CSH PRESS GENOME RESEARCH

Resource

GAGE: A critical evaluation of genome assemblies and assembly algorithms

Steven L. Salzberg,[1,7] Adam M. Phillippy,[2] Aleksey Zimin,[3] Daniela Puiu,[1] Tanja Magoc,[1] Sergey Koren,[2,4] Todd J. Treangen,[1] Michael C. Schatz,[5] Arthur L. Delcher,[6] Michael Roberts,[3] Guillaume Marçais,[3] Mihai Pop,[4] and James A. Yorke[3]

[1]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA; [2]National Biodefense Analysis and Countermeasures Center, Battelle National Biodefense Institute, Frederick, Maryland 21702, USA; [3]Institute for Physical Sciences and Technology, University of Maryland, College Park, Maryland 20742, USA; [4]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742, USA; [5]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA; [6]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA

J Club
13.02.2012

The GAGE study

- The GAGE (**G**enome **A**ssembly **G**old-standard **E**valuations) study was designed to provide a snapshot of how the latest genome assemblers compare on a sample of large-scale next-generation sequencing projects.

- http://gage.cbcb.umd.edu

- What will an assembly based on short reads look like?
- Which assembly software will produce the best results?
- What parameters should be used when running the software?

Key results

- Comparison of 8 open-source assemblers on 4 different data sets

- Describing all procedures and parameters and providing the complete data sets used for each assembly in study

- Evaluations used real sequence data from high-throughput sequencing machines

Assemblers

- **ABySS 1.2.7** (Simpson et al. 2009)
- **ALLPATHS-LG 3-35218** (Gnerre et al. 2011)
- **Bambus2 3.0.1** (Koren et al. 2011) (http://www.cbcb.umd.edu/software/bambus)
- **CABOG 6.1** (Miller et al. 2008)
- **MSR-CA 1.0** (http://www.genome.umd.edu/MSR_CA_MANUAL.htm)
- **SGA 0.9.8** (Simpson and Durbin 2012)
- **SOAPdenovo 1.0.5** (Li et al. 2010b)
- **Velvet 1.0.13** (Zerbino and Birney 2008)

Data sets

Table 1. Details of the four next-generation sequence data sets used for the GAGE assembly comparison

Species	*S. aureus*	*R. sphaeroides*	Human Chr14	*B. impatiens*
Size (Mb)	2.90	4.60	88.29	250 (est.)
Read length	101, 37	101	101	124
Fragment size, Library 1	180	180	155	400
Number of reads, Library 1	1,294,104	2,050,868	36,504,800	303,118,594
Fragment size, Library 2	3500	3500	2280–2800	3000–4000
Number of reads, Library 2	3,494,070	2,050,868	22,669,408	129,118,270
Fragment size, Library 3			35 kb	8 kb
Number of reads, Library 3			2,405,064	65,081,280

- Reads were error-corrected with Quake or ALPATHS-LG error corrector. SGA used uncorrected reads.

- S. aureus – 45x genome coverage
- R. sphaeroides – 45x genome coverage
- Human Chr14 – 60x genome coverage

Assembly results (1)

Table 2. Assemblies of *S. aureus* (genome size 2,872,915)

	Contigs				Scaffolds			
Assembler	**Num**	**N50 (kb)**	**Errors**	**N50 corr. (kb)**	**Num**	**N50 (kb)**	**Errors**	**N50 corr. (kb)**
ABySS	302	29.2	19	24.8	246	34	1	28
ALLPATHS-LG	**60**	96.7	20	**66.2**	**12**	1,092	**0**	**1,092**
Bambus2	109	50.2	190	16.7	17	1,084	**0**	1,084
CABOG	Could not run: incompatible read lengths in one library							
MSR-CA	94	59.2	34	48.2	17	**2,412**	3	1,022
SGA	252	4.0	**10**	4.0	456	208	1	208
SOAPdenovo	107	**288.2**	65	62.7	99	332	8	284
Velvet	162	48.4	42	41.5	45	762	17	126

The best value for each column is shown in bold. For all assemblies, N50 values are based on the same genome size. The Errors column contains the number of misjoins plus indel errors >5 bp for contigs, and the total number of misjoins for scaffolds. Corrected N50 values were computed after correcting contigs and scaffolds by breaking them at each error. See the evaluation section in the text for details on how errors were identified.

Table 3. Assemblies of *R. sphaeroides* (genome size 4,603,060)

	Contigs				Scaffolds			
Assembler	**Num**	**N50 (kb)**	**Errors**	**N50 corr. (kb)**	**Num**	**N50 (kb)**	**Errors**	**N50 corr. (kb)**
ABySS	1915	5.9	76	4.2	1701	9	3	5
ALLPATHS-LG	204	42.5	49	**34.4**	**34**	**3192**	**0**	**3192**
Bambus2	**177**	93.2	373	12.8	92	2439	2	2419
CABOG	322	20.2	44	17.9	130	66	5	55
MSR-CA	395	22.1	52	19.1	43	2,976	5	2966
SGA	3067	4.5	**12**	2.9	2096	51	**0**	51
SOAPdenovo	204	**131.7**	422	14.3	166	660	3	658
Velvet	583	15.7	43	14.5	178	353	6	270

Columns are the same as in Table 2.

Assembly results (2)

Table 4. Assemblies of human chromosome 14 (ungapped size 88,289,540)

	Contigs				Scaffolds			
Assembler	Num	N50 (kb)	Errors	N50 corr. (kb)	Num	N50 (kb)	Errors	N50 corr. (kb)
ABySS	51,924	2.0	**704**	2.0	51,301	2.1	**9**	2
ALLPATHS-LG	4529	36.5	2760	21.0	**225**	**81,647**	45	**4702**
Bambus2	13,592	5.9	11,943	4.3	1792	324	143	161
CABOG	**3361**	**45.3**	3181	**23.7**	479	393	597	26
MSR-CA	30,103	4.9	5550	4.3	1425	893	1068	94
SGA	56,939	2.7	981	2.7	30,975	83	19	79
SOAPdenovo	22,689	14.7	6424	7.4	13,502	455	268	214
Velvet	45,564	2.3	4910	2.1	3,565	1190	9156	27

Columns are the same as in Table 2.

Table 5. Assemblies of the bumble bee, *B. impatiens* (estimated size 250 Mb)

	Contigs			Scaffolds		
Assembler	Num	N50 (kb)	E-size (kb)	Num	N50 (kb)	E-size (kb)
ALLPATHS-LG	Could not run: incompatible library types					
CABOG	22,107	23.5	34.2	**1191**	1125	1367
MSR-CA	21,885	32.4	46.9	2551	1246	1528
SGA	Program crashed: cause unclear					
SOAPdenovo	**15,957**	**57.1**	**78.2**	5800	**1374**	**1608**
Velvet	Program crashed: insufficient memory (256 GB)					

Column headers have the same meanings as in Table 2.

Assembly accuracy (1)

Table 6. Statistics showing bases that failed to align or were present in different copy numbers in the reference genomes and the assemblies of *S. aureus*, *R. sphaeroides*, and Hs14

Assembler	Assembly size (%)	Chaff size (%)	Unaligned ref bases (%)	Unaligned asm bases (%)	Duplicated ref bases (%)	Compressed ref bases (%)
S. aureus (2.87 Mb)						
ABySS	127.0	66.00	1.37	<0.01	23.30	0.99
ALLPATHS-LG	99.9	0.03	0.62	<0.01	0.03	1.27
Bambus2	98.5	0	1.32	<0.01	<0.01	1.29
MSR-CA	99.6	0.02	1.30	<0.01	0.83	1.01
SGA	98.5	21.38	1.91	<0.01	0.03	1.30
SOAPdenovo	101.3	0.35	0.38	0.01	1.44	1.41
Velvet	99.2	0.45	0.79	0.03	0.10	1.28
R. sphaeroides (4.60 Mb)						
ABySS	108.0	1.65	3.01	0.15	10.04	0.04
ALLPATHS-LG	99.7	0.01	0.47	0.01	0.38	0.30
Bambus2	94.9	0	4.93	<0.01	<0.01	0.24
CABOG	92.1	<0.01	7.51	0.01	0.12	0.70
MSR-CA	96.9	0.02	3.52	0.04	1.04	0.49
SGA	97.8	4.95	2.31	0.02	0.06	0.92
SOAPdenovo	99.9	0.45	0.88	0.02	1.07	0.51
Velvet	97.8	0.54	1.60	0.01	0.29	0.92
Human chromosome 14 (88.29 Mb)						
ABySS	83.1	41.37	17.78	0.03	0.59	0.52
ALLPATHS-LG	95.6	0.03	2.76	0.03	0.27	2.57
Bambus2	77.3	<0.01	20.55	0.07	0.14	4.04
CABOG	97.7	0.03	1.68	0.06	0.16	1.71
MSR-CA	92.5	0.18	8.10	0.57	1.69	2.27
SGA	93.3	107.82	6.97	0.06	0.14	2.14
SOAPdenovo	104.9	3.77	1.83	0.60	6.76	3.76
Velvet	84.7	6.25	15.12	0.31	0.09	0.64

The true size of each genome is shown next to the species name. All table values are expressed as a percentage of the true genome size. Column headers are defined in the main text. Additional statistics are provided in the Supplemental Material.

Assembly accuracy (2)

Table 7. Statistics on insertions, deletions, and misassembly errors in the various assemblies of *S. aureus*, *R. sphaeroides*, and Hs14

		Indels		Contigs			Scaffolds		
Assembler	SNPs	≤5 bp	>5 bp	Misjoins	Inv	Reloc	Misjoins	Inv	Reloc/indel
S. aureus (2.87 Mb)									
ABySS	258	20	9	5	3	2	1	1	0
ALLPATHS-LG	79	4	12	4	0	4	0	0	0
Bambus2	28	56	164	13	2	11	0	0	0
MSR-CA	191	23	10	12	6	6	3	3	0
SGA	32	2	2	4	1	3	—	—	—
SOAPdenovo	246	25	31	17	1	16	8	1	7
Velvet	217	6	14	14	5	9	17	5	12
R. sphaeroides (4.60 Mb)									
ABySS	692	288	34	21	2	19	3	0	3
ALLPATHS-LG	218	150	37	6	0	6	0	0	0
Bambus2	189	149	363	5	0	5	2	0	2
CABOG	536	145	24	10	1	9	5	4	1
MSR-CA	807	179	32	10	1	9	5	2	3
SGA	336	116	4	4	0	4	-	-	-
SOAPdenovo	527	155	406	8	0	8	3	1	2
Velvet	413	148	27	8	0	8	6	6	7
Human chromosome 14 (88.29 Mb)									
ABySS	60,408	9987	678	13	6	7	9	9	0
ALLPATHS-LG	55,317	27,559	2558	101	44	57	45	0	45
Bambus2	64,869	17,141	5411	3266	1722	1544	143	37	106
CABOG	81,125	28,420	2883	149	46	103	597	389	208
MSR-CA	153,104	21,933	3082	1234	653	581	1068	210	858
SGA	70,976	15,483	681	150	90	60	—	—	—
SOAPdenovo	98,185	21,347	3902	1261	520	741	268	17	251
Velvet	79,399	17,505	4172	369	199	170	9156	3824	5332

Column headers are defined in the main text.

Assembly accuracy (3)

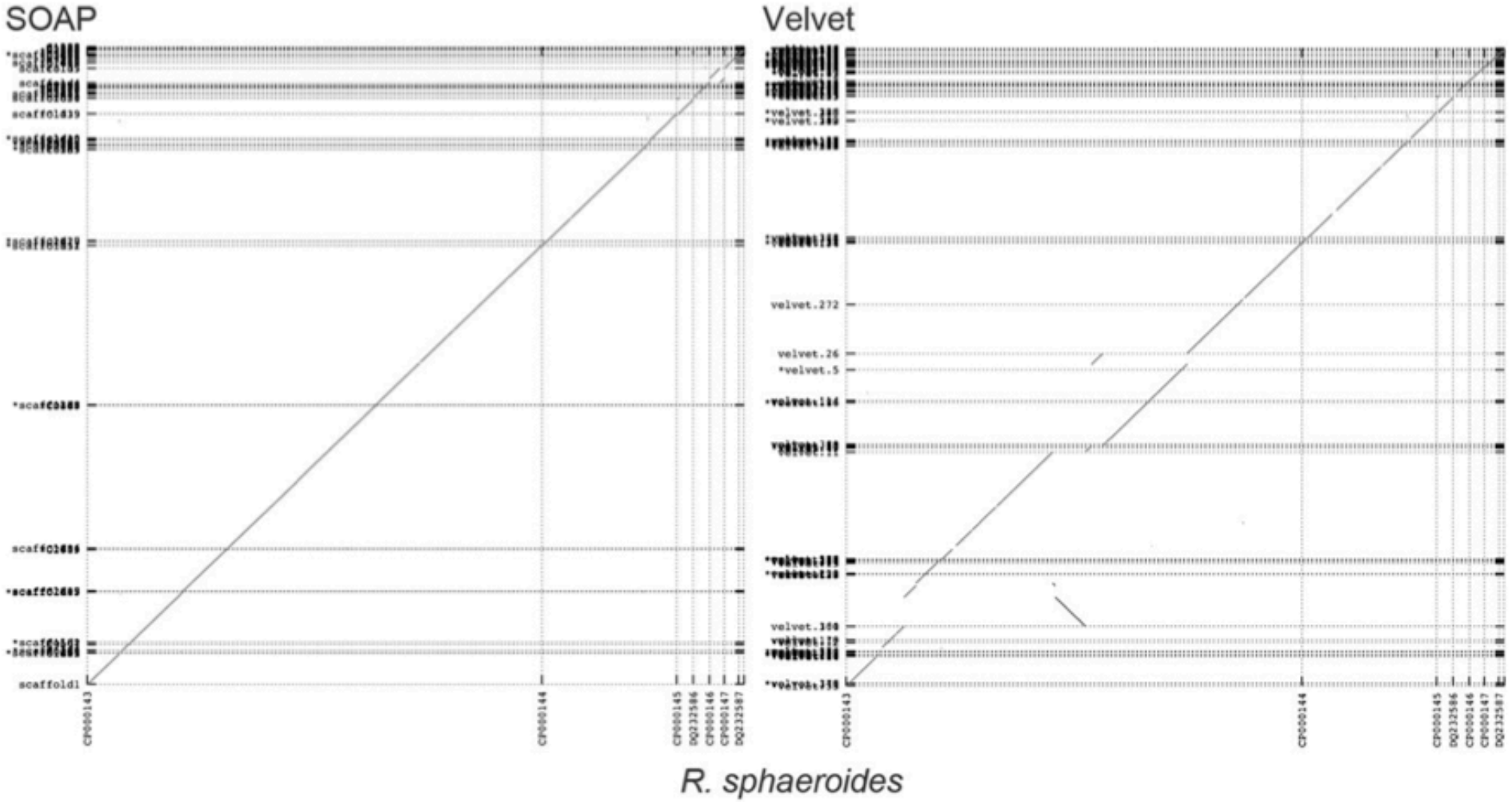

Figure 2. A dot-plot comparison of the SOAPdenovo and Velvet scaffolds of *R. sphaeroides*. The finished reference chromosomes are plotted on the *x*-axis and the assembly scaffolds on the *y*-axis. Dotted lines indicate scaffold or chromosome boundaries. The apparent rearrangement at the *top right* of the SOAPdenovo plot is an artifact of the circular reference plasmid.

Effect of the used libraries

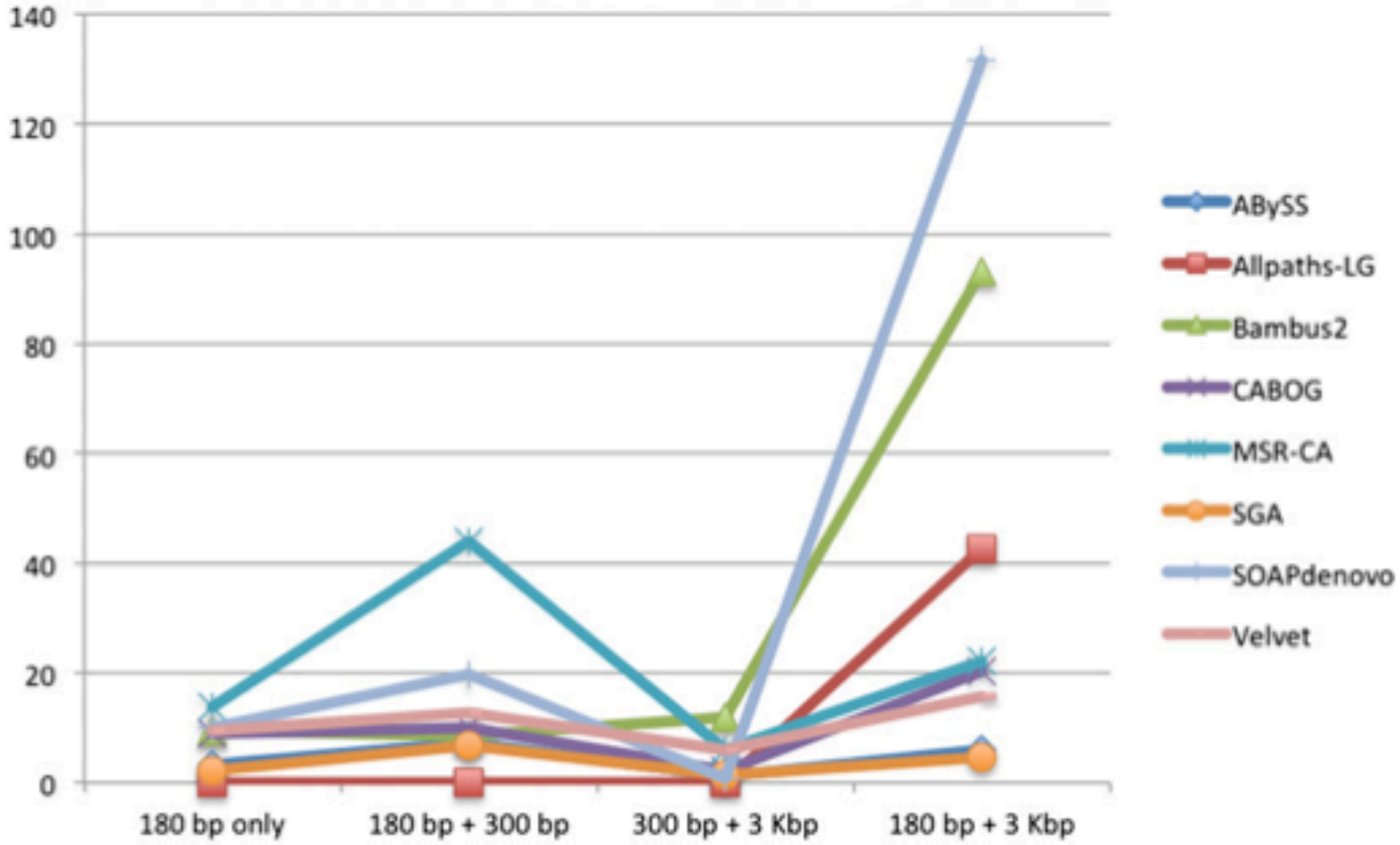

Figure 3. Assemblies of *R. sphaeroides* using four different combinations of paired-end libraries as input to the assemblers. Each run used either one library (180 bp only) or a different combination of two libraries from 180 to 3000 bp. Note that N50 values are uncorrected; see Table 3 for the true N50 sizes for the 180 bp + 3 kb combination, which are much lower in some instances; e.g., SOAPdenovo has a corrected N50 of 14.3 kb (rather than 131.7 kb) for assembly with the 180-bp and 3-kb libraries.

Indel errors among assemblers

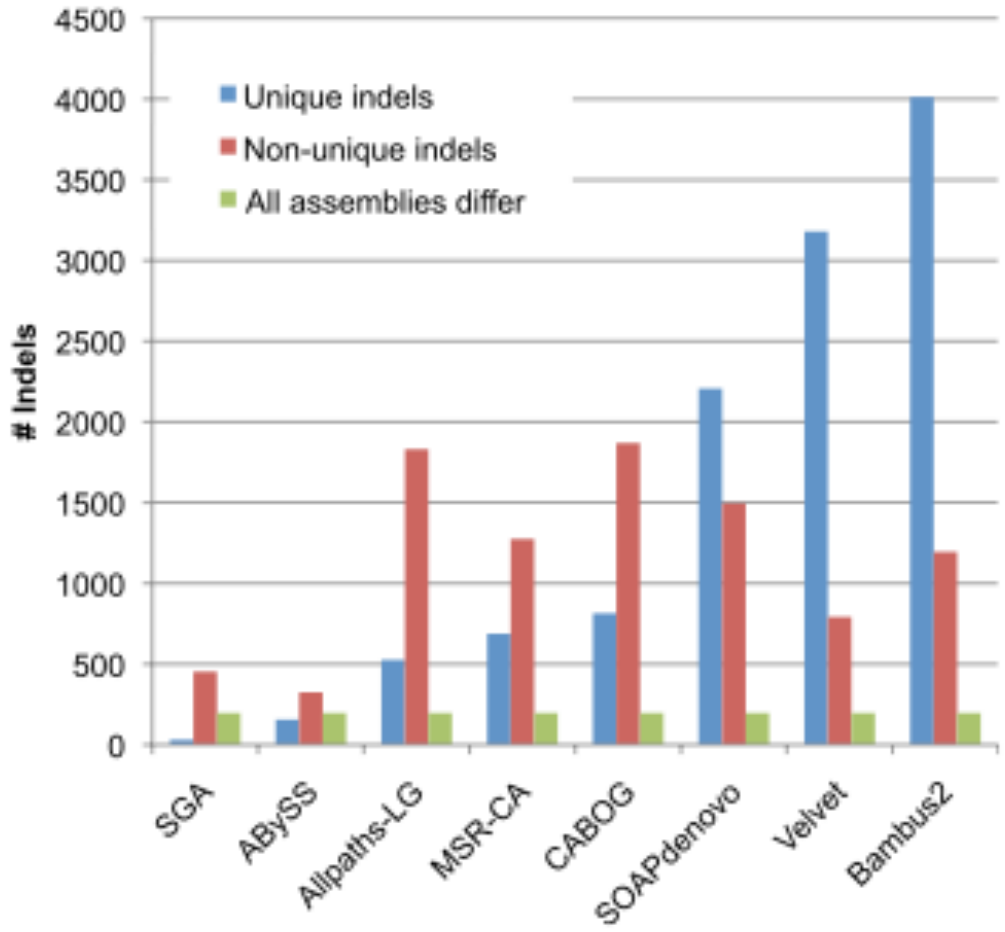

Figure 5. Comparison of insertion and deletion errors among all eight assemblers for human chromosome 14. (Blue) The indel errors >5 bp in length that are unique to each assembler. (Red bars) Indel errors made by at least one other assembler. (Green bars) Indels shared by all assemblers, which might represent true differences between the target genome and the reference.

Average N50 for 3 genomes

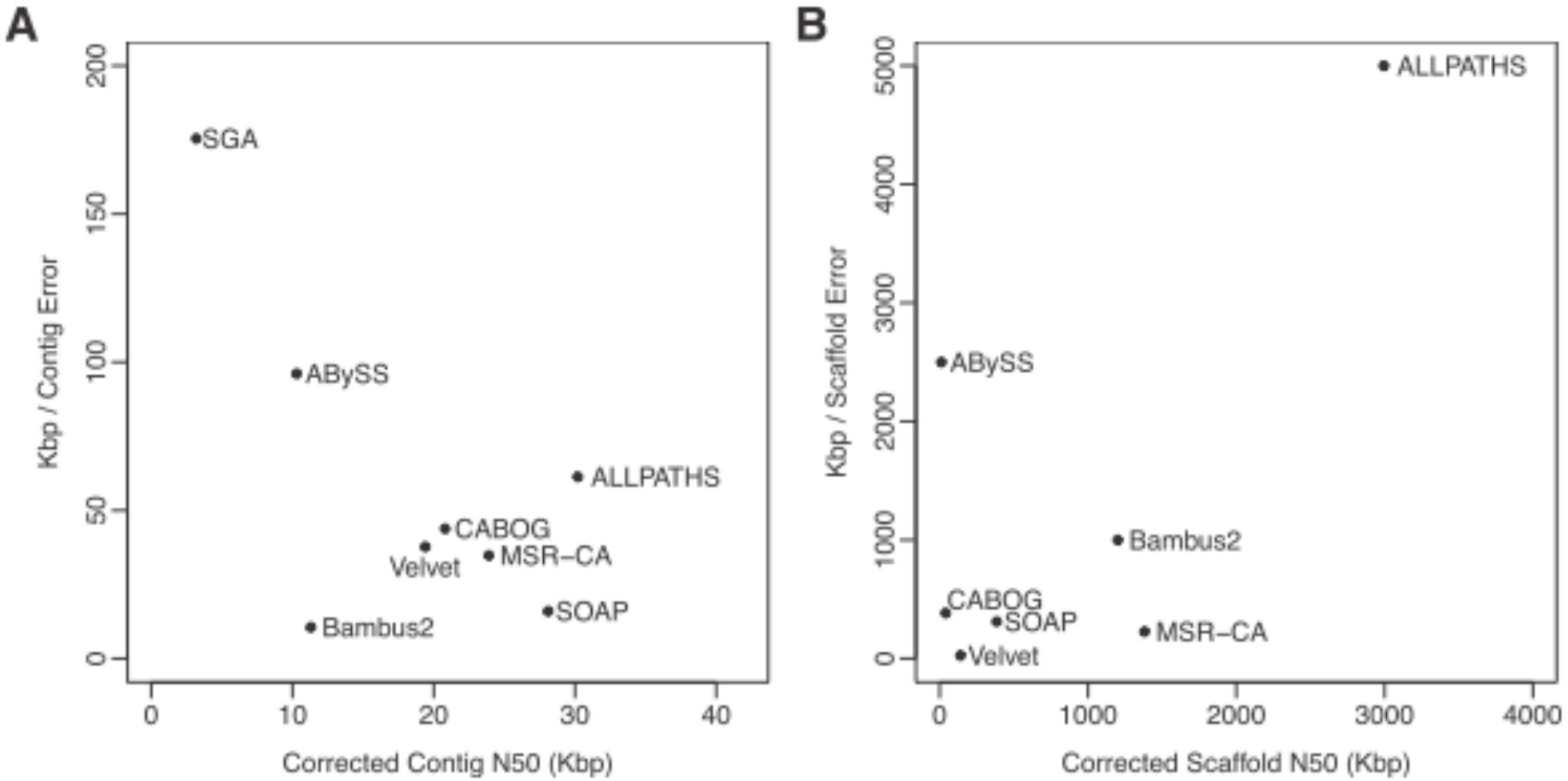

Figure 6. Average contig *(A)* and scaffold *(B)* sizes, measured by N50 values, versus error rates, averaged over all three genomes for which the true assembly is known: *S. aureus*, *R. sphaeroides*, and human chromosome 14. Errors (vertical axis) are measured as the average distance between errors, in kilobases. N50 values represent the size *N* at which 50% of the genome is contained in contigs/scaffolds of length *N* or larger. In both plots, the best assemblers appear in the *upper right*.

Remarks

- Read error correction is a key part of the assembly process.
- One of the most important variables in predicting assembly contiguity may be the genome itself, which is an element that cannot be controlled.
- Despite all efforts at error correction and repeat identification, assembly of a mammalian genome from NGS data remains an extremely challenging problem.
- For larger genomes, the choice of assemblers is often limited to those that will run without crashing.