# CloG: a pipeline for closing gaps in a draft assembly using short reads

Xing Yang, Daniel Medvin, Giri Narasimhan
Bioinformatics Research Group (BioRG)
School of Computing and Information Sciences
Miami, FL 33140, USA.
e-mails: {xyang006,daniel.medvin,giri}@fiu.edu

Deborah Yoder-Himes, Stephen Lory
Microbiology and Molecular Genetics Department
Harvard Medical School, Boston, MA 02115, USA
e-mails: {Deborah_Yoder-Himes,
Stephen_Lory}@hms.harvard.edu

*Abstract* — **Closing of gaps in draft assemblies using Next Generation Sequencing (NGS) data is becoming increasingly important. In this paper we present CloG, a software pipeline that uses NGS data to close gaps in draft assemblies. Firstly, CloG uses the VELVET assembler to generate a hybrid assembly from a mixture of reads: short reads from the NGS data and the original draft assembly (treated as long reads). It then closes gaps between adjacent contigs by reconciling (i.e., "stitching") the two assemblies. By exploiting the strengths of both hybrid assembly and stitching reassembly, CloG is able to outperform its contemporaries in closing gaps in the draft assembly of the bacterium *Burkholderia dolosa*.**

***Keywords-sequencing, assembly, gaps, contigs, repeat regions, finishing, seeds***

## I. Introduction and Motivation

Elucidation of the complete genome sequence of a variety of organisms is the foundation of current genomic research. In addition, the evolution of sequencing technology has been a major driving force for progress in the field. The first generation of sequencing technology was based on the chain-termination method developed by Fred Sanger [2]. Fully automated high throughput implementations of this method utilizing parallel capillary electrophoresis have been developed. These sequencers represent the state of the art in Sanger sequencing and continue to hold many advantages over newer technologies. One advantage of the Sanger sequencing method is that it is capable of producing reads with lengths of up to 1000 bp [2], thus covering many large repeat regions and producing assemblies with less repeat-induced fragmentation.

Still, the Sanger method has been limited by the cost of sequencing and long run times, encouraging the development of many next-generation sequencing (NGS) technologies such as: 454 pyrosequencing, Illumina Solexa, ABI SOLiD, and Helicos [7, 10]. Although NGS approaches offer reduced costs and run times compared to Sanger sequencing, they also produce shorter reads, complicating the assembly of repetitive genomic regions [15]. Nevertheless, NGS technologies have been successfully used in applications such as: resequencing, variant discovery, transcriptome sequencing, finishing, and even in *de novo* sequencing [10].

Despite improvements in sequencing technology, no sequencer produces sufficient data to assemble a complete genome in a single experiment [9]. Instead, sequencing reads are often assembled into a set of contiguous fragments called contigs and stitched together into longer scaffolds (ordered contigs with gaps of known or unknown length between them). Draft assemblies are therefore incomplete assemblies with gaps. Gaps in assemblies from Sanger sequencing data have been attributed to secondary structure formation and other technical issues related to its clone-based approach [4]. Gaps in assemblies from NGS sequencing data can be attributed to (a) the presence of repeat regions that are considerably longer than individual reads, and (b) to other technology-specific reasons. The bane of all sequencing technologies that involve sequencing fragments followed by an assembly process is the presence of repeat regions in genomes. Repeat regions tend to grow in size, number, and complexity in more evolved organisms [16].

Draft assemblies can be completed in a process known as finishing where the missing sequence (gaps) between the contigs is obtained, low quality regions of contigs are improved, misassemblies are resolved, and the scaffolds are ordered [9]. Finishing typically accounts for the majority of the labor and cost of genome projects, presenting a challenge for most projects, a point which is illustrated by the fact that as of writing this paper (December 2010), roughly 70% of the 4928 assemblies available from the NCBI Microbial Genome Sequencing database are draft assemblies. Many of the draft assemblies have gaps of varying length and quantity. Incomplete assemblies prevent comprehensive comparative genomic analyses to be performed. Furthermore, these gaps prevent the accurate mapping of short read sequencing data generated to address a wide variety of questions including genetic variation, RNA expression, protein-DNA interactions and chromosome conformation [15].

There are many different approaches to finishing draft assemblies. One common but labor intensive approach is PCR amplification of the sequence spanning gaps [9]. Many tools such as Consed [5], Dupfinisher [6], and ABACAS [3] have been developed to facilitate PCR primer design and to assist in finishing the sequence. Another approach that facilitates contig ordering is a technology called optical mapping which produces a map of all the restriction sites across the genome to which contigs can be aligned [14]. Other existing tools include OSLay [13] and Reconciliator [19]. Approaches such as AMOScmp facilitate scaffolding improvement without the need for any further experimental data generation by producing a guided assembly in which reads are aligned to the genome of a related organism [12].

To address the task of filling in missing sequence (gap closure), further sequencing is often required. One approach is to sequence amplified PCR fragments spanning gaps; however, if the assembly is highly fragmented this option may be costly [9]. Another approach is to resequence the entire DNA library, but using a different sequencing platform in order to avoid the biases from previous sequencing efforts. The cost of finishing has been considerably lowered with the advent of NGS technologies [10]. One study found that NGS resequencing is well suited for finishing, completely closing all gaps in two out of six organisms and significantly reducing the number of gaps in the others [4]. Specialized tools have been developed to fill in the sequence gaps using NGS resequencing reads. One such tool is IMAGE which performs NGS resequencing gap closure by producing local assemblies of NGS reads corresponding to gap regions [17]. Also, some "hybrid" assemblers such as VELVET [18] and CABOG [11] indirectly accomplish gap closure by allowing multiple read libraries with varying read lengths and library properties to be utilized for assembly.

Although many finishing tools already exist, the excess of draft assemblies in public repositories makes it clear there is still a need for more practical finishing approaches and better bioinformatic tools to support them. In this paper we present an algorithm for the specialized finishing task of **Clo**sing **G**aps (**CloG**) that are caused by sequencing bias (not repeats) using NGS resequencing and we show that our approach closes more gaps than VELVET and IMAGE in the draft assembly of *Burkholderia dolosa* AUO158.

## II. The CloG Pipeline

Our Closing Gap (CloG) pipeline consists of several stages: trimming contig ends, creating hybrid assemblies, and stitching.

### *A. Trimming Contig Ends*

It is useful to trim contig ends before a "hybrid" assembly is created because sequencing or assembly technology limitations often cause these ends to be of low quality. This idea has been explored before where methods trim a fixed length from both ends of every contig. We offer a more refined approach here. The idea is to trim portions at the ends of contigs that have low quality. Quality of bases in a contig is assessed by looking at coverage values, which can be obtained by using tools such as Bowtie [8]. Contig ends are then scanned, starting from each end until a region of sufficiently high quality is reached. The low quality tips are then trimmed.

### *B. Hybrid Assembly*

A "hybrid" assembly is first created using one of many different *de novo* assemblers. It is called a hybrid assembly because the input consists of sequence data from two sources – the trimmed contigs of the original draft assembly and the millions of short reads from the NGS data set. *De novo* assemblers, such as VELVET [10, 18] and CABOG [11], are able to work with a mixture of different read types. CloG uses the VELVET assembler to generate a hybrid assembly.

*De novo* assemblers often produce flawed assemblies because of non-uniform coverage by the reads and errors in reads. Note that "long reads" from the draft assembly have to be dealt with in a special manner since otherwise any gap region covered by a long read would correspond to a "low coverage" region. VELVET handles them by assembling from short reads first and then using the draft assembly contigs as long reads to help resolve branching problems.

While the above process may generate new portions of the sequence, it may widely disagree with the draft assembly and the resulting hybrid assembly can be even more fragmented than the original draft assembly. Instead of taking the hybrid assembly as the final one, CloG applies "stitching" to derive maximum benefit from both assemblies.

### *C. Stitching*

Stitching is the process of closing gaps between adjacent contigs in the draft assembly using contigs from the hybrid assembly. The basic idea behind stitching is to generate a consensus sequence by finding overlapping regions of the two assemblies. As mentioned above, due to sequencing or assembly technology limitations often these ends tend to be of low quality. Stitching is not straightforward because there may be large regions of disagreement between the two assemblies. The final consensus assembly sequence in such regions may be reached by choosing either one of the two input assemblies.

CloG addresses the difficulties in stitching by introducing the concept of common "seed" sequences. "Seeds" are specific length (default: 200 bp) sequences located at a specific distance (default: 800 bp) away from (untrimmed) contig ends. Stitching is performed by first identifying a hybrid assembly contig that shares common seeds with two different draft assembly contigs. Seeds are said to be shared if the alignment score is above a specified threshold. Consensus sequences are constructed by then stitching together appropriate fragments from the two assemblies.

Note that this step uses "untrimmed" contigs from the draft assembly because the stitching process forces the ends to get deleted anyways. Also note that in this version of CloG, seeds are at a fixed location from either end. In the next version of CloG, we intend to experiment with a variable position of the seed based on the quality of the alignment between the two contigs in question.

**Constructing Consensus Assemblies by Stitching** As shown below in Fig. 1 stitching is implemented with the help of seeds. Since contig ends usually have problems, as mentioned earlier, if there is disagreement at the draft assembly's contig end, CloG chooses the corresponding hybrid assembly as the consensus. Otherwise the original draft assembly is used.

There are several reasons to choose the original draft assembly sequences as consensus in non-end regions. First,

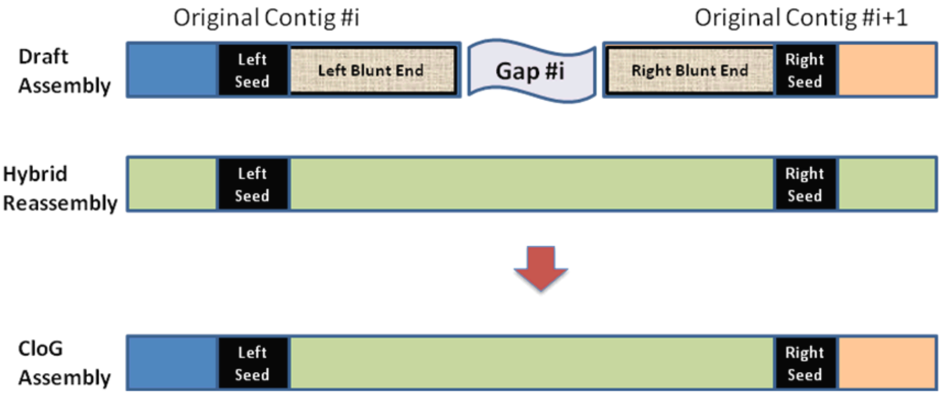

Figure 1. Consensus is taken from hybrid assembly sequence for draft assembly contig end regions.

we assume that in the non-end regions the draft assembly is of good quality. Second, annotations and analysis may have been done for the existing draft assembly and it is preferable to retain the old sequences and annotations as much as possible. Third, the mapping of reads to the sequence is not uniform and reads are not error free. Finally, the performance and parameter settings of *de novo* assemblers have to be considered. For example, VELVET outputs scaffolds with Ns instead of contigs for paired-end reads.

**Matching Hybrid Contigs to Seed Pairs** The primary reason for introducing the concept of seeds is that it reduces the problem of finding overlaps between assemblies to that of identifying pairs of short seed sequences in the hybrid assembly, thus avoiding the bad contig end problem. However, specificity is compromised for sensitivity since seeds may be aligned at multiple locations in the hybrid assembly.

Candidate hybrid assembly segments are identified by aligning a pair of seeds associated with a gap to the hybrid assembly. If both seeds of a pair get aligned to the same hybrid assembly contig with the right orientation, the corresponding hybrid assembly segment is considered a candidate for closing that gap. In some cases, a long candidate hybrid assembly segment helps to close multiple gaps. Usually at most one such candidate hybrid assembly segment can be found for a gap. Due to repeats in the genome or errors in the hybrid assembly, however, it is possible that multiple candidate assembly segments can be found for a single gap. In such situations, the best hybrid assembly fragment that closes a gap is defined as the one that closes the most number of adjacent gaps. If there is more than one fragment that fits the description then ties are broken arbitrarily. Fig. 2 shows an example.

## III. Experiments and Results

CloG was applied to close gaps for the bacterium *B. dolosa* AUO158 [1] draft assembly using Illumina paired-end reads. We compared the performance of CloG with that of two other popular tools: VELVET [18] and IMAGE [17].

### *A. Data*

**Draft Assembly** *Burkholderia dolosa* AUO158 contig/ scaffold files were downloaded [www.broadinstitute.org]. The draft genome was sequenced by 454 WGS (coverage: 7.63X) and assembled with Newbler. The assembly of the 3-chromosome genome of *B. dolosa* consists of 233 contigs. Gap length between two contigs is indicated by the number of Ns between them; or marked as “unknown length” if the number is 100. The draft assembly had the following characteristics: Total length = 6,247,594 bp; Max contig length = 209,563 bp; N50 length = 50,165; Total gap length = 172,806. Note that N50 is the length of the smallest contig such that 50% of the length of the genome is contained in contigs of that size or greater. For a fair comparison, a uniform estimate of the size of the draft genome was used to compute N50 for all assemblies in Table I.

**NGS Reads** The Illumina reads had the following characteristics: Number of paired reads = 7,728,520; Read length = 40 bp; Average insert length = ~200 bp. (Note that every paired read corresponds to a DNA fragment whose terminating regions are sequenced. The total length of the DNA fragment is referred to as the "insert length".)

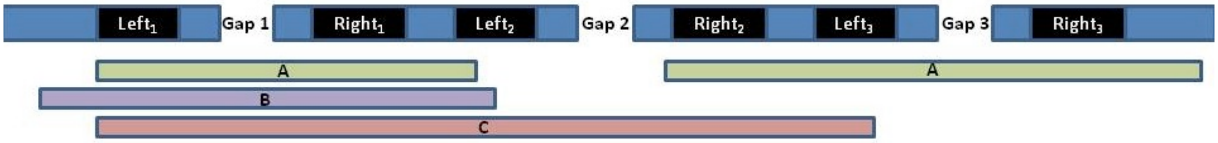

Figure 2. Example of matching hybrid contigs to gaps. Four ordered draft assembly contigs and three hybrid assembly contigs (A, B, and C) are shown. Although Contig A covers more seeds, Contig C will be chosen to close Gap 1 because it closes more consecutive gaps than the other contigs.

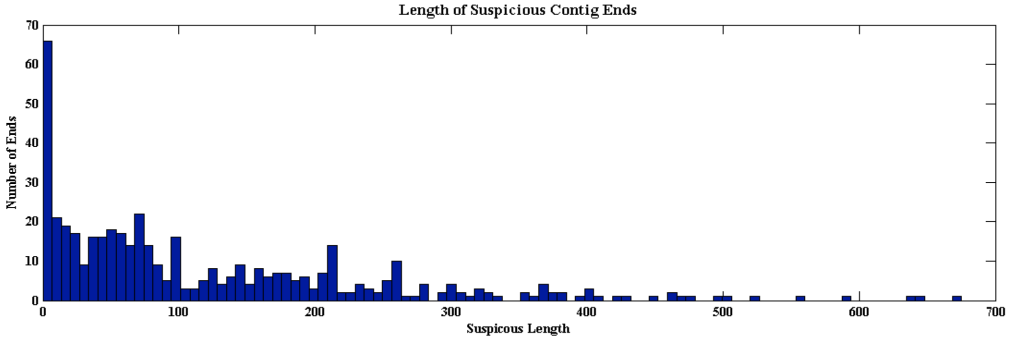

Figure 3. Distribution of the locations of low coverage segments relative to the ends of 233 contigs (466 ends).

### B. *Suspicious Contig Ends*

By aligning Illumina paired-end (PE) reads to the *B. dolosa* draft assembly, we were able to obtain the coverage information at each base position. Low PE read coverage can be the result of non-uniform sampling or an error in the assembly. Although one cannot be sure which of the two cases apply to each low coverage segment, it is worthwhile noting that contig ends are highly prone to low coverage. Fig. 3 shows that most low coverage segments are within a few hundred bases from contig ends. As discussed in Section 2A, to avoid errors, suspicious ends in draft assembly contigs were trimmed before they were assembled into a hybrid assembly.

### C. *Hybrid Assembly Using VELVET*

A reference-guided assembly was generated by passing the following input to VELVET: (a) original draft assembly as long reads and (b) Illumina short reads. The assembly result statistics are shown in Table 1. The assembly statistics for using only short reads is provided in Table 1 as well. By comparing these two assemblies we can see that reference-guided hybrid assembly produces much longer contigs than those assembled from resequencing short reads only.

Since VELVET generates scaffolds instead of contigs when the input is PE reads, contigs were extracted from VELVET scaffolds to get the statistics for contigs without Ns. From Table 1 we can see that VELVET produces longer contigs than the original draft assembly. The improvement in contig length, however, is compromised by a few other factors. First, the total number of contigs is increased from 233 to 2,842, drastically increasing the number of gaps. Second, the total assembly size is unexpectedly increased by ~1 Mbp. Also, while contigs in the original draft assembly are well ordered, the ordering of the VELVET scaffolds is not clear. We show in the next section that CloG was able to take advantage of the hybrid assembly while avoiding these drawbacks by stitching.

### D. *Stitching Together Draft Assembly with VELVET Hybrid Assembly*

For each gap, the left and right 200 bp seed sequences were extracted 800 bp away from the left and right ends of the flanking contigs. These seed pairs were then blasted against the original draft assembly to identify candidate hybrid assembly fragments that close gaps. For each gap at most one candidate was chosen using the criteria mentioned in Section 2C. The CloG assembly was then generated by stitching the original draft assembly with the candidate hybrid assembly fragments.

For 180 out of 233 gaps, CloG was able to find candidate hybrid assembly fragments to perform the stitching. As mentioned earlier, the hybrid assembly generated using

TABLE I. ASSEMBLY STATISTICS

| Assembly | N50<br>(bp) | Longest<br>(bp) | Total<br>(bp) | # Contigs<br>or Scaffolds |
|---|---|---|---|---|
| Draft Assembly Contig | 50,165 | 209,563 | 6,247,594 | 233 |
| VELVET Scaffold: short | 31,867 | 106,721 | 6,984,760 | 2,624 |
| VELVET Contig: long + short | 85,405 | 228,675 | 7,241,371 | 2,842 |
| VELVET Scaffold: long + short | 174,839 | 671,795 | 7,307,958 | 2,244 |
| IMAGE Contig | 71,006 | 209,805 | 6,239,725 | 175 |
| CloG Contig | 91,940 | 227,575 | 6,309,721 | 198 |
| CloG Scaffold | 190,379 | 656,073 | 6,316,430 | 53 |

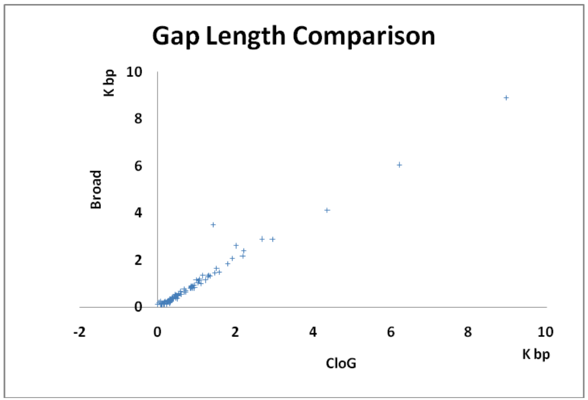

Figure 4. Gap length comparison between the draft assembly and CloG assembly. There was widespread agreement.

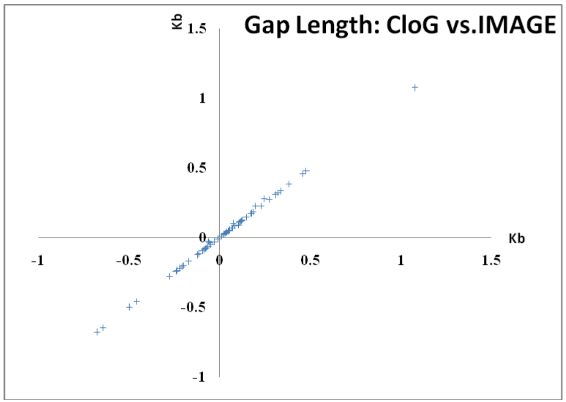

Figure 5. Gap length comparison between IMAGE (X-axis) and CloG (Y-axis).

VELVET outputs scaffolds instead of contigs for paired-end reads and thus, candidate hybrid assembly fragments may contain short gaps of length less than that of the insert length. As a result, the immediate products of the stitching step of CloG are scaffolds instead of contigs. A total of 53 ordered scaffolds were generated by CloG, stitching 180 gaps. We say “stitching” instead of “closing” because of the Ns in the hybrid assembly fragment that wind up introducing small gaps. By extracting contigs from CloG scaffolds, statistics of CloG contigs were obtained and are shown in Table 1. Three VELVET assemblies are listed in Table 1. The word “short” in all of them indicates that the short Illumina reads were used in all of the VELVET assemblies. Two of the three VELVET assemblies also used the draft contigs as long reads and are indicated by the word “long”.

The N50 length of CloG scaffolds is longer than those in the VELVET hybrid assembly, while the longest contigs were comparable. Also, the CloG assembly resulted in fewer contigs and scaffolds and the total genome length is comparable to that of the reference.

Among the 180 gaps that were stitched, estimated gap lengths for 99 gaps were provided by the draft assembly while the remaining 75 were indicated as being of unknown length. We compared the gap lengths indicated by CloG with those indicated by the draft assembly in Fig. 4. We can see that, except for a few cases, most gap lengths agreed well for the two methods. For the more interesting case of the 75 gaps with no reference lengths, we examined their length distribution as indicated by CloG in Fig. 6. It appears that most such gaps are of short length or are overlapped by their flanking contigs.

Although the total number of contigs was not greatly reduced, 145 out of the 198 gaps in the CloG assembly were newly introduced small gaps from the hybrid assembly with lengths shorter than the insert length (200 bp). In fact, the total length of the 145 small gaps was only 6,709 bp. Since these gaps are short, it is likely that most of them can be closed by reads that were unused in the draft assembly. This can be obtained from the Broad Institute website.

### *E. Comparing IMAGE and CloG Assemblies*

Next, we compared the performance of CloG with that of IMAGE [17], a tool dedicated to closing gaps in draft assemblies using Illumina paired-end reads. The best result we were able to obtain was the following: using the pre-trimmed draft assembly as the reference, after 20 iterations, 48 out of 233 gaps were closed. Of the 48 gaps closed by IMAGE, CloG failed to stitch only one. The gap length of the 47 gaps that were closed by both tools agreed well and is shown in Fig. 5. It is worth noting that except for one gap with length ~1.2 Kbp, IMAGE mostly closed gaps with short

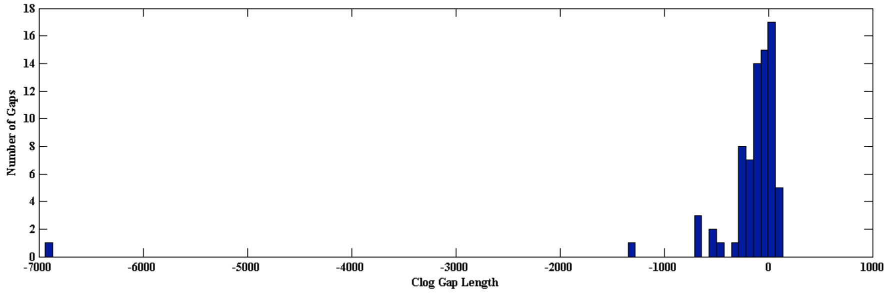

Figure 6. CloG’s gap length for 75 gaps with unknown Broad gap lengths. Most such gaps are reported as short gaps in the CloG assembly. In some cases, they are reported to be overlapping flanking reference contigs, indicated by negative lengths in the CloG assembly. We verified the rare case with ~7000 bp overlap by blasting the flanking contigs.

length or overlapping flanking contigs.

The assembly statistics for IMAGE and CloG are shown in Table 1. The CloG assembly has a larger N50 and longer contigs than IMAGE. Although the number of contigs in the IMAGE assembly was smaller than in the CloG assembly, the actual number of gaps closed by CloG is larger than that of IMAGE because many of the gaps in the CloG assembly were newly introduced small gaps, caused by the Ns in the corresponding hybrid assembly fragment. The resulting total genome length of IMAGE is shorter than the original draft assembly, partly due to the fact that most of the gaps that were closed by IMAGE were of short length or are actually overlapped by flanking contigs, and also because the pre-trimming of draft assembly contigs to remove low coverage contig ends results in a shorter assembled genome.

## IV. Conclusion

We have presented our pipeline, CloG, for closing gaps in draft assemblies using short resequencing NGS data. We were able to show that CloG outperformed VELVET and IMAGE in our experiments on the *B. dolosa* draft assembly with Illumina resequencing reads. The CloG assembly resulted in longer N50 and fewer gaps than the VELVET hybrid assembly, while still maintaining the ordering information of the original draft assembly. The CloG assembly also resulted in longer N50 than the IMAGE assembly and CloG was able to stitch 180 gaps, while IMAGE closed only 48 gaps. Although the number of contigs in the CloG assembly is larger than that of the IMAGE assembly, CloG resolved a larger fraction of the gaps than IMAGE. Thus, the CloG assembly produced a larger number of smaller gaps with a smaller total length, suggesting that CloG performed better than IMAGE and VELVET at the task of gap closure using short resequencing reads. Furthermore, in future work we hope to improve CloG's performance by using a variable seed location based on read quality and coverage information.

## Acknowledgments

GN gratefully acknowledges the support of NIH Grants P01 DA15027-01 and NIH/NIGMS S06 GM008205, FREA Award from NICHD/EARDA grant G11HD038341, and a grant from Florida Department of Health. We thank the Broad Institute for sequencing the *Burkholderia dolosa* genome and making the draft assembly available for download.

## References

[1] *Burkholderia dolosa Sequencing Project, Broad Institute of Harvard and MIT.* [cited; Available from: http://www.broadinstitute.org/annotation/genome/burkholderia_dolosa/MultiDownloads.html.

[2] Adams, J.U., *DNA Sequencing Technologies.* Nature Education, 2008. **1**(1).

[3] Assefa, S., T.M. Keane, T.D. Otto, C. Newbold, and M. Berriman, *ABACAS: algorithm-based automatic contiguation of assembled sequences.* Bioinformatics, 2009. **25**(15): p. 1968-9.

[4] Goldberg, S.M., et al., *A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes.* Proc Natl Acad Sci U S A, 2006. **103**(30): p. 11240-5.

[5] Gordon, D., C. Abajian, and P. Green, *Consed: a graphical tool for sequence finishing.* Genome Res, 1998. **8**(3): p. 195-202.

[6] Han, C.S. and P. Chain. *Finishing Repetitive Regions Automatically with Dupfinisher.* in *International Conference on Bioinformatics and Computational Biology (BIOCOMP'06)*. 2006.

[7] Hert, D.G., C.P. Fredlake, and A.E. Barron, *Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods.* Electrophoresis, 2008. **29**(23): p. 4618-26.

[8] Langmead, B., C. Trapnell, M. Pop, and S.L. Salzberg, *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.

[9] Lapidus, A.L., *Genome Sequence Databases (Overview): Sequencing and Assembly*. 2009, Lawrence Berkeley National Laboratory.

[10] Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

[11] Miller, J.R., et al., *Aggressive assembly of pyrosequencing reads with mates.* Bioinformatics, 2008. **24**(24): p. 2818-24.

[12] Pop, M., A. Phillippy, A.L. Delcher, and S.L. Salzberg, *Comparative genome assembly.* Brief Bioinform, 2004. **5**(3): p. 237-48.

[13] Richter, D.C., S.C. Schuster, and D.H. Huson, *OSLay: optimal syntenic layout of unfinished assemblies.* Bioinformatics, 2007. **23**(13): p. 1573-9.

[14] Samad, A., E.F. Huff, W. Cai, and D.C. Schwartz, *Optical mapping: a novel, single-molecule approach to genomic analysis.* Genome Res, 1995. **5**(1): p. 1-4.

[15] Shendure, J. and H. Ji, *Next-generation DNA sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1135-45.

[16] Smit, A.F., *The origin of interspersed repeats in the human genome.* Curr Opin Genet Dev, 1996. **6**(6): p. 743-8.

[17] Tsai, I.J., T.D. Otto, and M. Berriman, *Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps.* Genome Biol, 2010. **11**(4): p. R41.

[18] Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.

[19] Zimin, A.V., D.R. Smith, G. Sutton, and J.A. Yorke, *Assembly reconciliation.* Bioinformatics, 2008. **24**(1): p. 42-5.