Supplementary material for "GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies"

Alexey A. Gritsenko*†‡ Jurgen F. Nijkamp*‡
Marcel J.T. Reinders*†‡ Dick de Ridder*†‡

1 Sequence assembly

To select the k-mer length for *de novo* genome assembly using Velvet we tried different values of k and calculated length and accuracy statistics for the resulting assemblies. We measured the number of contigs, maximum contig length, the N50 statistic and total assembly length to get a feel of assembly completeness and contiguity. We also measured *coverage* as percentage of reads mapping to the genome, and *accuracy* as the percentage of paired reads with proper pairing (as defined by BWA, [Li and Durbin, 2009]). To measure accuracy and coverage, single- and paired-end mapping of the reads to the assembled contigs was performed using BWA. Tables S1, S3 and S2 show these statistics for different k for *E. coli*, *P. syringae* and *P. suwonensis* assemblies correspondingly.

2 Phylogenetic tree construction

The phylogenetic tree for *E. coli* stains MG1655, BW2952 and DH10B was constructed using the SplitsTree 4 package [Huson and Bryant, 2006] and the coverage distance function from [Henz *et al.*, 2004]. Genome alignments were obtained using MUMmer [Delcher *et al.*, 2002] with settings from [Auch *et al.*, 2010].

*The Delft Bioinformatics Lab, Department of Mediamatics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands.
†Platform Green Synthetic Biology, P.O. Box 5057, 2600 GA Delft, The Netherlands.
‡Kluyver Centre for Genomics of Industrial Fermentation, P.O. Box 5057, 2600 GA Delft, The Netherlands.

Table S1: *E. coli* assembly statistics for different k-mer lengths of Velvet. Assembly for the chosen k is highlighted.

k	Contigs	N50	Maximum	Total length	Coverage	Accuracy
19	4,180	1,621	9,259	4,505,092	91.66%	72.38%
21	1,485	5,466	40,066	4,516,751	93.86%	87.63%
23	951	9,181	41,213	4,521,870	94.47%	90.93%
25	722	12,114	55,230	4,527,423	94.83%	92.51%
27	581	15,644	73,054	4,529,084	95.00%	93.50%
29	512	18,358	71,241	4,531,657	95.16%	94.00%
31	481	19,872	73,062	4,535,181	95.26%	94.21%
33	586	15,104	62,943	4,541,512	95.38%	93.75%
35	11,079	445	2,853	4,245,608	82.96%	36.44%

Table S2: *P. syringae* assembly statistics for different k-mer lengths of Velvet.

k	Contigs	N50	Maximum	Total length	Coverage	Accuracy
19	5,059	1,892	12,464	5,846,661	84.47%	64.51%
21	1,926	7,024	42,317	5,886,062	86.70%	80.78%
23	1,560	8,599	46,055	5,902,217	87.20%	82.93%
25	1,990	5,977	24,056	5,930,228	87.55%	81.27%
27	3,829	2,623	13,478	5,946,020	87.32%	72.76%
29	8,825	865	8,433	5,592,074	81.59%	45.63%
31	5,523	343	2,676	1,755,054	28.42%	6.57%
33	57	500	3,166	21,040	1.10%	0.61%
35	15	244	448	3,588	0.24%	0.04%

3 Scaffolder running time

Scaffolding and mapping running times were measured for all experiments. This data is presented in Table S4. Scaffolding time for Velvet and mapping time for SSPACE have been calculated from the programs' output. Preprocessing of reads prior to mapping and post-processing of the mapper's output was counted as mapping time.

References

[Auch *et al.*, 2010] Auch, A.F., Klenk, H.-P. and Göker, M. (2010) Standard operating procedure for calculating genome-to-genome distances based on

Table S3: *P. suwonensis* assembly statistics for different *k*-mer lengths of Velvet.

k	Contigs	N50	Maximum	Total length	Coverage	Accuracy
21	798	178	672	148,597	1.16%	0.70%
23	3,640	194	609	724,989	6.90%	6.73%
25	6,457	222	900	1,451,717	15.79%	16.58%
27	8,045	264	1,273	2,084,930	25.28%	28.08%
29	8,538	313	1,793	2,522,405	32.97%	37.82%
31	8,306	385	2,421	2,846,252	39.62%	46.66%
33	7,520	482	3,595	3,069,871	45.06%	54.49%
35	6,391	635	3,505	3,220,911	49.30%	61.05%
37	5,270	857	5,770	3,321,047	52.65%	66.44%
39	3,978	1,223	7,233	3,371,436	55.17%	70.96%
41	2,939	1,706	11,487	3,396,276	56.95%	74.35%
43	2,039	2,721	16,786	3,407,475	58.35%	77.06%
45	1,435	3,959	16,772	3,408,865	59.12%	78.75%
47	1,020	5,818	23,722	3,408,282	59.68%	79.90%
49	697	9,367	36,131	3,405,741	60.05%	80.72%
51	537	12,638	46,479	3,402,802	60.21%	81.10%
53	427	16,065	64,878	3,400,488	60.33%	81.40%
55	351	19,866	87,700	3,399,187	60.42%	81.60%
57	308	24,193	87,698	3,396,963	60.49%	81.74%
59	303	26,043	90,572	3,394,128	60.47%	81.74%
61	309	24,862	90,573	3,392,147	60.46%	81.73%
63	301	24,005	78,697	3,386,612	60.46%	81.74%
65	334	21,764	78,707	3,380,022	60.38%	81.63%
67	380	17,029	78,569	3,372,389	60.26%	81.44%
69	462	13,262	74,778	3,363,394	60.10%	81.18%
71	648	9,303	54,433	3,351,627	59.81%	80.67%
73	1,088	5,308	22,390	3,338,680	59.36%	79.71%
75	4,214	933	13,128	3,082,996	53.13%	68.00%

high-scoring segment pairs, *Standards in Genomic Sciences*, **2**, 142–148, doi:10.4056/sigs.541628.

[Delcher *et al.*, 2002] Delcher, A.L., Phillippy, A., Carlton, J. and Salzberg, S.L. (2002) Fast algorithms for large-scale genome alignment and comparison, *Nucleic Acids Research*, **30**, 2478–2483, doi:10.1093/nar/30.11.2478.

Table S4: Scaffolder and mapping running time. For *E. coli* "(all)" denotes usage of paired reads and related genomes of *E. coli* strains DH10W and BW2952 for scaffolding.

Dataset	Scaffolder	Mapping time, min	Scaffolding time, min	Total time, min
E. coli	Velvet	N/A	8 sec	8 sec
	SSPACE	2 m 48 sec	1 m 7 sec	3 m 11 sec
	GRASS	29 m 55 sec	23 sec	30 m 18 sec
	GRASS+	29 m 55 sec	53 sec	30 m 48 sec
(all)	GRASS+	47 m 16 sec	40 sec	47 m 56 sec
	MIP Scaffolder	68 m 49 sec	2 m 2 sec	70 m 52 sec
SRR001665	OPERA	21 m 11 sec	27 m 45 sec	48 m 56 sec
SRR001666	OPERA	27 m 49 sec	30 sec	28 m 19 sec
P. suwonensis	Velvet	N/A	13 sec	13 sec
	SSPACE	5 m 8 sec	7 m 22 sec	12 m 3 sec
	GRASS	139 m 59 sec	23 sec	140 m 23 sec
	GRASS+	139 m 59 sec	45 sec	140 m 44 sec
	MIP Scaffolder	95 m 37 sec	1 m 1 sec	96 m 37 sec
	OPERA	125 m 28 sec	8 m 19 sec	133 m 47 sec
SRR097515	OPERA	74 m 56 sec	25 sec	75 m 22 sec
SRR191848	OPERA	75 m 32 sec	1 m 53 sec	77 m 25 sec
P. syringae	Velvet	N/A	1 sec	1 sec
	SSPACE	1 m 6 sec	27 sec	1 m 33 sec
	GRASS	13 m 20 sec	15 sec	13 m 35 sec
	GRASS+	13 m 20 sec	3 m 7 sec	16 m 27 sec
	MIP Scaffolder	9 m 19 sec	27 sec	9 m 46 sec
	OPERA	10 m 38 sec	72 m 22 sec	83 m 1 sec

[Henz *et al.*, 2004] Henz, S.R., Huson, D.H., Auch, A.F., Nieselt-Struwe, K. and Schuster, S.C. (2004) Whole-genome prokaryotic phylogeny, *Bioinformatics*, **21**, 2329–2335, doi:10.1093/bioinformatics/bth324.

[Huson and Bryant, 2006] Huson, D.H. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies, *Molecular Biology and Evolution*, **23**, 254–267, doi:10.1093/molbev/msj030.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754–1760, dpi:10.1093/bioinformatics/btp324.