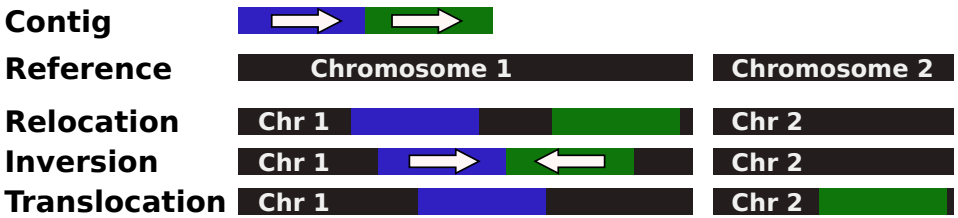
Supplement for "QUAST: Quality Assessment Tool for Genome Assemblies"

1 Supplementary Methods

1.1 Misassemblies

QUAST evaluates the number of misassemblies in an assembly, using Plantagora's definition. Plantagora defines a *misassembly breakpoint* as a position in the assembled contigs where the left flanking sequence aligns over 1 kb away from the right flanking sequence on the reference, or they overlap by more than 1 kb, or the flanking sequences align on opposite strands or different chromosomes. QUAST also generates a report with the number of misassemblies due to each of these reasons. QUAST's classification of all three cases is shown in Supplementary Fig. S1.



Supplementary Fig. S1: Classification of misassemblies. A *relocation* is a misassembly in which the left and right flanking sequences both align to the same chromosome on the reference, but they are either over 1 kb apart or they overlap by over 1 kb. As long as these distance thresholds are exceeded, the strands on which they align are not considered. An *inversion* is a misassembly in which the left and right flanking sequences both align to the same chromosome but on opposite strands, and further, they do not meet the distance thresholds to be classified as a relocation. A *translocation* is a misassembly in which the flanking sequences align on different chromosomes.

In addition, the QUAST report contains information about *local misassemblies*. We define a *local misassembly breakpoint* as a breakpoint that satisfies these conditions:

1. Two or more distinct alignments cover the breakpoint.
2. The gap between left and right flanking sequences is less than 1 kb.
3. The left and right flanking sequences both are on the same strand of the same chromosome of the reference genome.

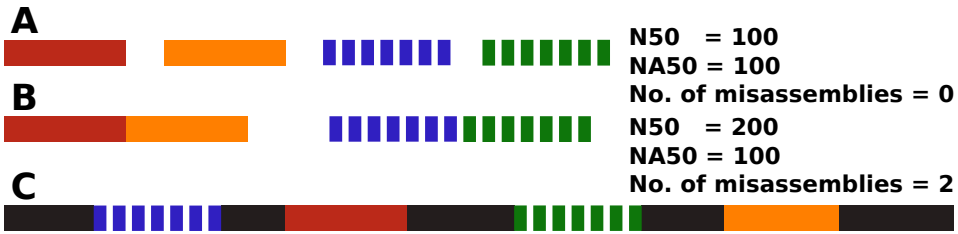
1.2 NAx and NGAx metrics

NAx (*A* stands for *aligned*, and *x* is a number from 0 to 100): This is a combination of the well-known Nx metric and Plantagora's number of misassemblies metric.

It is computed in two steps. First, we break the contigs into aligned blocks. If a contig has misassembly breakpoints, it is broken into multiple blocks at these breakpoints. Additionally, if there are unaligned regions within a contig, these regions are removed and the contig is split into blocks.

Next, we compute the ordinary Nx statistic on these blocks instead of on the original contigs. Note that the length of a block may differ slightly from the length of the reference region to which it aligns, due to small indels; we use the length of the block in the contig, rather than the aligned reference region. Thus, all four of Nx, NGx, NAx, and NGAx (defined below) are based on lengths of contigs or portions of contigs, ensuring that NAx $\leq$ Nx and NGAx $\leq$ NGx.

We illustrate the NAx metric in Supplementary Fig. S2. We compare the contigs of two assemblers, A and B, by focusing on N50 (to measure contig sizes) and the number of misassemblies. Suppose that assemblers A and B generate nearly identical assemblies, except that some contigs that are separate in assembly A are merged together in



Supplementary Fig. S2: Example of computing NA50. From top to bottom: assembly A (4 contigs), assembly B (2 contigs), alignment of contigs to reference genome C.

assembly B in order to increase the N50 statistic and make assembly B appear to be better. Of course, this comes at the expense of increasing the number of misassemblies. It may thus be confusing that assembly B appears to be better than assembly A in terms of N50, but worse than assembly A in terms of number of misassemblies.

The NA50 metric resolves this confusion. This metric splits merged contigs into separate blocks that are equal for both assemblers, so the value of NA50 is the same for assemblies A and B, while the number of misassemblies is higher in assembly B. Thus, assembly A is the better assembly.

The NGAx metric is defined similarly to NAx: first break the contigs into aligned blocks, but then compute the NGx statistic (instead of Nx) on the aligned blocks. An NGAx plot for *E. coli* is shown in Supplementary Fig. S9. IDBA-UD and SPAdes are the winners on the NGAx metric. SPAdes is the winner for $x > 30$, including NGA50 and NGA75, but IDBA-UD is the winner on smaller x. This is because the six largest contigs in IDBA-UD are larger than in SPAdes, but the remaining contigs are larger in SPAdes than in IDBA-UD. Unlike the standard Nx or NGx metrics, the NGAx metric breaks contigs at misassembly events; thus, misassemblies do not artificially inflate these metrics. It turns out that the misassemblies in SPAdes and IDBA-UD are in shorter contigs. and their higher number of misassemblies as compared to the Velvet-based assemblers does not diminish the use of this comparison. Supplementary Table 4 also shows that the total lengths of all misassembled contigs of SPAdes and IDBA-UD are rather small (46584 bp and 57334 bp, respectively).

2 QUAST performance

We benchmarked QUAST on three data sets: *E. coli*, *H. sapiens* chromosome 14, and *B. impatiens*. See Supplementary Table 1 for performance results. All benchmarking was done on a 4 CPU (Intel Xeon X7560 2.27GHz) computer. When running QUAST with a reference genome (*E. coli* and *H. sapiens*), the most time-consuming step is alignment, performed by the Nucmer aligner from MUMmer v3.23 (Kurtz *et al.*, 2004). When we ran QUAST on *B. impatiens* assemblies, for which a finished assembly does not exist at the time of this writing, the most time-consuming step was gene prediction, performed by GlimmerHMM (Majoros *et al.*, 2004). Note that alignment and gene prediction are parallelized, with each assembly processed by a separate thread, so QUAST will produce better results on computers with more CPUs.

Supplementary Table S1: QUAST performance

Genome	Genome size	No. of assemblies	Total time	Most time-consuming step
E. coli	4.6 Mb	7	0:01:00	0:00:21 (alignment)
H. sapiens, chr. 14	88.3 Mb	8	3:53:40	3:34:33 (alignment)
B. impatiens	approx. 250 Mb	4	1:18:48	1:16:21 (gene prediction)

3 QUAST report on single-cell *E. coli*

Genome size: 4.6 **Mb.** The reference genome is *E. coli* str. K-12 substr. MG1655, available at the NCBI website.
Number of genes: 4324**.** Gene annotations were taken from http://www.ecogene.org/.
Number of operons: 884**.** Operon annotations were taken from http://csbl1.bmb.uga.edu/OperonDB/displayNC.php?id=215.
Only contigs of length $\geq$ 200 **bp are used.**

The data set and some of its assemblies are taken from Chitsaz *et al.* (2011). The SPAdes and IDBA-UD assemblies are new.

Supplementary Table S2: Main table for *E. coli*.

Assembly	EULER-SR	E+V-SC	IDBA-UD	SOAPdenovo	SPAdes	Velvet	Velvet-SC
No. of contigs	610	396	**283**	817	532	310	617
NGA50	26580	32051	90607	16606	**99913**	22648	19791
Largest contig	140518	132865	**224018**	87533	211020	132865	121367
Total length	4306898	4555721	4734432	4183037	4975641	3517182	4556809
Genome fraction (%)	86.544	93.577	95.896	81.360	96.993	75.528	93.309
No. of misassemblies	19	**2**	9	6	11	**2**	**2**
No. of complete genes	3442	3816	4030	3060	**4071**	3121	3662

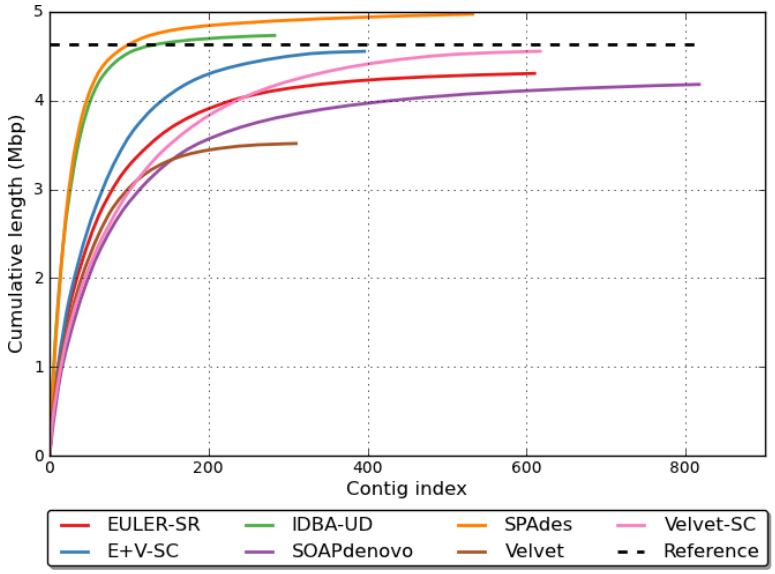
Supplementary Table S3: Misassemblies report for *E. coli*.

Assembly	EULER-SR	E+V-SC	IDBA-UD	SOAPdenovo	SPAdes	Velvet	Velvet-SC
No. of misassemblies	19	**2**	9	6	11	**2**	**2**
No. of relocations	16	**2**	6	5	5	**2**	**2**
No. of translocations	0	0	0	0	0	0	0
No. of inversions	3	**0**	3	1	6	**0**	**0**
No. of misassembled contigs	17	**2**	9	6	11	**2**	**2**
Misassembled contigs length	257468	23485	57334	96583	46584	**16522**	22359
No. of local misassemblies	89	6	15	216	8	**1**	2
No. of mismatches	390	89	**59**	4419	516	78	79
No. of indels	8916	35	**5**	437	34	44	44

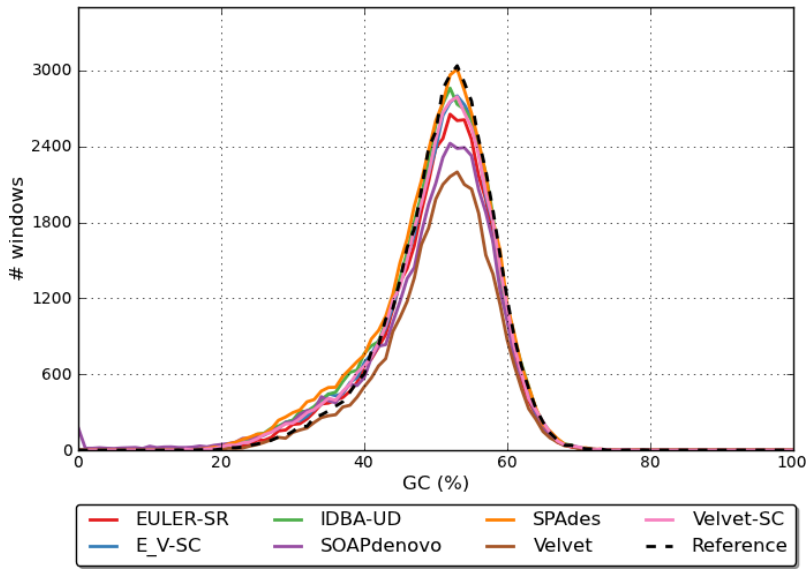
Supplementary Table S4: Extended report for *E. coli*.

Assembly	EULER-SR	E+V-SC	IDBA-UD	SOAPdenovo	SPAdes	Velvet	Velvet-SC
No. of contigs	610	396	**283**	817	532	310	617
Largest contig	140518	132865	**224018**	87533	211020	132865	121367
Total length	4306898	4555721	4734432	4183037	4975641	3517182	4556809
Reference length	4639675	4639675	4639675	4639675	4639675	4639675	4639675
N50	28697	32485	87102	22995	**96171**	32051	20445
NG50	26662	32051	90607	18468	**99933**	22648	19791
N75	11168	16969	**40948**	8601	38463	17097	8796
NG75	8392	15526	43468	4761	**44599**	1002	8051
No. of misassemblies	19	**2**	9	6	11	**2**	**2**
No. of misassembled contigs	17	**2**	9	6	11	**2**	**2**
Misassembled contigs length	257468	23485	57334	97258	46584	**16522**	22359
No. of unaligned contigs	89 + 29 part	67 + 3 part	74 + 17 part	165 + 37 part	226 + 59 part	**26 + 1 part**	82 + 3 part
Unaligned contigs length	218957	278346	346586	379343	462429	**75498**	277781
No. of ambiguously mapped contigs	10	22	24	**4**	30	21	27
Genome fraction (%)	86.544	93.577	95.896	81.360	**96.993**	75.528	93.309
Duplication ratio	1.021	**0.999**	1.002	1.009	1.025	**0.999**	1.001
GC (%)	50.21	50.08	49.92	49.92	49.69	50.47	50.08
Reference GC (%)	50.79	50.79	50.79	50.79	50.79	50.79	50.79
No. of mismatches per 100 kb	9.55	2.08	**1.34**	117.39	11.43	2.27	1.85
No. of indels per 100 kb	218.39	0.82	**0.11**	11.61	0.75	1.28	1.03
No. of genes	3442 + 408 part	3816 + 194 part	4030 + 127 part	3060 + 555 part	**4071 + 133 part**	3121 + 172 part	3662 + 348 part
No. of operons	641 + 174 part	739 + 100 part	800 + 62 part	527 + 235 part	**808 + 61 part**	605 + 100 part	650 + 190 part
NA50	28697	32485	87102	22249	**95929**	32051	20139
NGA50	26580	32051	90607	16606	**99913**	22648	19791
NA75	9943	16969	**40948**	6180	38463	17097	7745
NGA75	6991	15526	43468	2026	**44599**	335	7235

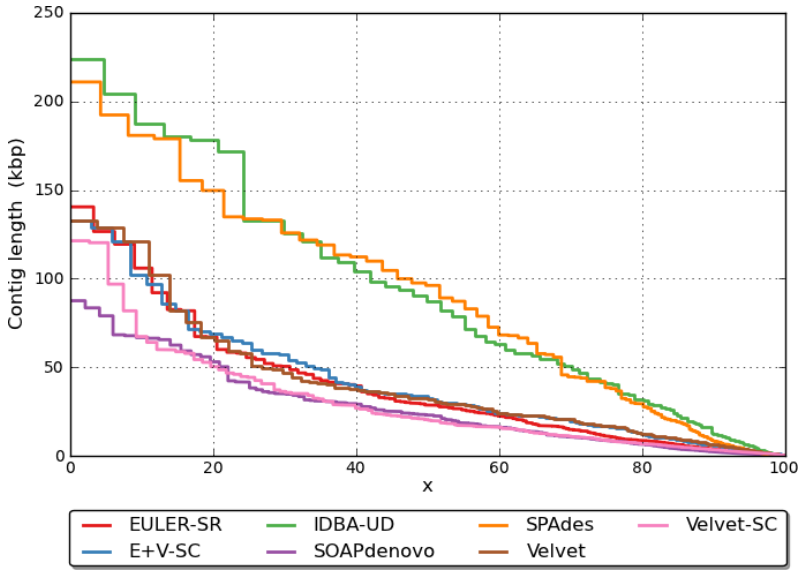
Basic plots for *E. coli*.



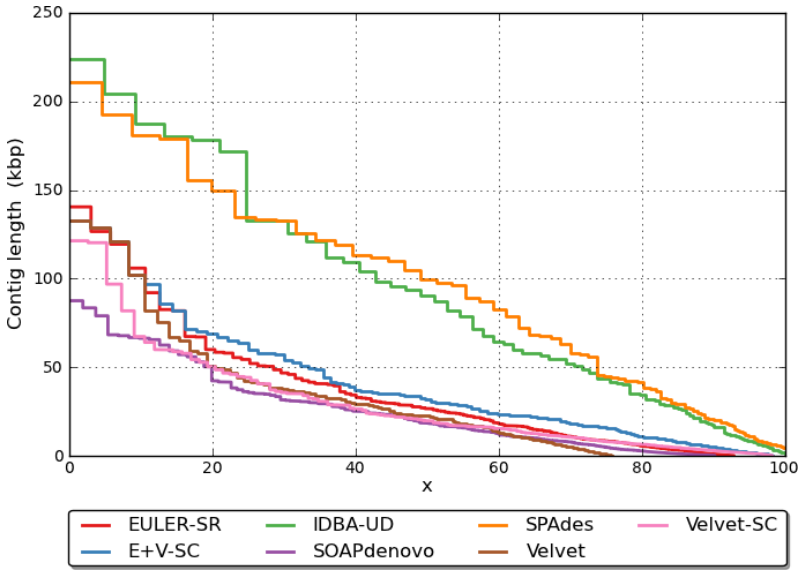
Supplementary Fig. S3: Cumulative length.



Supplementary Fig. S4: GC content.

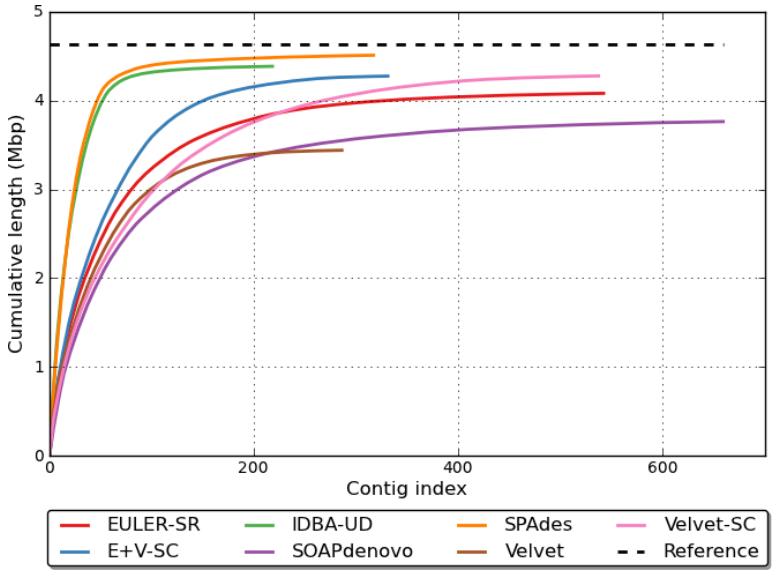


Supplementary Fig. S5: Nx.

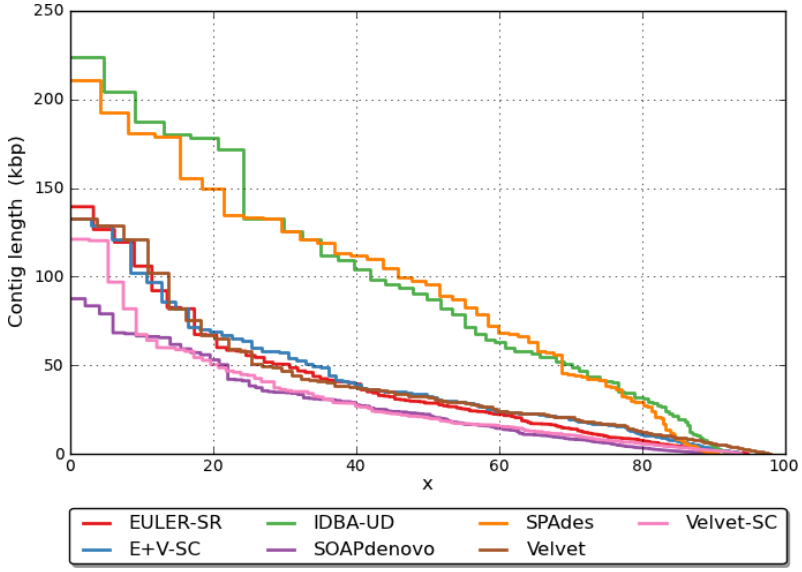


Supplementary Fig. S6: NGx.

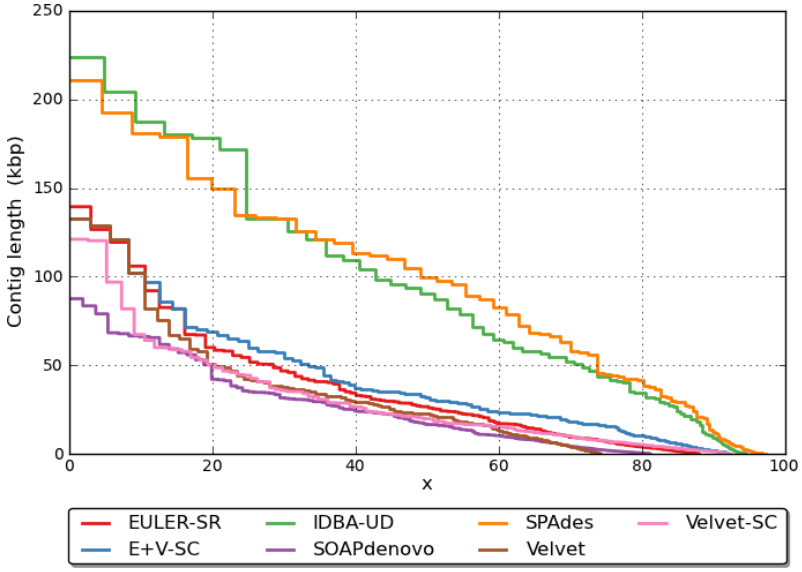
Aligned statistic plots for *E. coli*.



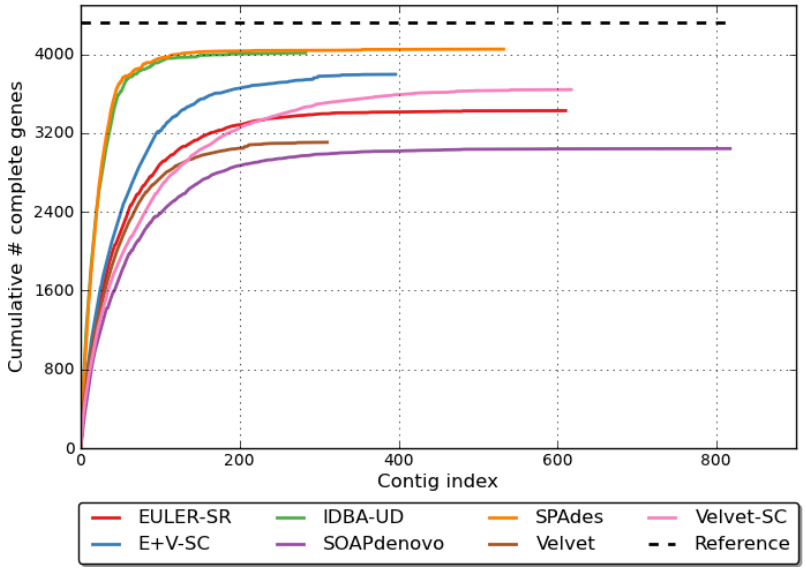
Supplementary Fig. S7: Cumulative length (aligned contigs).



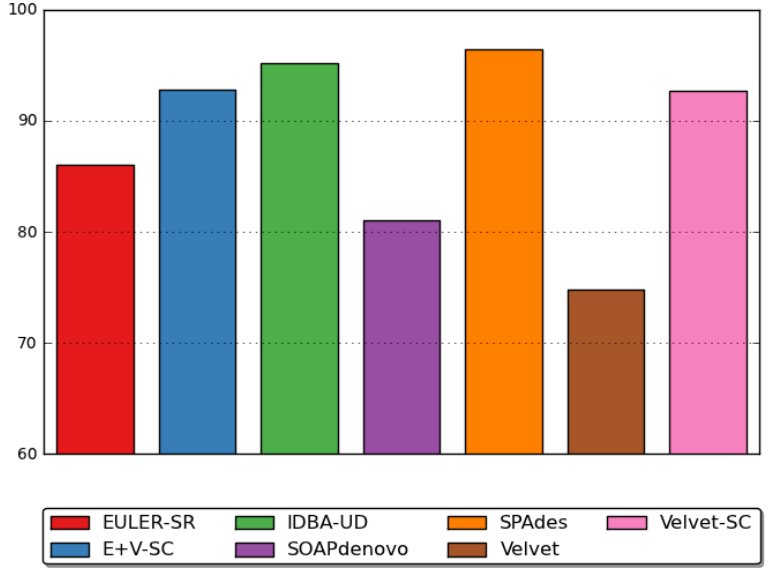
Supplementary Fig. S8: NAx.



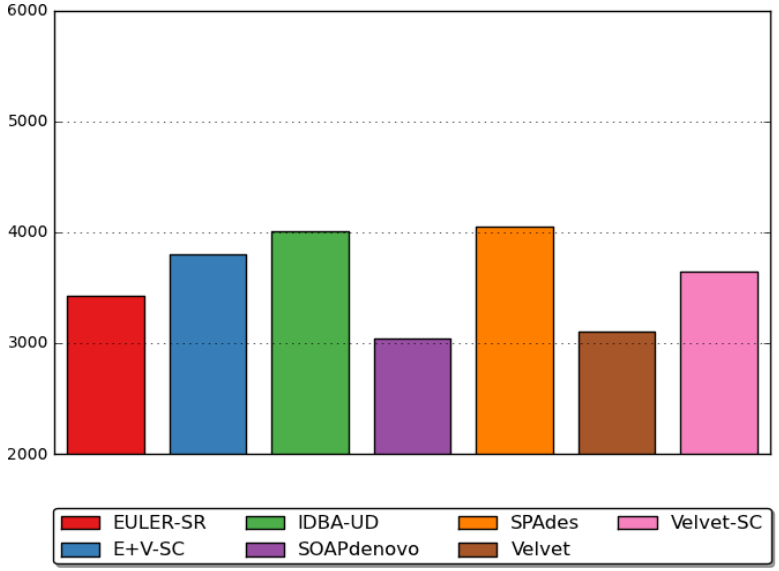
Supplementary Fig. S9: NGAx.



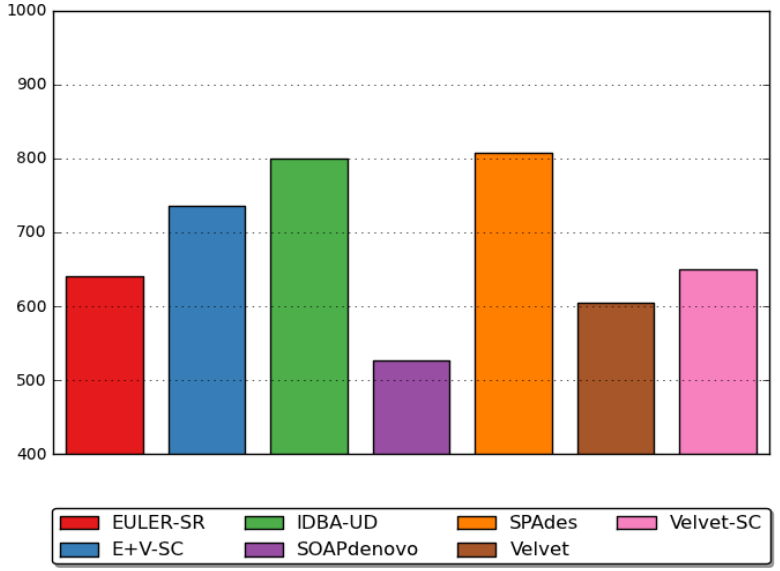
Supplementary Fig. S10: Cumulative No. of complete genes.



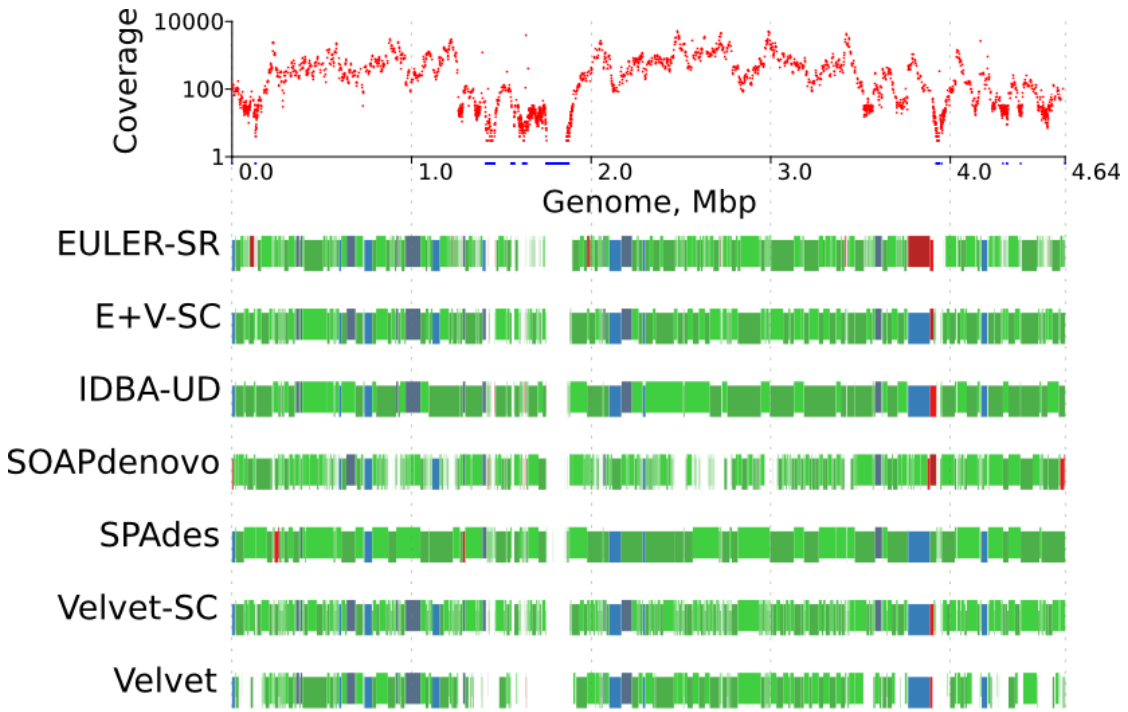
Supplementary Fig. S11: Genome fraction, %.



Supplementary Fig. S12: No. of complete genes.



Supplementary Fig. S13: No. of complete operons.



Supplementary Fig. S14: Contig alignment plot for *E. coli*.

Coverage plot (top):

- Coverage is averaged in 1000 bp bins.
- Blue dots below the x-axis indicate positions with zero coverage; this is not averaged over a bin.

Mapped contig plots:

- Contigs are mapped to the genome with Plantagora, and split into one or more blocks at misassembly breakpoints. Each colored rectangle represents a block.
- Rectangles are staggered vertically for better plot comprehension.

Contig colors:

- **Green bars**: contigs without misassemblies.
- **Blue bars**: contigs with similar boundaries in at least half of all assemblies.
- **Red bars**: blocks after splitting misassembled contigs at misassembly breakpoints.
- **Orange bars**: blocks of misassembled contigs when the boundaries of the parts are similar in at least half of all assemblies.

4 QUAST report on *H. sapiens*, chromosome 14

Genome size: 88.3 **Mb.** The reference genome is *H. sapiens* chr. 14 genomic contig, GRCh37.p10, available at the NCBI website.
Number of genes: 1449**.** Gene annotations were taken from http://genome.ucsc.edu/cgi-bin/hgTables.
Only contigs of length $\geq$ 200 **bp are used.**

The data set and all of its assemblies are taken from Salzberg *et al.* (2011).

Supplementary Table S5: Main table for *H. sapiens*, chromosome 14.

Assembly	ABySS	Allpaths-LG	Bambus2	CABOG	MSR-CA	SGA	SOAPdenovo	Velvet
No. of contigs	51924	4529	13592	**3361**	30103	56939	21818	45564
NGA50	2024	36284	5061	**44934**	4820	2710	15091	2289
Largest contig	30053	240773	**736657**	296904	53925	30350	147494	27872
Total length	73341066	84435699	68243005	86232753	83291373	82375466	92488595	74740589
Genome fraction (%)	82.228	95.255	79.334	**98.248**	94.111	93.557	98.094	84.415
No. of misassemblies	**20**	115	3372	174	2423	219	5311	385
No. of complete genes	205	669	311	**718**	324	259	437	212

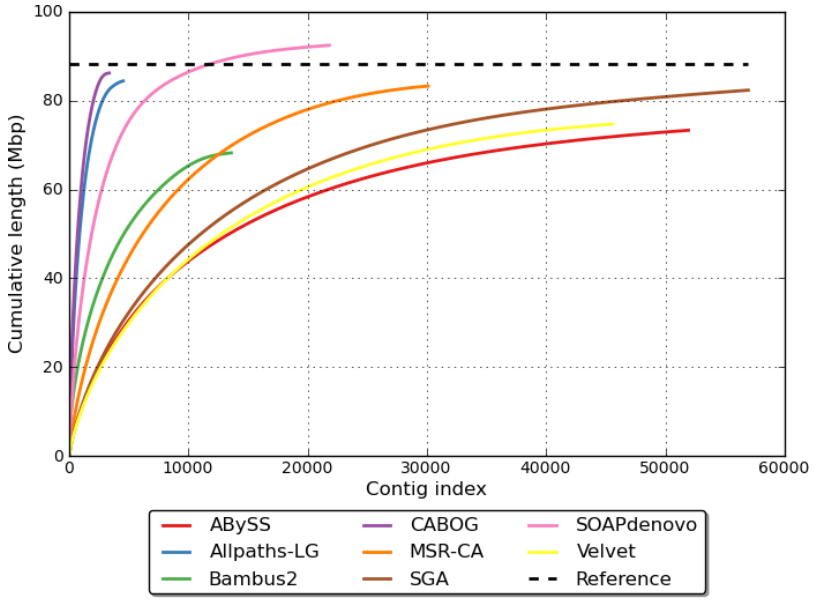
Supplementary Table S6: Misassemblies report for *H. sapiens*, chromosome 14.

Assembly	ABySS	Allpaths-LG	Bambus2	CABOG	MSR-CA	SGA	SOAPdenovo	Velvet
No. of misassemblies	**20**	115	3372	174	2423	219	5311	385
No. of relocations	**20**	114	3367	170	2411	219	5171	378
No. of translocations	0	0	0	0	0	0	0	0
No. of inversions	**0**	1	5	4	12	**0**	140	7
No. of misassembled contigs	**20**	100	1069	118	2297	197	2700	358
Misassembled contigs length	**98963**	1771098	12946575	5344856	6945874	287139	41103208	769621
No. of local misassemblies	158	244	2256	500	622	**31**	8180	867
No. of mismatches	**60453**	78645	65375	80517	146027	70823	104783	78110
No. of indels	**20672**	64632	93721	69534	55058	25345	75116	84544

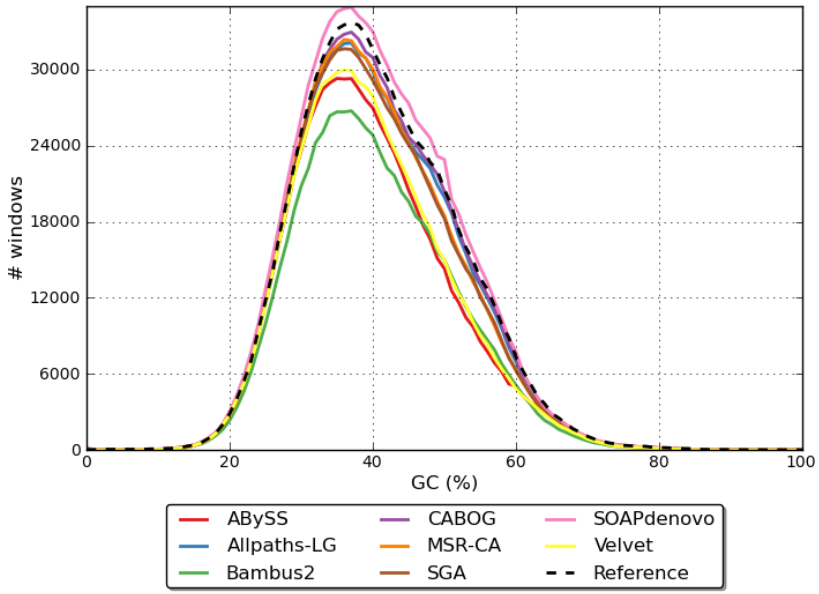
Supplementary Table S7: Extended report for *H. sapiens*, chromosome 14.

Assembly	ABySS	Allpaths-LG	Bambus2	CABOG	MSR-CA	SGA	SOAPdenovo	Velvet
No. of contigs	51924	4529	13592	**3361**	30103	56939	21818	45564
Largest contig	30053	240773	**736657**	296904	53925	30350	147494	27872
Total length	73341066	84435699	68243005	86232753	83291373	82375466	92488595	74740589
Reference length	88289540	88289540	88289540	88289540	88289540	88289540	88289540	88289540
N50	2776	38359	8430	**46694**	5345	2981	15762	2887
NG50	2024	36530	5851	**45336**	4906	2710	16662	2294
N75	1230	20271	4224	**26111**	2538	1376	6516	1435
NG75	550	17552	1336	**24395**	2112	1086	7626	784
No. of misassemblies	**20**	115	3372	174	2423	219	5311	385
No. of misassembled contigs	**20**	100	1069	118	2297	197	2700	358
Misassembled contigs length	**98963**	1771098	12946575	5344856	6945874	287139	41103208	769621
No. of unaligned contigs	35+86 part	**5+45 part**	211+426 part	19+39 part	319+1804 part	167+443 part	82+251 part	900+519 part
Unaligned contigs length	23686	**8529**	193119	19322	566781	63785	167308	564483
No. of ambiguously mapped contigs	168	6	15	3	12	**0**	**0**	**0**
Genome fraction (%)	82.228	95.255	79.334	**98.248**	94.111	93.557	98.094	84.415
Duplication ratio	1.015	1.004	**0.978**	1.008	0.998	1.003	1.078	1.001
GC (%)	39.69	40.77	40.20	40.80	40.45	40.46	40.80	39.78
Reference GC (%)	40.89	40.89	40.89	40.89	40.89	40.89	40.89	40.89
No. of mismatches per 100 kb	**82.45**	93.20	96.19	93.45	176.74	86.05	114.20	105.41
No. of indels per 100 kb	**28.20**	76.60	137.90	80.70	66.64	30.80	81.87	114.10
No. of genes	205+1203 part	669+766 part	311+1098 part	**718+729 part**	324+1120 part	259+1178 part	437+1008 part	212+1203 part
NA50	2776	37827	7066	**45936**	5232	2981	14188	2883
NGA50	2024	36284	5061	**44934**	4820	2710	15091	2289
NA75	1230	20014	3766	**26111**	2401	1376	5512	1433
NGA75	550	17382	1228	**24395**	2002	1086	6535	779

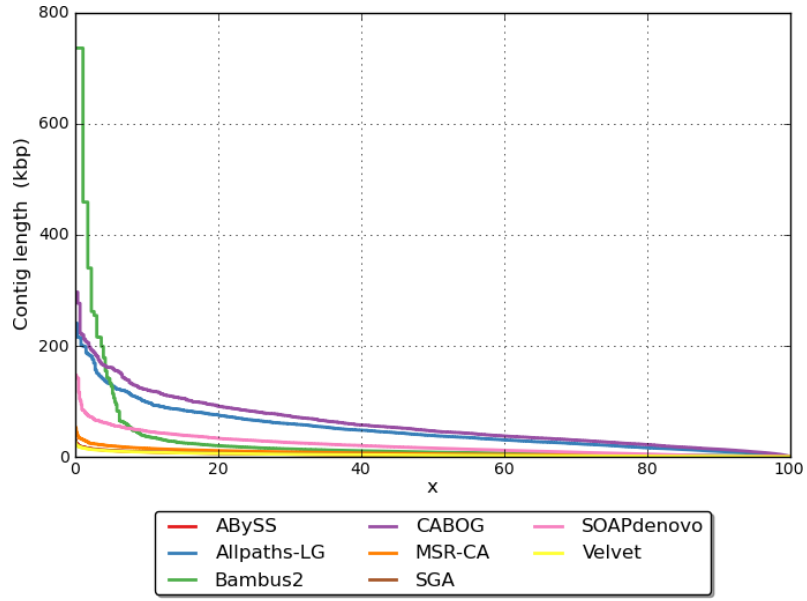
Basic plots for *H. sapiens*, chromosome 14.



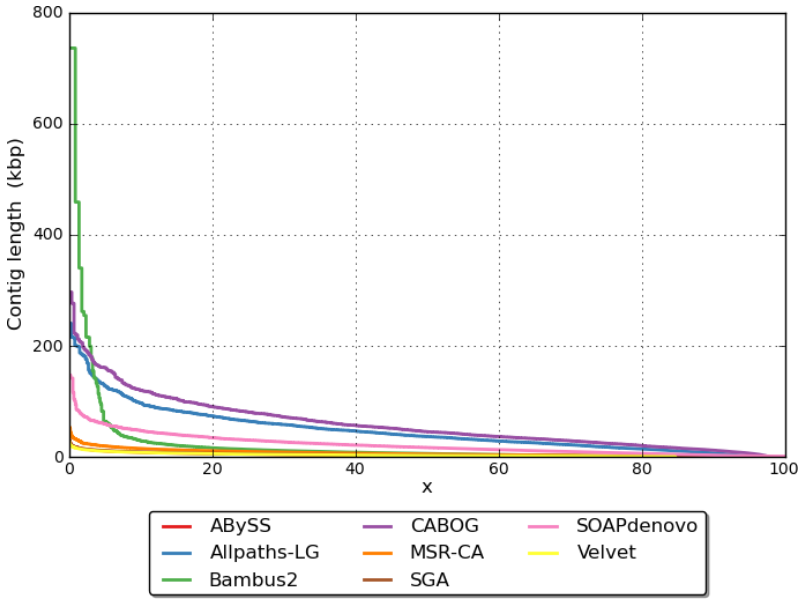
Supplementary Fig. S15: Cumulative length.



Supplementary Fig. S16: GC content.

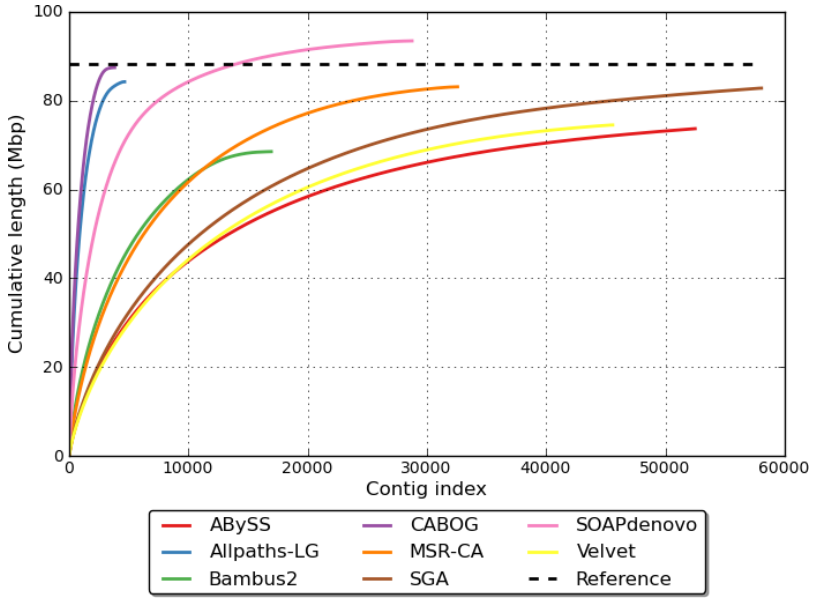


Supplementary Fig. S17: Nx.

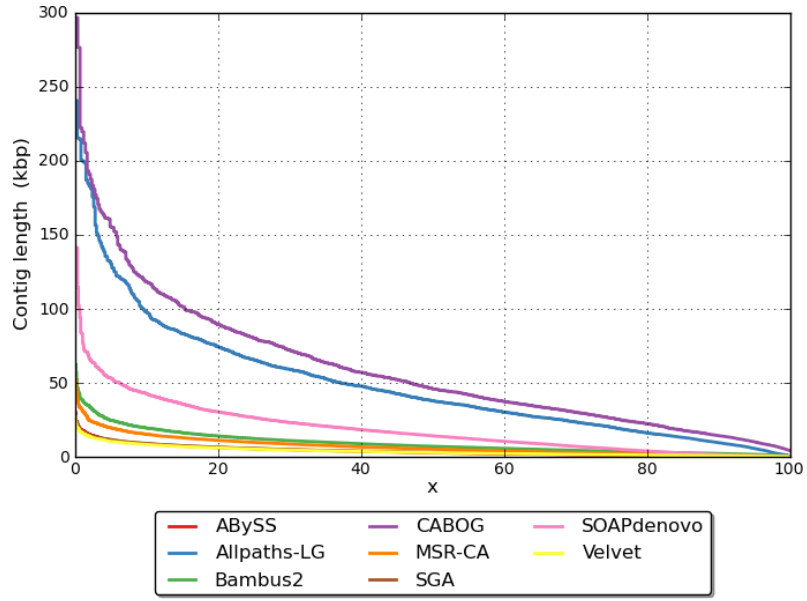


Supplementary Fig. S18: NGx.

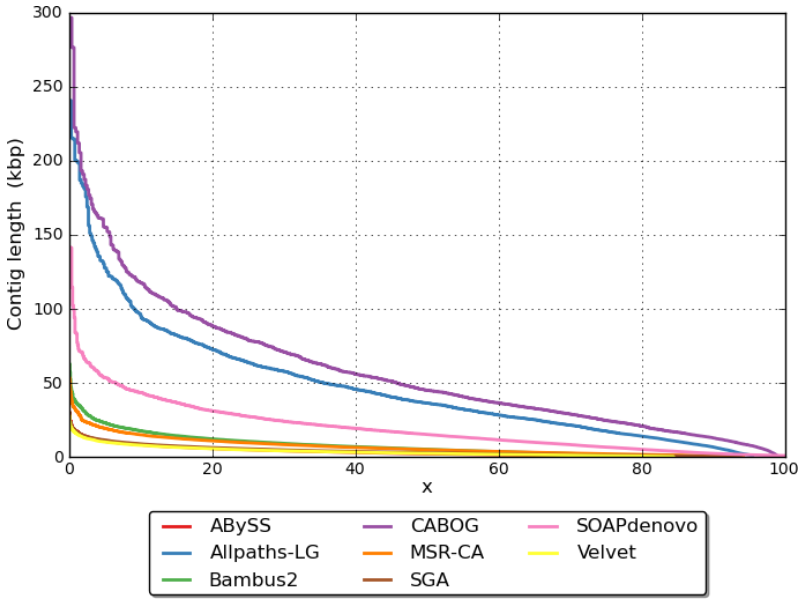
Aligned statistic plots for *H. sapiens*, chromosome 14.



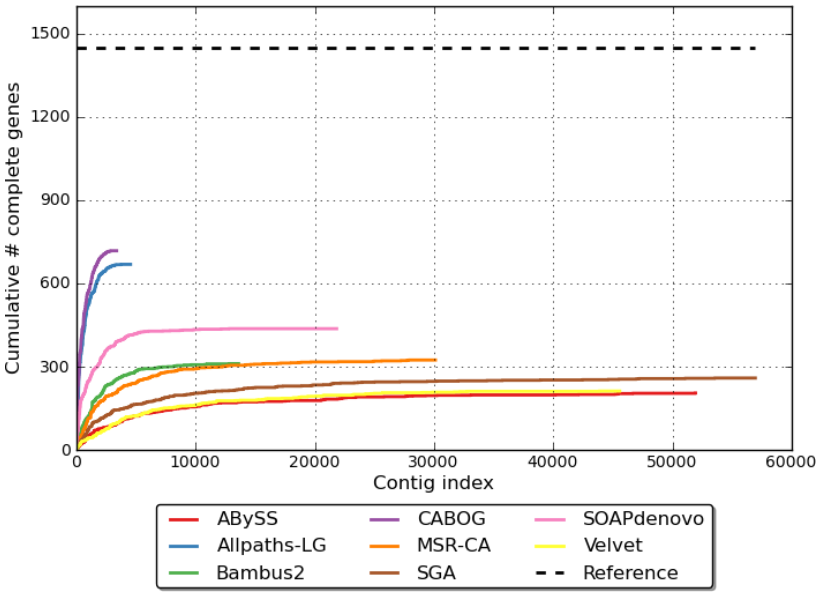
Supplementary Fig. S19: Cumulative length (aligned contigs).



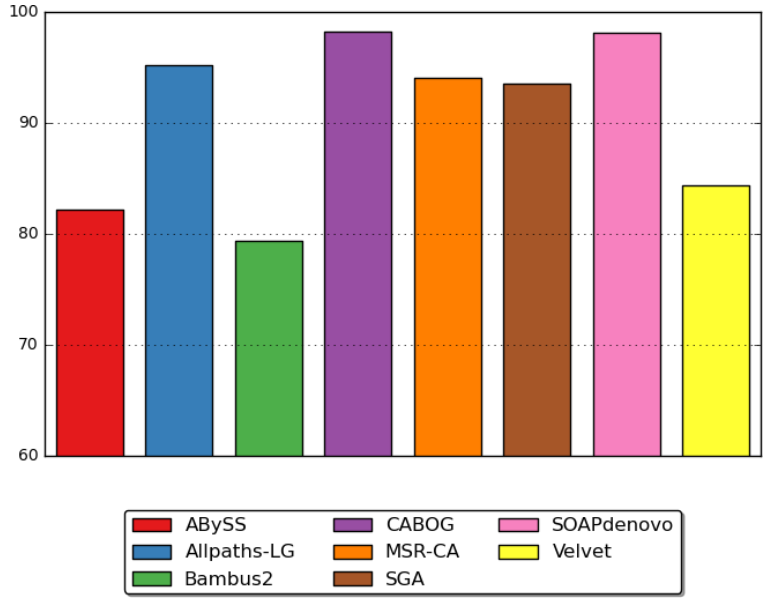
Supplementary Fig. S20: NAx.



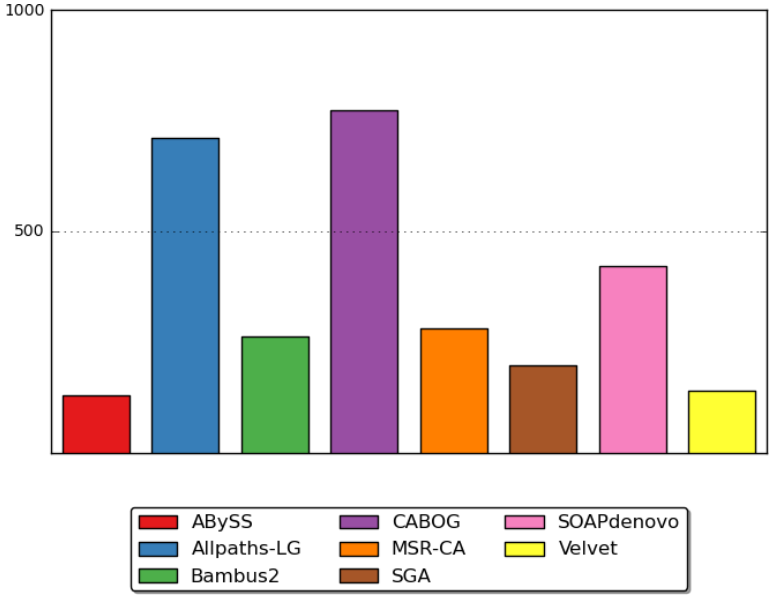
Supplementary Fig. S21: NGAx.



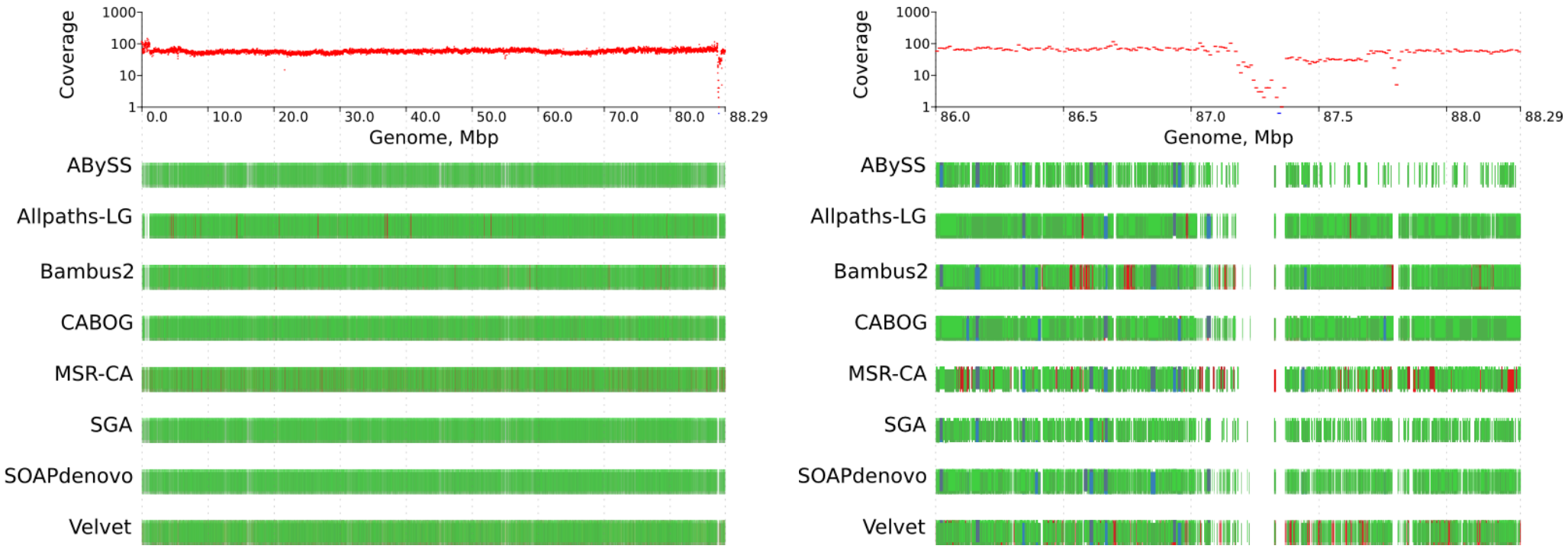
Supplementary Fig. S22: Cumulative No. of complete genes.



Supplementary Fig. S23: Genome fraction, %.



Supplementary Fig. S24: No. of complete genes.



Supplementary Fig. S25: Contig alignment plots for *H. sapiens*, chromosome 14. *Left:* The whole chromosome. *Right:* Region at the end of the chromosome.

Coverage plot (top):

- Coverage is averaged in 1000 bp bins.
- Blue dots below the x-axis indicate positions with zero coverage; this is not averaged over a bin.

Mapped contig plots:

- Contigs are mapped to the genome with Plantagora, and split into one or more blocks at misassembly breakpoints. Each colored rectangle represents a block.
- Rectangles are staggered vertically for better plot comprehension.

Contig colors:

- **Green bars**: contigs without misassemblies.
- **Blue bars**: contigs with similar boundaries in at least half of all assemblies.
- **Red bars**: blocks after splitting misassembled contigs at misassembly breakpoints.
- **Orange bars**: blocks of misassembled contigs when the boundaries of the parts are similar in at least half of all assemblies.

5 QUAST report on *B. impatiens*

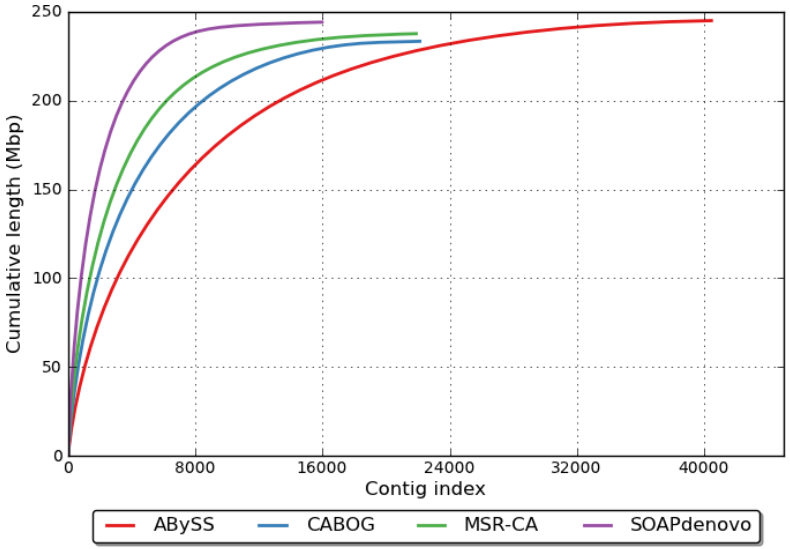
Estimated genome size: 250 **Mb (used for NGx statistics).**
Only contigs of length $\geq$ 200 **bp are used.**
Since there is no finished genome for *B. impatiens*, reference-based statistics cannot be computed. However, we ran GlimmerHMM (Majoros *et al.*, 2004) to find potential genes.

The data set and all of its assemblies are taken from Salzberg *et al.* (2011).

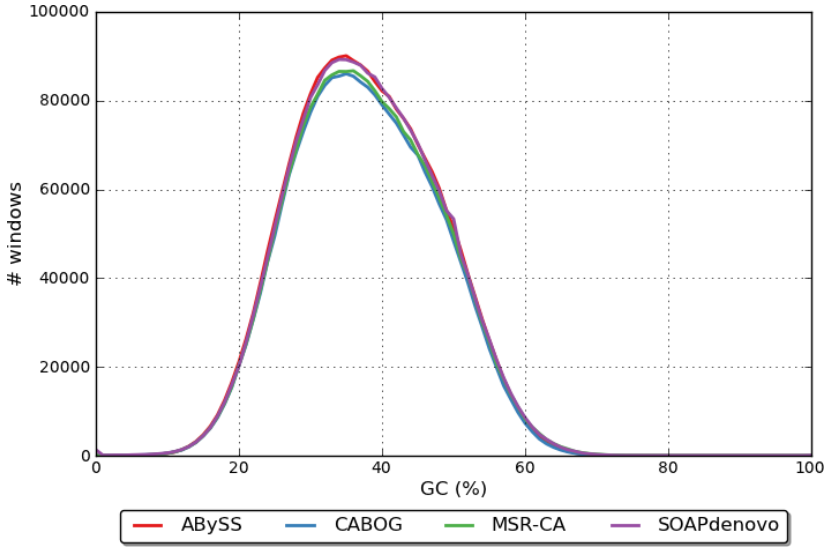
Supplementary Table S8: Extended report for *B. impatiens*.

Assembly	ABySS	CABOG	MSR-CA	SOAPdenovo
No. of contigs	40451	22107	21885	**15957**
Largest contig	177883	297795	311566	**523622**
Total length	245097922	233471993	237757432	244248450
Estimated reference length	250000000	250000000	250000000	250000000
N50	14791	25852	34797	**58901**
NG50	14383	23515	32431	**57117**
N75	6767	11811	15431	**26464**
NG75	6365	9467	12779	**24625**
GC (%)	37.65	37.60	37.77	37.75
No. of predicted genes (unique)	**39905**	34113	36753	34319
No. of predicted genes ($\geq$ 0 bp)	**41421**	34143	36863	34388
No. of predicted genes ($\geq$ 300 bp)	**23504**	20779	22912	22645
No. of predicted genes ($\geq$ 1500 bp)	**6453**	6034	6240	6407
No. of predicted genes ($\geq$ 3000 bp)	2917	2775	2919	**2992**

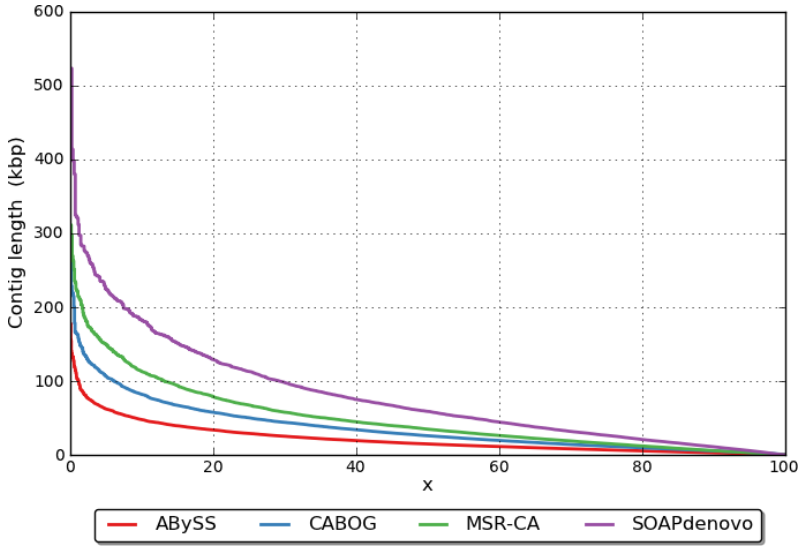
Basic plots for *B. impatiens*.



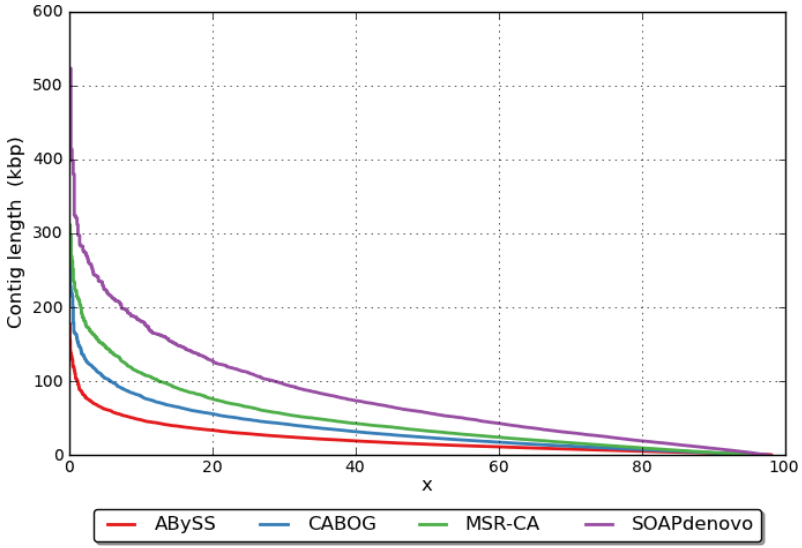
Supplementary Fig. S26: Cumulative length.



Supplementary Fig. S27: GC content.



Supplementary Fig. S28: Nx.



Supplementary Fig. S29: NGx.

6 List of used assemblers

We used thirteen assemblers in the comparisons:

- ABySS (Simpson *et al.*, 2009)
- Allpaths-LG (Gnerre *et al.*, 2011)
- Bambus2 (Koren *et al.*, 2011)
- CABOG (Miller *et al.*, 2008)
- EULER-SR (Pevzner *et al.*, 2001)
- IDBA-UD (Peng *et al.*, 2012)
- MSR-CA (http://www.genome.umd.edu/)
- SGA (Simpson and Durbin, 2012)
- SOAPdenovo (Li *et al.*, 2010)
- SPAdes (Bankevich *et al.*, 2012)
- Velvet (Zerbino and Birney, 2008)
- Velvet-SC, EULER+Velvet-SC (Chitsaz *et al.*, 2011)

References

Bankevich, A. *et al.* (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol*, **19**(5), 455–477.

Chitsaz, H. *et al.* (2011). Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nature Biotechnol*, **29**(10), 915–921.

Gnerre, S. *et al.* (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A.*, **108**(4), 1513–1518.

Koren, S. *et al.* (2011). Bambus 2: scaffolding metagenomes. *Bioinformatics*, **27**(21), 2964–2971.

Kurtz, S. *et al.* (2004). Versatile and open software for comparing large genomes. *Genome Biol*, **5**(2), R12.

Li, R. *et al.* (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, **20**(2), 265–272.

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics*, **20**(16), 2878–2879.

Miller, J. R. *et al.* (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, **24**(24), 2818–2824.

Peng, Y. *et al.* (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**(11), 1–8.

Pevzner, P. A. *et al.* (2001). An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A*, **98**(17), 9748–9753.

Salzberg, S. L. *et al.* (2011). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res*, **22**(3), 557–567.

Simpson, J. *et al.* (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res*, **19**(6), 1117–1123.

Simpson, J. T. and Durbin, R. (2012). Efficient de novo assembly of large genomes using compressed data structures. *Genome Res*, **22**(3), 549–556.

Zerbino, D. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, **18**(5), 821–829.