Metassembler: A pipeline for improving *de novo* genome assembly

Paul Baranay, Scott Emrich, Michael Schatz

Department of Biological Sciences and Department of Computer Science and Engineering, University of Notre Dame
Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory

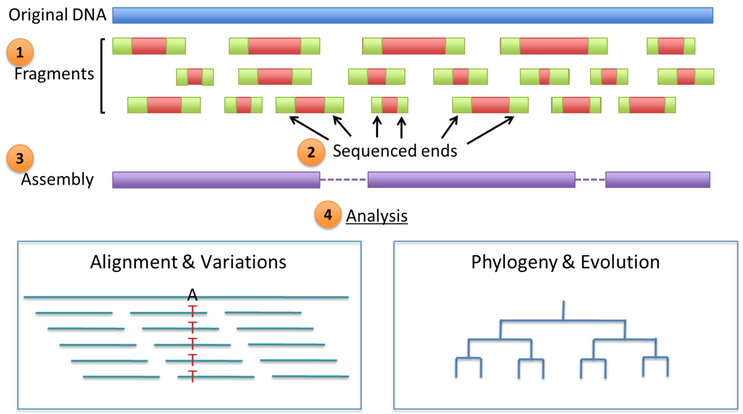
pbaranay@nd.edu · http://metassembler.sf.net

Summary

Sequencing projects typically create several draft assemblies of the genome under consideration, either by employing several different assemblers or by incrementally improving the input parameters of a single assembler. Usually, a single draft assembly is selected as the candidate for publication. Instead of discarding the extra assemblies, we propose using them in the process of "metassembly," which combines information from several input assemblies into a single output assembly that is superior to either of its substituents.

Genome assembly

Genome assembly is the process of determining an organism's DNA sequence from a library of sequenced reads. Frequently, next-generation sequence methods are employed to create libraries of millions of very short reads, about 100 base pairs in length.



Schematic representation of a next-generation assembly project

While this method of sequencing is very cost effective, assembling these short reads into a representation of the original genome is computationally challenging. Several assembly programs have already been created to work with short reads, employing a variety of specialized techniques such as error correction, read doubling, and clever graph construction methods.

Mathematical justification

The CE statistic or compression-expansion statistic can be used to identify regions of a genome that are likely to be misassembled. By mapping the library of reads to the genome, we can compute the mean M of the implied insert lengths l_i. Given local coverage N and an expected mean insert size of μ with standard deviation $\sqrt{\sigma/N}$, the CE statistic Z identifies separation between mate-pairs at a given region:

$$M = \frac{1}{N}\sum_{i=1}^{N} l_i \qquad Z = \frac{M-\mu}{\sigma/\sqrt{N}}$$

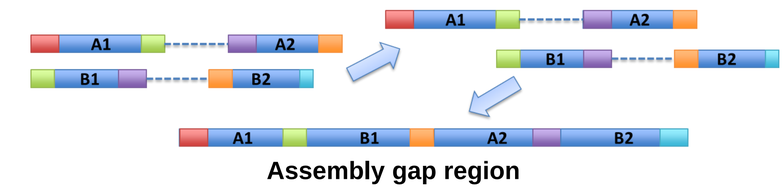
Large positive Z values (higher separation) indicate expansion errors caused by erroneously inserting sequence, whereas negative Z values indicate compression errors caused by erroneously deleting sequence. Approximating Z by the normal distribution and with the threshold for misassembly set to a cutoff of Z_0 = ±3.0, we can detect misassemblies that have only a 0.2% chance of occurring randomly.

Metassembler: assembly reconciliation

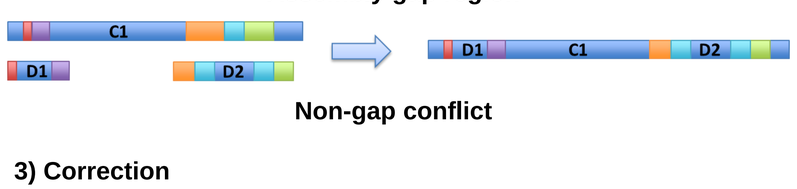
Metassembler is a pipeline for combining two existing assemblies into a single, more accurate assembly.

1) Assembly generation
Metassembler can work with any pair of assemblies, whether generated by the end user or an intermediate party. One assembly is designated as the reference or primary assmebly, and the other is designated as the supplementary assembly. Metassembler works best when it has access to the initial libraries of read data used to create the reference assembly so it can compute the CE statistic.

2) Identify candidate compressions
Metassembler employs Nucmer to align the two assemblies together and identify regions of similarity. These alignments are filtered to ensure a 1-to-1 match of regions in each assembly. Potential conflicts between the two assemblies are identified forensically and the CE statistic for each conflict is computed.



Assembly gap region



Non-gap conflict

3) Correction
Conflicts can consist of gaps in one assembly that are spanned by the other, or of non-gap regions where one assembly has a more reasonable CE statistic value than the other. Gaps were closed if the CE statistic fell within the range from -3 to 3. Non-gap regions were used to patch the assembly if the region's CE statistic would improve by at least 3 standard errors when patched.

4) Cleanup
To ensure that no sequence is "lost", large contigs from each assembly that were not corrected are added to the final assembly in their original form.

Acknowledgments

- Cold Spring Harbor Laboratory Undergraduate Research Program
- Zachary Lippman and Keisha John (CSHL)
- Dominic Chaloner and Thomas Burish (Notre Dame)
- Matt Titmus, Mitch Bekritsky, and Hayan Lee (CSHL Schatz Lab)

References

- Gnerre, S., et al. "High-quality draft assemblies of mammalian genomes from massively parallel sequence data." PNAS 2011.
- Kelley, D.R.; Schatz, M.C.; Salzberg, S,L. "Quake: quality-aware detection and correction of sequencing errors." Genome Biology 2010.
- Kurtz, S., et al. "Versatile and open software for comparing large genomes." Genome Biology 2004.
- Li., R., et al. "De novo assembly of human genomes with massively parallel short read sequencing." Genome Research 2010.
- Zimin A., et al. "Assembly reconciliation." Bioinforamtics 2008.

Results

Fish genome (Assemblathon 2)

Metassembly improvements	
Gaps closed	595
Errors corrected	28
Total improvements	623

We created two assemblies using the fish data sequenced by the Broad Institute and provided by the Assemblathon 2 organizers. The data were raw Illumina reads with about 190x total coverage, consisting of several fragment libraries (180bp fragment size) and several sheared jumping libraries (fragment sizes ranging from 2500bp to 9000bp). The overall genome size is estimated as 1 GB.

ALLPATHS: The libraries were used as raw input to the ALLPATHS assembler, which performs its own error correction and requires no other configuration apart from setting the K-mer size (K=96).

SOAPdenovo: For assembly using SOAPdenovo, overlapping reads were first joined using FLASH, essentially doubling the read length. Fragments were error corrected using Quake and assembled into contigs with SOAPdenovo. Mate pairs were then aligned to the contigs, duplicates were removed, and reads were extended to 100bp. Finally, SOAPdenovo was used to create the final scaffolds.

Metassembler: Applying Metassembler to these genomes identified 623 regions as candidates for correction via metassembly. Adjusting these regions resulted in in an appreciable increase in the assembly's overall coverage (span) of the genome. On this data set as well as others presently under consideration, Metassembler substantially improves the quality, contiguity, and accuracy of the genome assembly.

Assembly	Scaff. #	N50 (MB)	Span (MB)
ALLPATHS	2791	3.71	8.44
SOAP	1.89M	0.06	5.46
Metassembler	3188	3.71	8.45

Future Work

Metassembler is broadly applicable to any sequencing project where multiple draft assemblies are created. In particular, we hope to apply Metassembler to the following projects in the near future:

- *Anopheles gambiae*, the mosquito vector of malaria
- Fish and snake genomes from the Assemblathon competition
- Human genome, including tumor cell sequences