Supplement
Miller et al. 2008 "Aggressive Assembly of Pyrosequencing Reads with Mates"

Data sets that included both 454 FLX reads and Sanger reads were selected to test the assemblers. The 454 SFF files for P*orphyromonas gingivalis W*83 and *Psychromas sp.* CNPT3 included the line: "qualityScoreVersion>1.1.03</qualityScoreVersion>" while *Escherichia coli* K-12 and *Cryptosporidium muris* RN66 did not. In the tables below, clear ranges indicate the usable base range provided with each read. Mate distance indicates the predicted mate separation as determined during library preparation. The 454 mates derive from circularized DNA fragments such that reads can include the joined fragment ends. Plates from these libraries include both mated and unmated reads. Coverage shows the number of times the genome would be covered by all clear ranges laid end to end.

Data from the bacterium *Escherichia coli* K-12 was used to test assemblers. The unmated reads are from a 454 FLX full-plate (SRA000156). The "F1" data set (see Table 1) is from the ETX12BK01.sff half-plate file. The "F2" data set is from the ETX12BK02.sff half-plate file. The mated reads are from a 454 FLX half-plate (SRA001028) as found in the E6PIHNP01.sff file. No Sanger reads are available from public sources. The reference, GenBank NC_000913, contained 4,639,675 bases in one circular contig. The reference had been created by the 13 Janus shotgun strategy.

Data from the bacterium *Psychromas* sp. CNPT3 included Sanger and FLX reads. The "S1" data set of Sanger mated reads derived from a 40Kbp library (MOORE_M063-F-01-40KB). The "S2" data set of Sanger reads derived from two 4Kbp libraries (MOORE_M063-G-01-4KB). No assemblies with the "S2" data set were described in the paper so "S2" was omitted from Table 1; "S2" assemblies are included in this supplement, however. The statistics for "S2" are: 539 unmated reads, average length 621; 13,684 mated reads, average length 799; genome coverage 3.8X. The unmated FLX reads derived from a single full-plate. The "F1" data set was from the EE4TTLG01.sff half-plate file. The "F2" data set was from the EE4TTLG02.sff half-plate file. The reference, RefSeq NZ_AAPG00000000, contained 2,945,265 bases in one linear contig.

Data from the bacterium *Porphyromonas gingivalis* W83 was used. The "S1" data set included Sanger mated reads from a 40Kbp fosmid library. As stated in the paper, the "S1" data set provided 1.0X genome coverage by reads. The "S2" data set is merely a random sample of "S1" that provides 0.4X coverage. No assemblies with the "S2" data set were described in the paper so "S2" was omitted from Table 1; "S2" assemblies are included in this supplement, however. FLX reads derived from two full plates, one unmated and one mated (SRA001027). The "F1" unmated data set is from the E9T0MN001.sff half-plate file. The "F2" data set is from the E9T0MN002.sff half-plate file. The "M1" mated data set is from the E8YURXS01.sff half-plate file. The "M2" data set is from the E8YURXS02.sff half-plate file.. The reference, GenBank NC_AE015924, containing 2343476 bases in one circular contig, had been sequenced by Sanger chemistry, assembled with TIGR Assembler, and finished at TIGR (Nelson 1995). Our test data, generated at JCVI in 2008, was distinct.

Data from the eukaryote *Cryptosporidium muris* RN66 was used though no independent, finished reference was available. The "S1" dataset is derived from four Sanger libraries: a 40Kbp fosmid library; two 6-8Kbp plasmid libraries; and a 2-3Kbp plasmid library. The "F1" data set is an FLX full plate of unmated reads (SRA001029) delivered as ER3CQ4U01.sff and ER3CQ4U02.sff half-plate files.

Runtime and memory usage for the assemblers was compared using two datasets derived for *Cryptosporidium muris* RN66. The first dataset consisted exclusively of the Pyro data while the second set included all available data, comprising 510,758 reads. The experiment was run on a dual-processor, dual-core 2.2GHz AMD Opteron 64-bit machine with 32GB RAM and SuSE Linux.

S 1. Assembly runtime and memory usage for *Cryptosporidium muris* RN66

	Runtime on Pyro data	Memory on Pyro data	Runtime on All data	Memory on All Data
CABOG	26.3	864	58.4	2,944
Newbler	17.9	652	34.1	827
PCAP	17.4	1,229	81.3	3,198
Euler-SR	1,092.4	1,715	247.6	1,907

Runtimes are in minutes of wall time. Memory consumption is in MB. Euler-SR did not run to completion on either data set.

Contig and scaffold size statistics were collected by automatic and uniform processing of the FASTA outputs of all assemblers on all datasets. The results were dumped into a spreadsheet provided separately. In addition to a contigs FASTA file, CABOG generates FASTA files of degenerate unitigs (not placed in the assembly) and surrogate unitigs (placed multiple times in the assembly); these

FASTA files were not included. Some assemblers generate large numbers of very small contigs, which may be equivalent to CABOG's degenerates; counting these would bias the "# Contigs" statistic in favor of CABOG. Therefore, for all assemblers, the statistics were based on contigs of length 2Kbp or more.

S 2. Contig and scaffold size statistics for all assemblies. See Excel spreadsheet provided separately.

CABOG and Newbler assemblies were compared to annotation of repeats (see Methods) on the *P. gingivalis* genome reference sequence. In calculating the contig size statistics elsewhere, a 2Kbp cutoff was applied to Newbler contigs and CABOG contigs. Here, that cutoff would bias statistics in favor of CABOG, since Newbler generated many small contigs containing repeat sequence. Therefore, all contigs generated by both assemblers were included here. (In other words, the following files were analyzed: *.ctg*.FASTA from CABOG and 454AllContigs.fna from Newbler).

S 3. *P. gingivalis* repeat coverage by CABOG and Newbler.

|Assembler	# Repeats Spanned	Average Spanned Repeat Length
A1-p.gingivalis-f1		
CABOG	20	360.20
Newbler	18	656.39
B1-p.gingivalis-m2		
CABOG	23	1,586.13
Newbler	9	486.33
B2-p.gingivalis-m1-m2		
CABOG	40	1,119.06
Newbler	14	536.07
B3-p.gingivalis-f1-m2		
CABOG	34	1,098.81
Newbler	14	702.60
C1-p.gingivalis-f1-fos1.0		
CABOG	30	1,241.06
Newbler	21	660.86
D1-p.gingivalis-m2-fos1.0		
CABOG	26	1,981.46
Newbler	9	814.89

CABOG and Newbler assemblies were compared to the reference sequences to measure genome coverage. The mapping of scaffolds to reference, generated by Snapper (http://kmer.sf.net), were filtered to retain the longest, best placement for each scaffold or scaffold segment along the genome.

S 4. Genome coverage by 2Kbp scaffolds generated by CABOG and Newbler.

Assembly	CABOG Coverage	Newbler Coverage
A1-e.coli-F1	97.06%	97.36%
A1-e.coli-F2	96.76%	97.25%
A1-p.gingivalis-F1	94.35%	94.29%
A1-p.gingivalis-F2	94.21%	94.33%
A1-psychromonas-F1	97.63%	97.91%
A1-psychromonas-F2	97.59%	97.91%
A2-e.coli-F1-F2	98.00%	97.90%
A2-p.gingivalis-F1-F2	94.63%	94.26%
A2-psychromonas-F1-F2	97.64%	97.89%
B1-e.coli-M1	98.93%	96.16%
B1-p.gingivalis-M1	98.69%	94.10%
B1-p.gingivalis-M2	98.74%	94.08%
B2-p.gingivalis-M1-M2	98.87%	93.88%
B3-e.coli-F1-M1	99.54%	97.54%
B3-e.coli-F2-M1	99.66%	97.59%
B3-p.gingivalis-F1-M1	98.54%	94.08%
B3-p.gingivalis-F1-M2	98.80%	94.18%
B3-p.gingivalis-F2-M1	98.90%	93.88%
B3-p.gingivalis-F2-M2	98.88%	94.02%
C1-p.gingivalis-F1-S1	96.62%	94.34%
C1-p.gingivalis-F2-S1	96.37%	94.08%

C2-p.gingivalis-F1-S2	94.99%	94.45%
C2-p.gingivalis-F2-S2	95.25%	94.19%
C3-psychromonas-F1-S1	97.82%	97.84%
C3-psychromonas-F2-S1	97.61%	97.50%
C4-psychromonas-F1-S2	98.91%	98.26%
C4-psychromonas-F2-S2	98.76%	98.33%
D1-p.gingivalis-M1-S1	98.38%	94.01%
D1-p.gingivalis-M2-S1	98.72%	93.92%
D2-p.gingivalis-M1-S2	97.26%	94.08%
D2-p.gingivalis-M2-S2	97.87%	94.00%