its complexity. Few software solutions exist that are capable of automating all steps in the process of de novo genome assembly from Illumina data.

2 Results:
A5-miseq can produce high quality and microbial genome assemblies on a laptop computer without any parameter tuning. A5-miseq does this by automating the process of adapter trimming, quality filtering, error correction, contig and scaffold generation, and detection of misassemblies. Unlike the original A5 pipeline, A5-miseq can use long reads from the Illumina MiSeq, use read pairing information during contig generation, and includes several improvements to read trimming. Together these changes result in substantially improved assemblies that recover a more complete set of reference genes than previous methods.

3 Availability:
A5-miseq is licensed under the GPL open source license. Source code and precompiled binaries for Mac OS X 10.6+ and Linux 2.6.15+ are available from http://code.google.com/p/ngopt

4 Contact:
aaron.darling@uts.edu.au

5 Introduction
Genome assembly involves an entire data processing workflow starting with raw sequence data and ending with scaffolded contigs. The steps often consist of adapter trimming, quality filtering, error correction, creation of contigs, verification of contigs by mapping reads to the assembly, and the creation/verification of scaffolds.

We previously published A5, a pipeline that automated all the steps to generate bacterial genome assemblies from raw Illumina data (Tritt et al., 2012). The workflow included five steps and the parameters for each were optimized on assemblies of Halophilic archaea and tested on E. coli.

Since the publication of A5, Illumina’s chemistry has significantly advanced and the MiSeq instruments are now capable of producing reads in excess of 400nt long, which is fourfold longer than what was previously possible on a HiSeq 2000. The original A5 could not process reads longer than 150nt. The longer reads make it possible to assemble genomes from less data overall, but doing so requires major revisions to the data processing algorithms in A5.

We introduce a revised pipeline called A5-miseq which replaces several components of the original A5 pipeline with new software modules and produces substantially improved assemblies. The A5-miseq workflow is shown in Figure 1 and has several improvements to the read cleaning and contig assembly steps (1 & 2).

---

* Genome Center, University of California Davis, USA
† three institute, University of Technology Sydneey, Australia, to whom correspondence should be addressed: aaron.darling@uts.edu.au
6 Methods

The A5-miseq pipeline consists of five steps: (1) **Read cleaning** – sequence adapters and low quality regions are removed by Trimmomatic [Lohse et al., 2012] followed by by Scythe (Vince Buffalo, unpublished). Errors in the reads are then corrected using SGA’s k-mer based error correction algorithm [Simpson et al., 2011]. (2) **Contig assembly** – paired and unpaired reads are used for assembly with the IDBA-UD algorithm [Peng et al., 2012]. (3) **Crude scaffolding** – contigs are scaffolded with any available large insert libraries using permissive parameters. (4) **Misassembly correction** – misassemblies are detected on the basis of read pairs that do not map within the expected distance. Contigs and scaffolds found to contain misassemblies are broken. (5) **Final scaffolding** – a final round of scaffolding with stringent parameters repairs any previously broken contigs.

A5-miseq substantially revises steps (1) and (2) relative to A5. In step (1) A5 would discard entire reads found to contain any amount of adapter readthrough. The standard Nextera XT protocol results in libraries where a large fraction of the reads (>50% in extreme cases) contain adapter readthrough when sequenced with the currently standard paired-end 300nt read chemistry. Instead of discarding such reads, A5-miseq trims out only the contaminated portion of the read. A5-miseq applies two adapter trimming algorithms in succession because we found that neither method alone is able to eliminate all adapter contamination from the assembled sequence. This could be due to errors in the reads which prevent detection of the adapter sequence but which are corrected later in the assembly process.

In step (2) A5-miseq employs a contig assembly algorithm (implemented in the IDBA-UD software) that exploits read pairing information during contig generation. This major improvement reduces the frequency with which misassembled contigs are formed during the contig generation step. Thus fewer misassemblies must be detected and corrected in step (4) of the pipeline. IDBA-UD required extensive source code revision to operate on Mac OS X and reduce its memory usage for laptop hardware. These changes are available in the A5-miseq source code repository.

We benchmarked A5-miseq and A5 on the raw GAGE-B MiSeq datasets. GAGE-B includes paired-end 250nt MiSeq reads at 100x coverage for four organisms. We ran A5-miseq and A5 assemblies and obtained results for other assemblers from the GAGE-B publication [Magoc et al., 2013]. The original A5 can not assemble 250nt reads, so we used Trimmomatic to cut the reads down to 150nt (discarding 100nt) prior to assembly. Running time and peak memory were measured with `/usr/bin/time -v` on a 2012 MacBook Air running Ubuntu 13.04.

To evaluate assembly accuracy we ran QUAST v2.2 [Gurevich et al., 2013] with the following parameters: `quast.py --gage -u -G ref.gff -R ref.fa ass.fa`, where `ref.fa` is the reference assembly from NCBI, `ref.gff` are the reference’s annotated genes from NCBI, and `ass.fa` is the assembly. The `-u` option causes the genome fraction to be calculated in the manner used by QUAST v1.3. This yields results that are comparable to those in the GAGE-B paper.

7 Discussion

The benchmarking results for A5-miseq on the GAGE-B data (shown in Table 1) indicate that it offers substantial improvements over the original A5 pipeline. A5-miseq assemblies contain more full length reference genes than A5 or any of the GAGE-B assemblers. Relative to the original A5, the number of genes missing from A5-miseq assemblies is reduced by 1.6 to 3.8-fold on the GAGE-B datasets. In most cases, A5-miseq assemblies have higher NGA50 values, fewer misassemblies, and fewer base calling errors than the original A5 pipeline. A5-miseq produces results that are competitive with the best
Table 1: Comparison of assembly accuracy between A5-miseq and other assemblers.

| Organism     | Size | A5-miseq | A5 | GAGE-B |
|--------------|------|----------|----|--------|
|              | Frac | NGA50    | MA | MM     | Genes |
| B. cereus    | 99.9 | 486.3    |  6 | 1.6    | 5733  |
| R. sphaeroides| 99.9 | 130.5    |  6 | 3.4    | 4426  |
| M. abscessus | 99.4 | 225.6    | 11 | 3.2    | 4919  |
| V. cholerae  | 99.6 | 196.7    | 15 | 6      | 3645  |

Assembly accuracy for the A5-miseq and A5 pipelines measured on raw 100x coverage MiSeq PE250 GAGE-B data. Accuracy was measured by QUAST. Frac is the fraction of the reference genome represented in assembly scaffolds. NGA50: the N50 after breaking contigs at misassemblies. MA: number of misassemblies, MM: number of nucleotide mismatches per 100 kbp, Genes: number of intact full-length genes recovered. For GAGE-B results, the single best assembly result for each metric and genome reported in Magoc et al. (2013) is shown, with the assembler producing the best result indicated as follows: S=SPAdes 2.3, sdn=SOAPdenovo, M=MaSuRCA, A=ABySS, sga=SGA.

8 Conclusion

Genome assembly is a quickly evolving field and software has been advancing rapidly. Although A5-miseq produces assemblies that are competitive with results in a recently published assembler evaluation (Magoc et al., 2013), it is likely that versions of other software that have not yet been published in a peer-reviewed journal (notably SPAdes 3.0) might produce even better results. SPAdes 2.3 automates many of the same steps that A5-miseq automates (Bankevich et al., 2012) and in general produces excellent assemblies (Table 1). If possible, researchers interested in genome assembly should become acquainted with the various algorithms available before selecting a particular approach. A5-miseq should be particularly useful for researchers with limited bioinformatics experience or computing resources.

Acknowledgement

We would like to thank the many users of the A5 pipeline for reporting bugs and other usability problems with the software, in particular Piklu Roy Chowdhury.

Funding: This work was supported in part by a collaborative agreement with the NSW Department of Primary Industries.

References

Tritt, A., Eisen, J.A., Facciotti, M.F., and Darling, A.E. (2012) An Integrated Pipeline for de Novo Assembly of Microbial Genomes, *PLoS One*, 7(9): e42304.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. (2013) QUAST: quality assessment tool for genome assemblies, *Bioinformatics*, 29(8): 1072-5.

Magoc T. *et al.* (2013) GAGE-B: An Evaluation of Genome Assemblers for Bacterial Organisms. *Bioinformatics*, 29(14), 1718-25.

Bankevich, A *et al* (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput. Biol.*, 19(5):455-77.

Jared T Simpson and Richard Durbin (2012) Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.*, 22(3):549-56.

Peng, Y, Leung, HC, Yiu, SM, Chin, FY (2003) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*,28(11):1420-8.

Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B. (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, 40:W622-7.