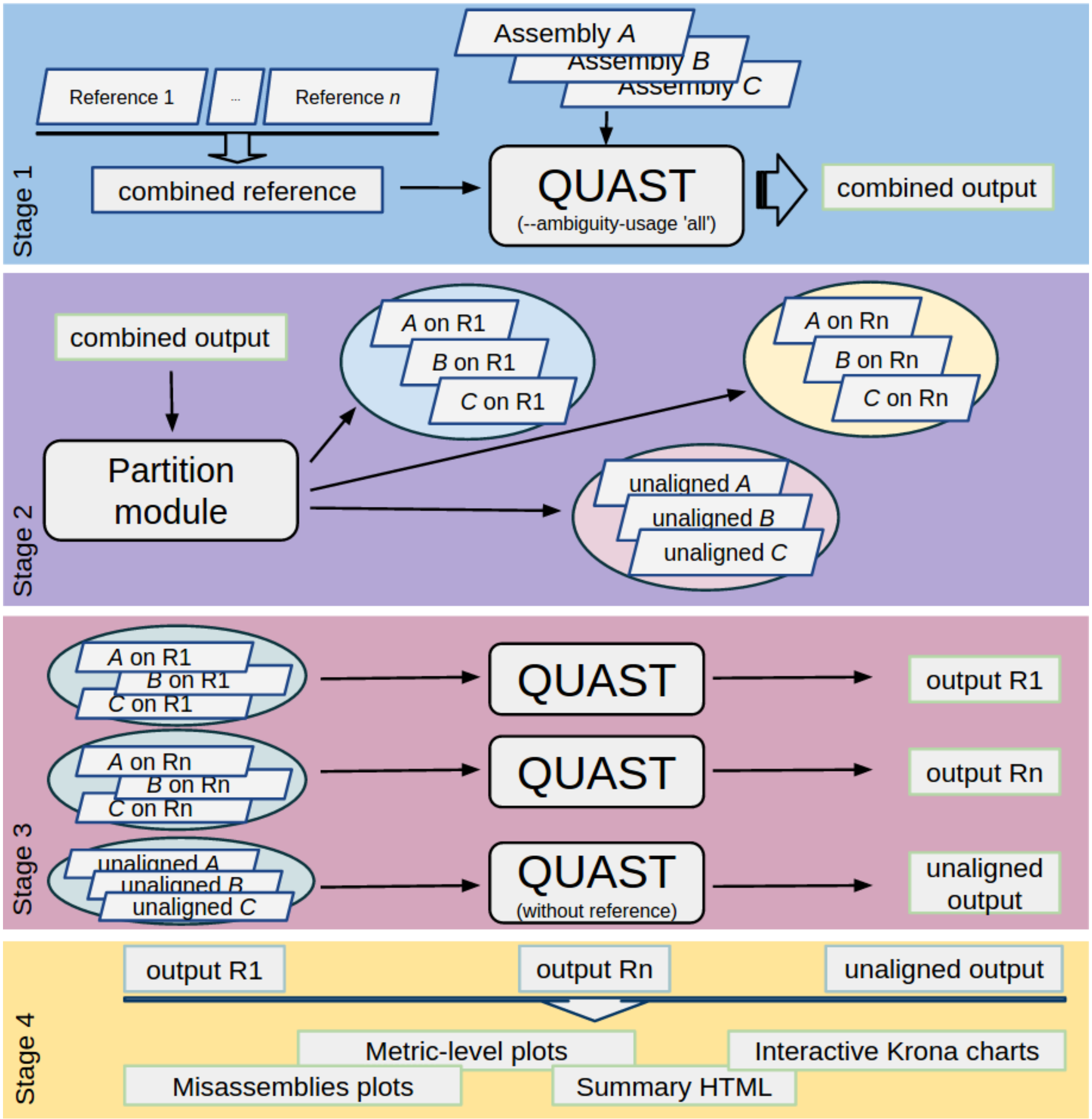
Supplementary Material for "MetaQUAST: evaluation of metagenome assemblies"

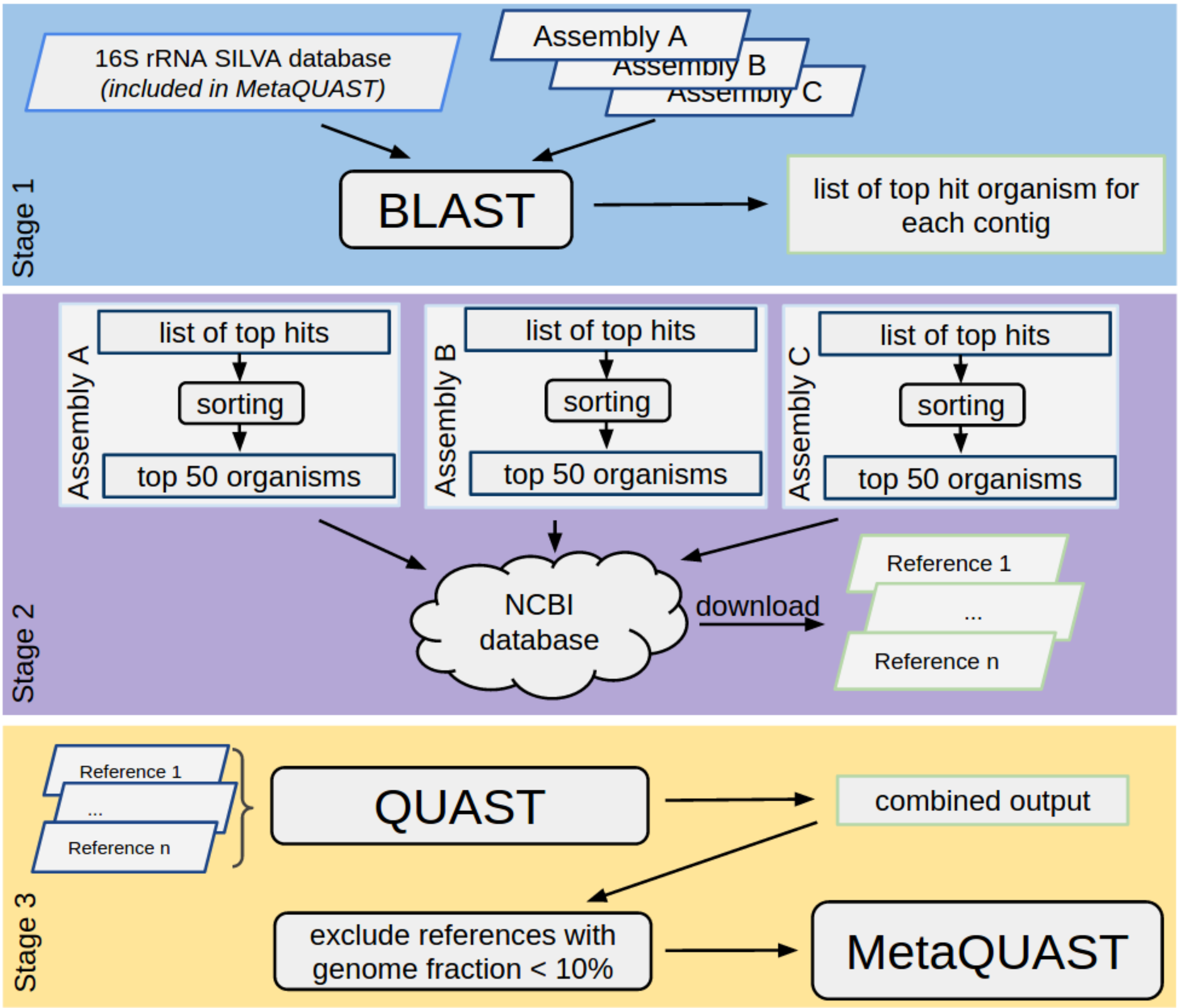
1 Supplementary Methods

1.1 Pipeline overview

MetaQUAST pipeline for reference-based evaluation is in Supplementary Fig. S1, pipeline for *de novo* evaluation is in Supplementary Fig. S2.



Supplementary Fig. S1: MetaQUAST pipeline for reference-based evaluation. Stage 1: All reference genomes are concatenated into a single file (*combined reference*). QUAST (Gurevich *et al.*, 2013) is launched with all input assemblies and the *combined reference*. We force QUAST to report all good ambiguous alignments per each contig instead of one (default behaviour) using "–ambiguity-usage all" option. Stage 2: MetaQUAST partitions all contigs into bins aligned to each input reference, plus a separate group for unaligned contigs. Stage 3: MetaQUAST launches QUAST for each input reference separately, feeding it with a corresponding group of contigs. The group of unaligned contigs is processed in QUAST *without reference* mode. Stage 4: The results of all QUAST runs are summarized. MetaQUAST makes plots and text reports for each key metric, plus a bird-eye overview of all references and assemblies features on one page - Summary HTML. When references are detected and downloaded by MetaQUAST (see Supplementary Fig. S2), it also creates a set of interactive Krona charts (Ondov *et al.*, 2011) based on the detected taxonomic classification.



Supplementary Fig. S2: MetaQUAST pipeline for *de novo* evaluation. Stage 1: BLASTn (Camacho *et al.*, 2009) aligns all assemblies to the 16S rRNA sequences from the SILVA database (Quast *et al.*, 2012). For each contig, we choose the top hit organism with the maximal BLAST score. Stage 2: 50 organisms with the maximal BLAST scores are picked for each assembly. MetaQUAST attempts to find and download reference genomes for all these organisms from the NCBI database. Stage 3: All downloaded files are concatenated into a single reference file. QUAST is used with this file to estimate individual references contig coverage fraction by each assembly. Then MetaQUAST excludes reference genomes with a low genome fraction (less than 10%) from further analysis, and launches the whole pipeline for the reference-based evaluation using the remaining files.

1.2 Structural variants detection and misassemblies refinement

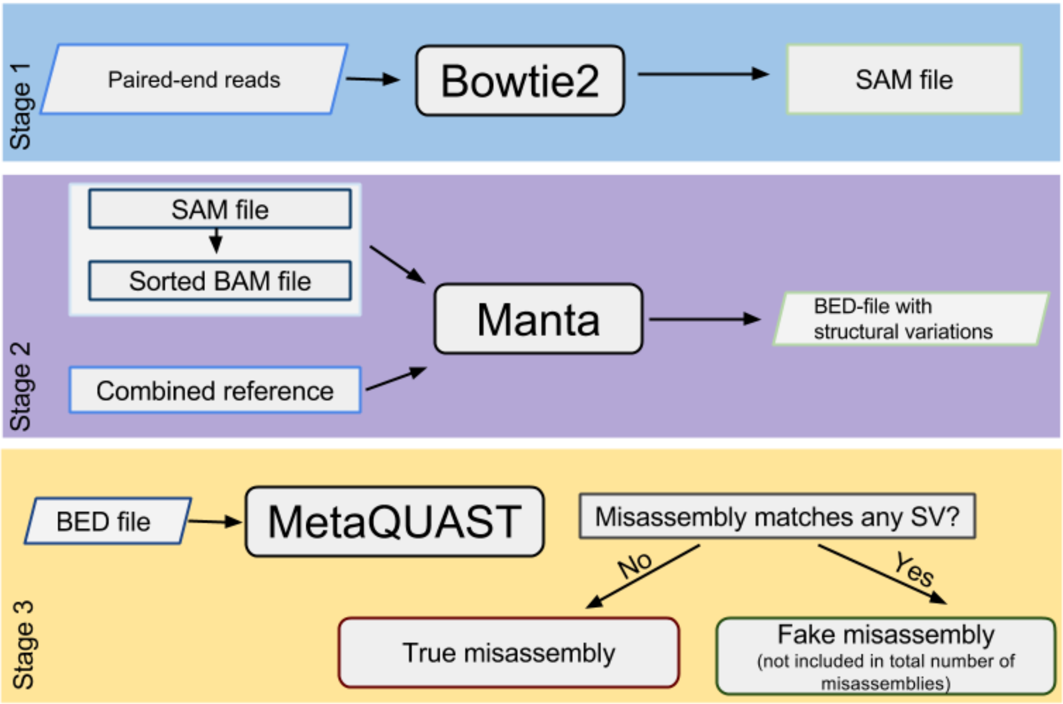
The general workflow of the MetaQUAST module for misassemblies refinement based on read mapping is presented in Supplementary Fig. S3. The processing starts with aligning reads on a reference genome, continues with structural variant (SV) calling, and finishes with comparing each misassembly with found SVs. The latter step is shown in details in Supplementary Fig. S4.

This approach allows us to significantly reduce the number of falsely reported misassemblies on all three test datasets. Supplementary Table S1 shows results of misassemblies refinement on the combined reference of the CAMI dataset. Online reports at http://bioinf.spbau.ru/metaquast include statistics for all references of three test datasets. For each assembly, *# structural variations* metric shows number of misassemblies matched with SVs and marked false while *# misassemblies* reports total number of real misassemblies. The performance of the refinement step is demostrated in section 2.

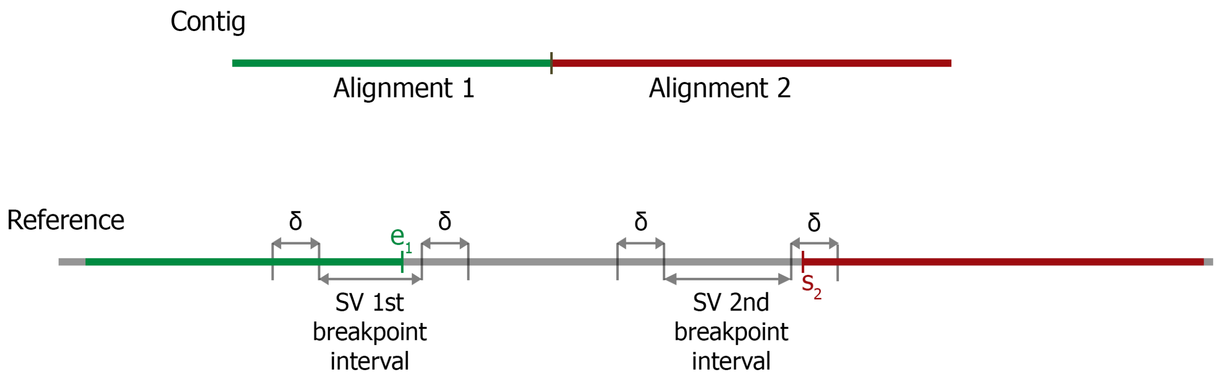
Supplementary Table S1: Misassemblies refinement results

Assembly	No. of misassemblies before refinement	 after refinement	Refinement rate (%)
Gold Assembly	453	342	24.5%
IDBA-UD	1469	1405	4.4%
Ray	247	187	24.3%
SOAPdenovo2	79	58	26.6%
SPAdes	1077	1017	5.6%

Misassemblies refinement results on the combined reference of the CAMI dataset. *No. of misassemblies before refinement* is the total number of misassemblies initially reported by MetaQUAST. *No. of misassemblies after refinement* is the number of misassemblies after excluding false positive ones (matched with SV). *Refinement rate* is calculated as the number of found false positive misassemblies divided by the total number of misassemblies before refinement.



Supplementary Fig. S3: The general workflow of the MetaQUAST misassemblies refinement module. Stage 1: bowtie2 (Langmead *et al.*, 2009) aligns reads against the combined reference genome. Stage 2: Manta (Chen *et al.*, 2015) looks through the resulting BAM file (Li *et al.*, 2009) and reports SVs based on discordant read-pairs. Stage 3: MetaQUAST compares each misassembly coordinates with the SVs list and classifies them into true ones or falsely reported ones.



Supplementary Fig. S4: Each misassembly reported by QUAST is compared with breakpoint confidence intervals of all discovered SVs. If both start and end coordinates of the misassembly lie within the SV calling intervals extended by a small δ (default value is 100 bp), MetaQUAST marks this misassembly fake and does not include into the final report. In the figure, the end of the first alignemnt (e_1) is the start of the misassembly (*relocation*), and it lies within the first confidence interval of an SV call. The starting point of the second alignement (s_2) is the end of the misassembly, and it lies within the second confidence interval of the same SV extended by δ. This misassembly is considered as caused by structural differences between the reference and the assembly and not reported.

2 MetaQUAST performance

We benchmarked MetaQUAST on three datasets: *CAMI* (http://cami-challenge.org) toy test dataset simulated from 30 publicly available genomes, MH0045 sample from the *MetaHIT* project (Qin *et al.*, 2010), and tongue dorsum female sample, SRS077736, from the *NIH Human Microbiome Project* (Consortium *et al.*, 2012). See Supplementary Table S2 for performance results. All benchmarking was done on a 4 CPU (Intel Xeon X7560 2.27GHz) computer. When running MetaQUAST without provided reference genomes (CAMI and HMP datasets), BLAST alignment and references downloading from NCBI were the most time-consuming steps (see column 5 in the table). This time may significantly vary depending on Internet connection bandwidth. Note that almost all processes (excluding reference downloading) are parallelized, with each assembly processed in a separate thread, so MetaQUAST works faster on computers with more CPUs.

We benchmarked the misassembly refinement step (see section 1.2) separately because it highly depends on the number of input reads and the combined reference features, e.g. number and length of repeated fragments. All benchmarking was done on the same computer as above. See Supplementary Table S3 for performance results. Note that bowtie2 used for reads alignment is effectively parallelized, thus the misassembly refinement step works significantly faster on computers with more CPUs.

Supplementary Table S2: MetaQUAST performance (excluding the misassemblies refinement step)

Dataset	No. of assemblies	Average assembly size	Total time	Time for searching references
CAMI	5	90 Mb	2h 09m	0h 51m
MetaHIT	4	80 Mb	1h 29m	–
HMP	4	120 Mb	2h 27m	1h 01m

Supplementary Table S3: Misassemblies refinement step performance

Dataset	No. of reads	Combined reference size	Reads aligment time	SV calling time	Total time
CAMI	147.8M	93 Mb	2h 19m	0h 35m	2h 54m
MetaHIT	20.8M	298 Mb	0h 58m	0h 31m	1h 29m
HMP	91.5M	59 Mb	3h 22m	0h 46m	4h 08m

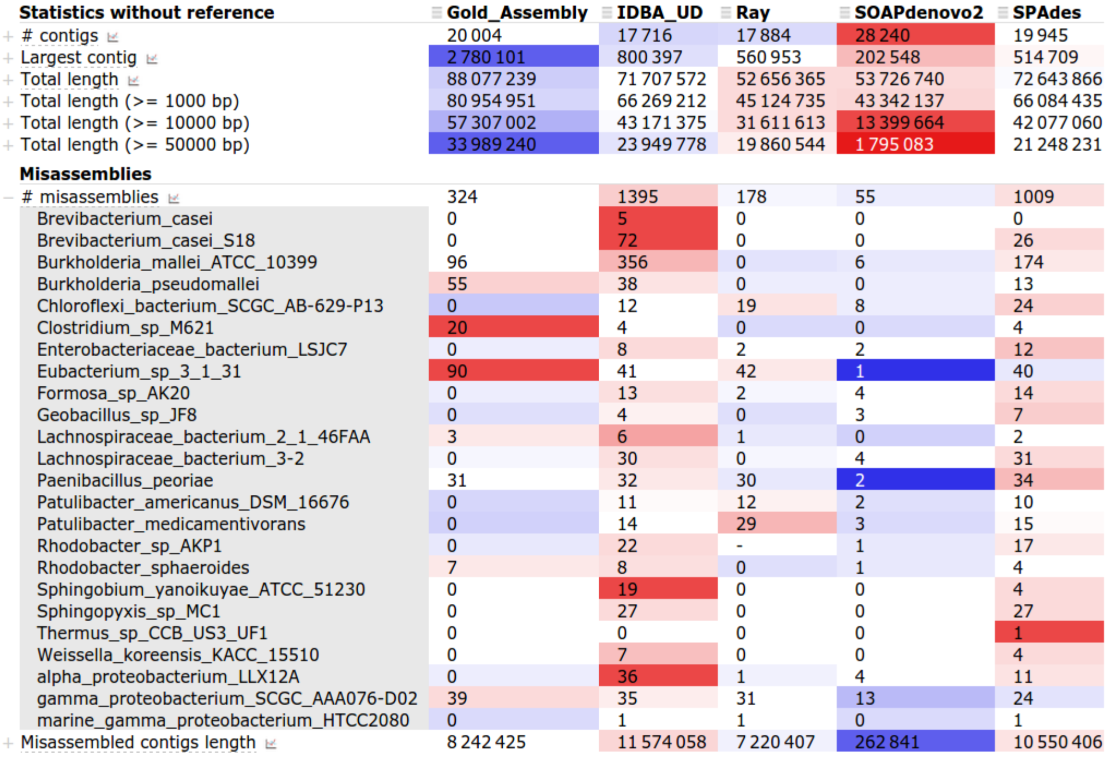
3 MetaQUAST report on CAMI toy dataset

We assembled the CAMI dataset using IDBA-UD (Peng *et al.*, 2012), SPAdes (Bankevich *et al.*, 2012), Ray Meta (Boisvert *et al.*, 2012), and SOAPdenovo2 (Luo *et al.*, 2012). The assemblies were complemented with the "Gold Assembly" provided by CAMI on their website. MetaQUAST was run in *de novo* evaluation mode and references detected by our algorithm were compared with the original ones used for the dataset simulation.

MetaQUAST detected 82 genomes, 50 of which were downloaded from the NCBI database, and 24 passed the filtering step (have a genome coverage fraction more than 10%). Average genome fraction for these 24 genomes based on "Gold Assembly" alignments is 65%, and 14 of these genomes are covered by more than 90%. 16 of these organisms precisely matched genomes used for simulating dataset, and 8 organisms represent another species of the same genus. Only 6 organisms from 30 were not found completely. Their alignments to corresponding rRNA sequences were detected but had a very low BLAST score and did not pass our threshold of the 50 best hits per assembly.

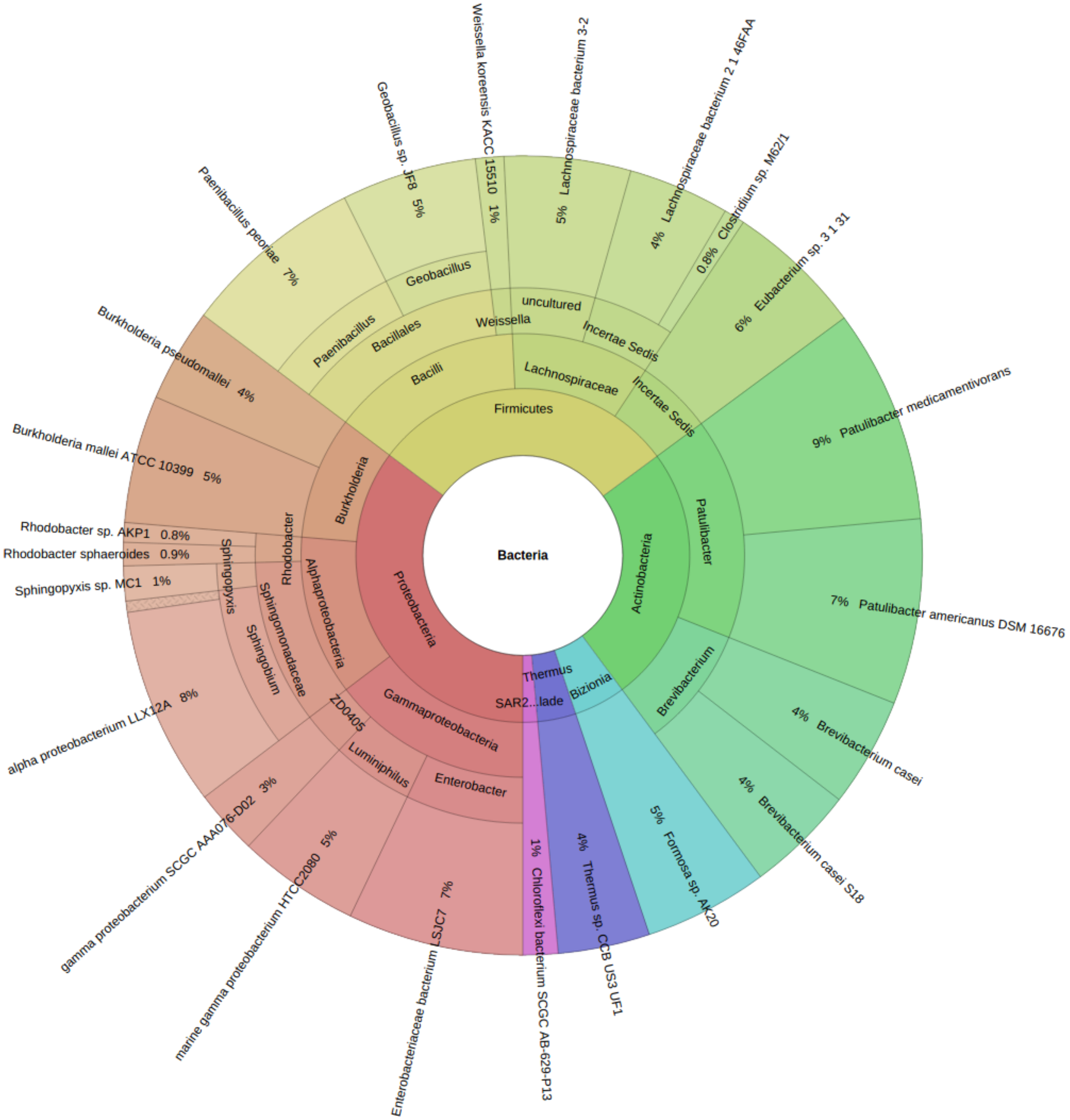
All five assemblies were evaluated against downloaded 24 references. Supplementary Fig. S5 shows that "Gold Assembly", provided by the CAMI team, has the best results in majority of metrics. A high number of misassemblies for some references possibly indicate the presence of other organisms, closely related to the downloaded genomes.

None of the assemblers may be called the best (or the worst) with regards to the majority of metrics. IDBA-UD assembled the largest contig (800 397 bp). SPAdes has a slightly larger total length than IDBA-UD (72 643 866 bp versus 71 707 572 bp), and a significantly fewer number of misassemblies (1009 versus 1395). SOAPdenovo2 has a very low number of misassemblies (only 55) but its genome fraction is twice smaller than IDBA-UD and SPAdes, and it has a very low length of contigs larger than 50 kbp (1 795 083 bp).

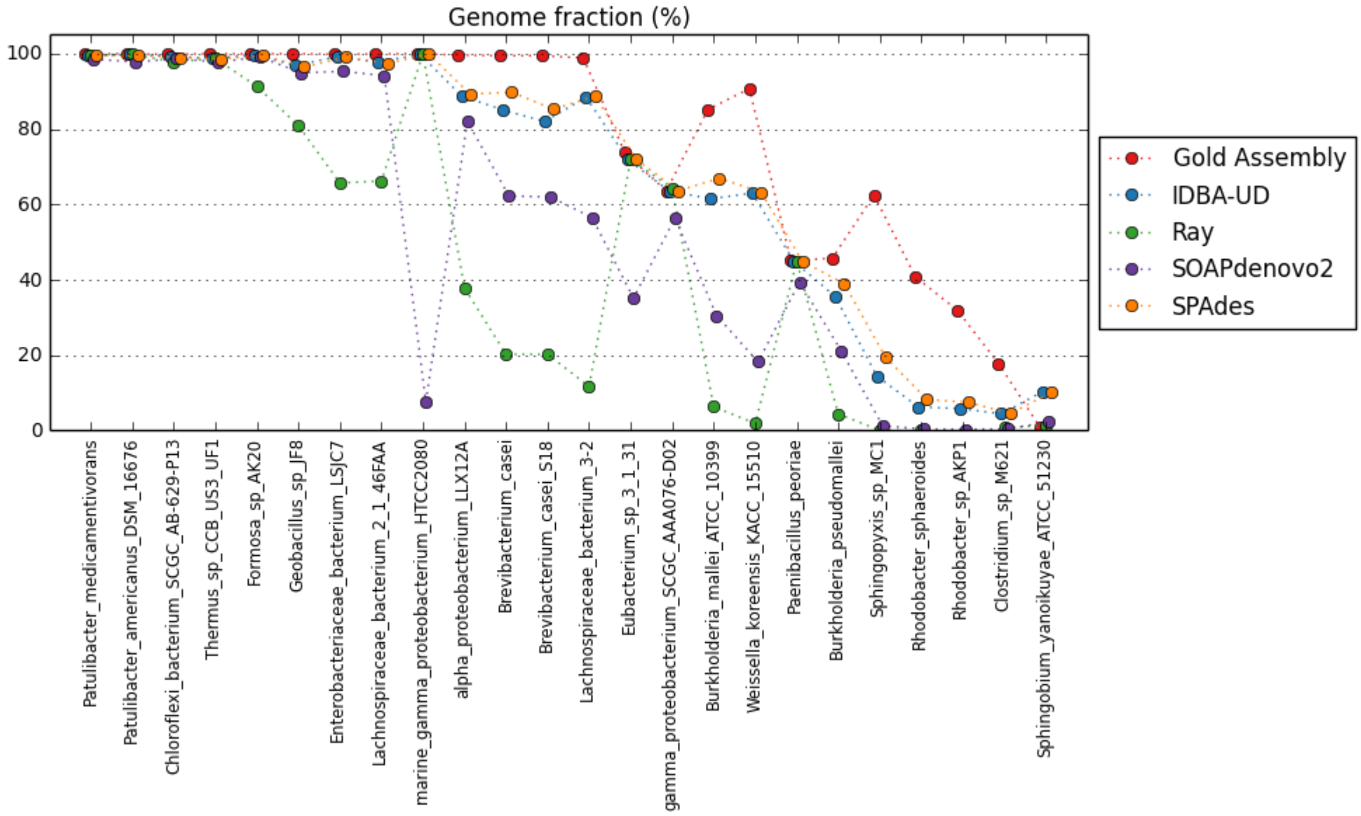


Statistics without reference	Gold_Assembly	IDBA_UD	Ray	SOAPdenovo2	SPAdes
# contigs	20 004	17 716	17 884	28 240	19 945
Largest contig	2 780 101	800 397	560 953	202 548	514 709
Total length	88 077 239	71 707 572	52 656 365	53 726 740	72 643 866
Total length (>= 1000 bp)	80 954 951	66 269 212	45 124 735	43 342 137	66 084 435
Total length (>= 10000 bp)	57 307 002	43 171 375	31 611 613	13 399 664	42 077 060
Total length (>= 50000 bp)	33 989 240	23 949 778	19 860 544	1 795 083	21 248 231
Misassemblies					
# misassemblies	324	1395	178	55	1009
Brevibacterium_casei	0	5	0	0	0
Brevibacterium_casei_S18	0	72	0	0	26
Burkholderia_mallei_ATCC_10399	96	356	0	6	174
Burkholderia_pseudomallei	55	38	0	0	13
Chloroflexi_bacterium_SCGC_AB-629-P13	0	12	19	8	24
Clostridium_sp_M621	20	4	0	0	4
Enterobacteriaceae_bacterium_LSJC7	0	8	2	2	12
Eubacterium_sp_3_1_31	90	41	42	1	40
Formosa_sp_AK20	0	13	2	4	14
Geobacillus_sp_JF8	0	4	0	3	7
Lachnospiraceae_bacterium_2_1_46FAA	3	6	1	0	2
Lachnospiraceae_bacterium_3-2	0	30	0	4	31
Paenibacillus_peoriae	31	32	30	2	34
Patulibacter_americanus_DSM_16676	0	11	12	2	10
Patulibacter_medicamentivorans	0	14	29	3	15
Rhodobacter_sp_AKP1	0	22	-	1	17
Rhodobacter_sphaeroides	7	8	0	1	4
Sphingobium_yanoikuyae_ATCC_51230	0	19	0	0	4
Sphingopyxis_sp_MC1	0	27	0	0	27
Thermus_sp_CCB_US3_UF1	0	0	0	0	1
Weissella_koreensis_KACC_15510	0	7	0	0	4
alpha_proteobacterium_LLX12A	0	36	1	4	11
gamma_proteobacterium_SCGC_AAA076-D02	39	35	31	13	24
marine_gamma_proteobacterium_HTCC2080	0	1	1	0	1
Misassembled contigs length	8 242 425	11 574 058	7 220 407	262 841	10 550 406

Supplementary Fig. S5: A part of the summary HTML report for the CAMI dataset. The full version is available online at http://bioinf.spbau.ru/metaquast. The heatmap helps to visually pinpoint outliers. Cells containing median values are colored in white. The cells containing outliers are brightly colored (with blue corresponding to the best values, and red corresponding to the worst). SOAPdenovo2 shows the best results in misassemblies but had a low genome fraction and a low Total length value.



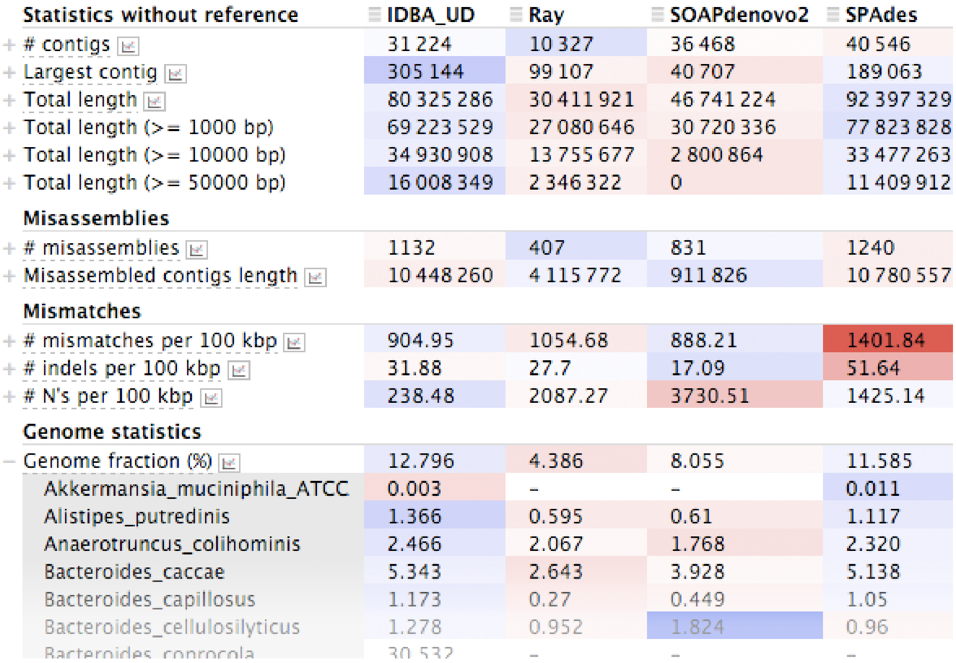
Supplementary Fig. S6: Krona round chart for the CAMI dataset. The chart demonstrates average abundance of every species in the dataset (based on all 5 assemblies). Relative species abundance is calculated based on the total length of contigs aligned to a corresponding reference genome.



Supplementary Fig. S7: Metric-level plot for genome fraction on the CAMI dataset. This plot shows genome coverage fraction for all assemblies versus all references. References on the plot are sorted by the mean value of this metric in all assemblies, starting from the best result to the worst one. The “Gold Assembly” has the highest values on almost all references, followed closely by SPAdes and IDBA-UD

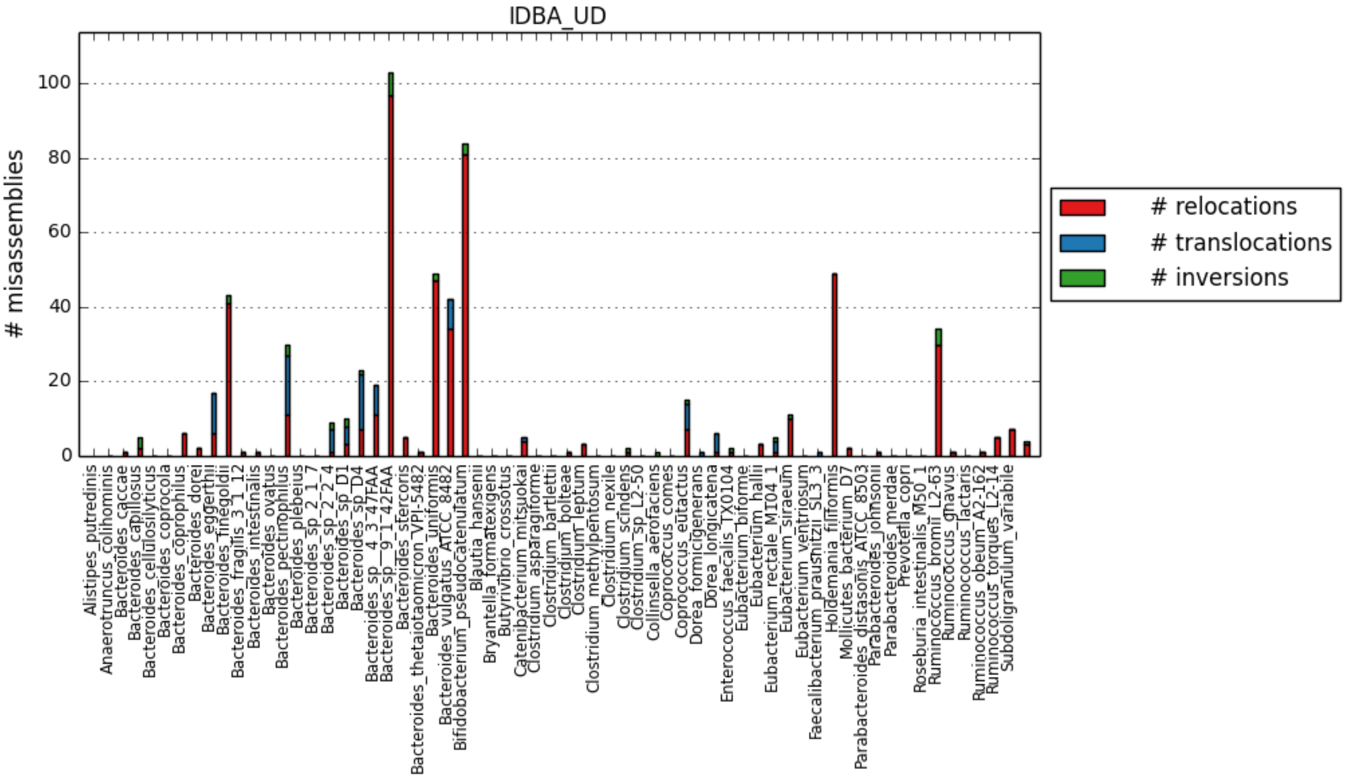
4 MetaQUAST report on MetaHIT dataset

MetaQUAST was ran on MH0045 sample from the MetaHIT project. We downloaded genomes of 75 species with >1% genome coverage by reads in >50% of the cohort individuals, participated in the project, as described in Qin *et al.* (2010). Most of the reference genomes (52 of 75) have a genome fraction less than 10% (see Supplementary Fig. S8). Also, more than half of assembly length were not aligned to any of the reference genomes, and, probably, contained a large number of unknown organisms.



Statistics without reference	IDBA_UD	Ray	SOAPdenovo2	SPAdes
# contigs	31 224	10 327	36 468	40 546
Largest contig	305 144	99 107	40 707	189 063
Total length	80 325 286	30 411 921	46 741 224	92 397 329
Total length (>= 1000 bp)	69 223 529	27 080 646	30 720 336	77 823 828
Total length (>= 10000 bp)	34 930 908	13 755 677	2 800 864	33 477 263
Total length (>= 50000 bp)	16 008 349	2 346 322	0	11 409 912
Misassemblies				
# misassemblies	1132	407	831	1240
Misassembled contigs length	10 448 260	4 115 772	911 826	10 780 557
Mismatches				
# mismatches per 100 kbp	904.95	1054.68	888.21	1401.84
# indels per 100 kbp	31.88	27.7	17.09	51.64
# N's per 100 kbp	238.48	2087.27	3730.51	1425.14
Genome statistics				
Genome fraction (%)	12.796	4.386	8.055	11.585
Akkermansia_muciniphila_ATCC	0.003	-	-	0.011
Alistipes_putredinis	1.366	0.595	0.61	1.117
Anaerotruncus_colihominis	2.466	2.067	1.768	2.320
Bacteroides_caccae	5.343	2.643	3.928	5.138
Bacteroides_capillosus	1.173	0.27	0.449	1.05
Bacteroides_cellulosilyticus	1.278	0.952	1.824	0.96
Bacteroides_coprocola	30.532	-	-	-

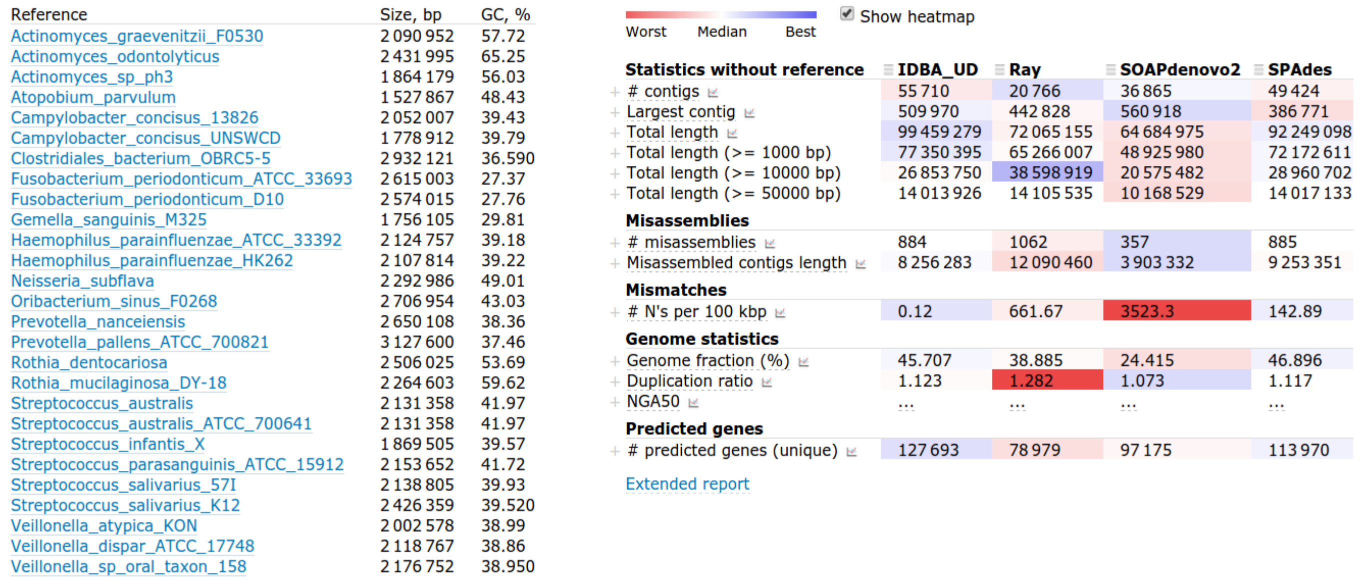
Supplementary Fig. S8: A part of the summary HTML report for the MetaHIT dataset. The full version is available online at http://bioinf.spbau.ru/metaquast. IDBA-UD and SPAdes assembled more genomes than Ray and SOAPdenovo2. At the same time, IDBA-UD and SPAdes demonstrated their best results on different organisms (dark blue cells in the expanded row).



Supplementary Fig. S9: Misassemblies plot of IDBA-UD assembly of the MetaHIT dataset. The plot shows the distribution of misassemblies by type (relocations, inversions, translocations) across all references for one assembly (IDBA-UD).

5 MetaQUAST report on HMP dataset

MetaQUAST was ran on the SRS077736 sample (tongue dorsum) from the HMP project without providing any references. 26 reference genomes with a genome coverage fraction >10% were found. However, all assemblies contain large fragments not aligned to the combined reference (43-66% of assemblies bases are unaligned). A short version of the summary HTML report is shown in Supplementary Fig. S10. IDBA-UD has the largest total length. IDBA-UD and SPAdes have a significantly higher genome fraction (45.7% and 46,9%) than Ray and, especially, SOAPdenovo2 (38,9% and 24,4% respectively). SOAPdenovo2 provides the most accurate assembly with a minimal number of misassemblies and mismatches, and has the largest contig. Ray demonstrates the lowest number of contigs, but the highest number of misassemblies.



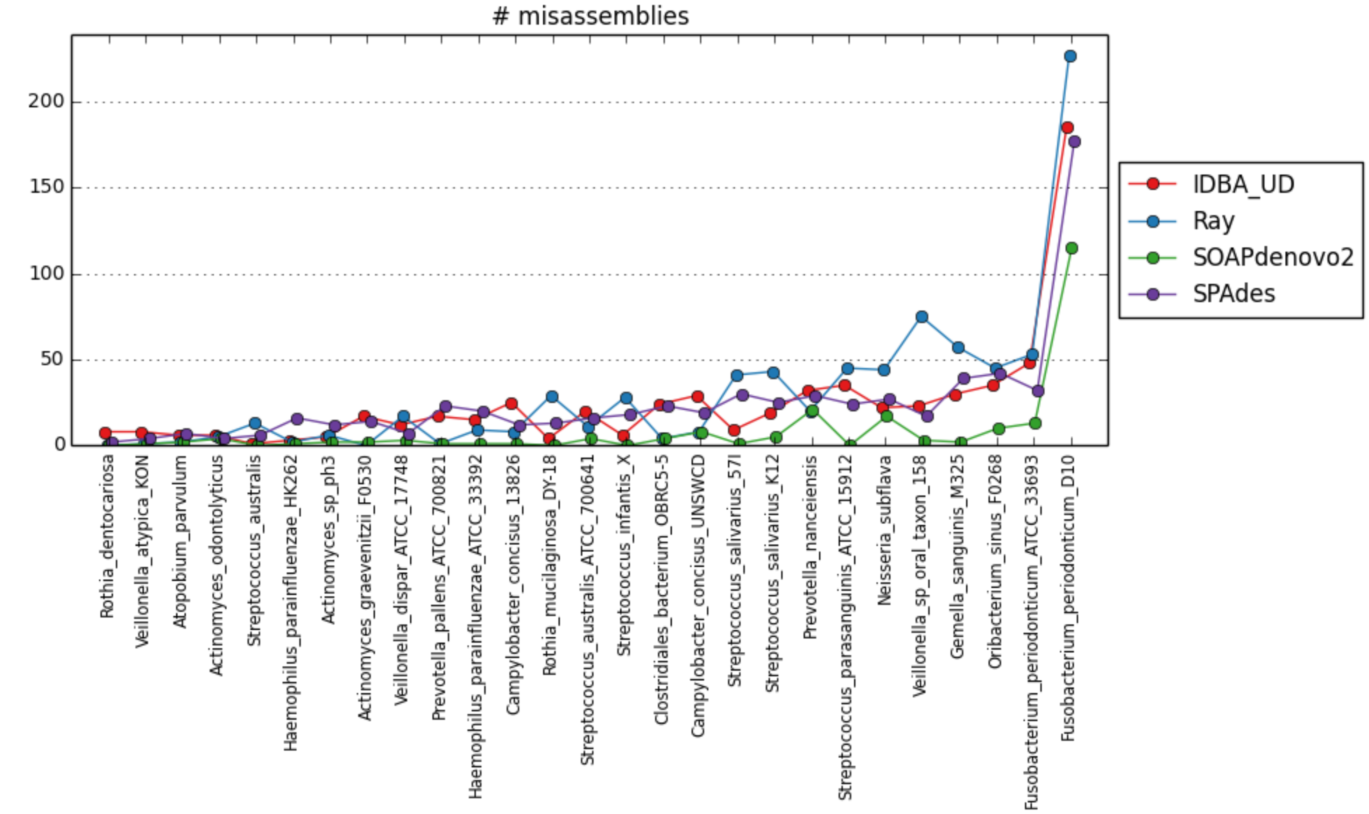
Reference	Size, bp	GC, %
Actinomyces_graevenitzii_F0530	2 090 952	57.72
Actinomyces_odontolyticus	2 431 995	65.25
Actinomyces_sp_ph3	1 864 179	56.03
Atopobium_parvulum	1 527 867	48.43
Campylobacter_concisus_13826	2 052 007	39.43
Campylobacter_concisus_UNSWCD	1 778 912	39.79
Clostridiales_bacterium_OBRC5-5	2 932 121	36.590
Fusobacterium_periodonticum_ATCC_33693	2 615 003	27.37
Fusobacterium_periodonticum_D10	2 574 015	27.76
Gemella_sanguinis_M325	1 756 105	29.81
Haemophilus_parainfluenzae_ATCC_33392	2 124 757	39.18
Haemophilus_parainfluenzae_HK262	2 107 814	39.22
Neisseria_subflava	2 292 986	49.01
Oribacterium_sinus_F0268	2 706 954	43.03
Prevotella_nanceiensis	2 650 108	38.36
Prevotella_pallens_ATCC_700821	3 127 600	37.46
Rothia_dentocariosa	2 506 025	53.69
Rothia_mucilaginosa_DY-18	2 264 603	59.62
Streptococcus_australis	2 131 358	41.97
Streptococcus_australis_ATCC_700641	2 131 358	41.97
Streptococcus_infantis_X	1 869 505	39.57
Streptococcus_parasanguinis_ATCC_15912	2 153 652	41.72
Streptococcus_salivarius_57I	2 138 805	39.93
Streptococcus_salivarius_K12	2 426 359	39.520
Veillonella_atypica_KON	2 002 578	38.99
Veillonella_dispar_ATCC_17748	2 118 767	38.86
Veillonella_sp_oral_taxon_158	2 176 752	38.950

Worst Median Best ☑ Show heatmap

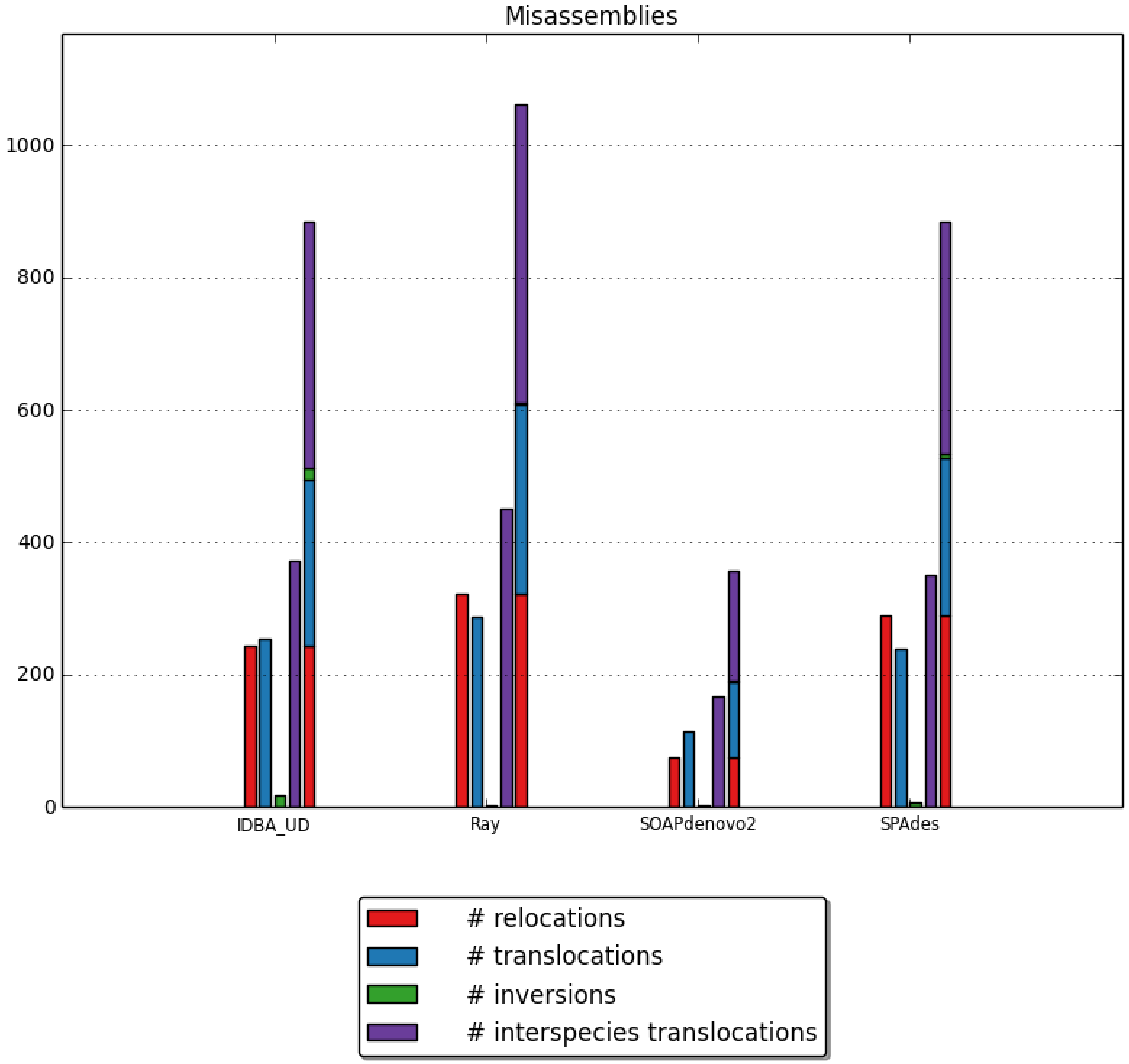
Statistics without reference	IDBA_UD	Ray	SOAPdenovo2	SPAdes
# contigs	55 710	20 766	36 865	49 424
Largest contig	509 970	442 828	560 918	386 771
Total length	99 459 279	72 065 155	64 684 975	92 249 098
Total length (>= 1000 bp)	77 350 395	65 266 007	48 925 980	72 172 611
Total length (>= 10000 bp)	26 853 750	38 598 919	20 575 482	28 960 702
Total length (>= 50000 bp)	14 013 926	14 105 535	10 168 529	14 017 133
Misassemblies				
# misassemblies	884	1062	357	885
Misassembled contigs length	8 256 283	12 090 460	3 903 332	9 253 351
Mismatches				
# N's per 100 kbp	0.12	661.67	3523.3	142.89
Genome statistics				
Genome fraction (%)	45.707	38.885	24.415	46.896
Duplication ratio	1.123	1.282	1.073	1.117
NGA50	...	...	...	...
Predicted genes				
# predicted genes (unique)	127 693	78 979	97 175	113 970

Extended report

Supplementary Fig. S10: Summary HTML report for the HMP dataset. Ray loses on Duplication ratio and the number of misassemblies. SOAPdenovo2 assembly contains a lot of undefined nucleotides (N). IDBA-UD and SPAdes assemblies are longer than Ray and SOAPdenovo2, though have similar numbers of contigs.



Supplementary Fig. S11: Metric-level plot for number of misassemblies on the HMP dataset. This plot shows the number of misassemblies for all assemblies versus all references. SOAPdenovo2 has the lowest number of misassemblies in all references. An unexpectedly high number of misassemblies in Fusobacterium periodonticum D10 indicate probable presence of other species closely related to Fusobacterium in the dataset.



Supplementary Fig. S12: Misassemblies plot for the HMP dataset. The plot shows a distribution of misassemblies by type (relocations, inversions, translocations, interspecies translocations) across all assemblies versus the combined reference. One additional column corresponds to the total number of misassemblies. A high number of interspecies translocations in assemblies is caused by presence of very closely related species in the dataset, and emphasizes difficulty of metagenome assembly.

6 List of used assemblers

We used four assemblers in the comparisons. All of them were launched with default parameters.

- IDBA-UD v.1.1.1 (Peng *et al.*, 2012)
- Ray v.2.3.1 (Boisvert *et al.*, 2012)
- SOAPdenovo2 v.2.04 (Li *et al.*, 2010)
- SPAdes v.3.5.0 (Bankevich *et al.*, 2012)

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.

Boisvert, S. *et al.* (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol*, **13**(12), R122.

Camacho, C. *et al.* (2009). Blast+: architecture and applications. *BMC bioinformatics*, **10**(1), 421.

Chen, X. *et al.* (2015). Manta: Rapid detection of structural variants and indels for clinical sequencing applications. *bioRxiv*.

Consortium, H. M. P. *et al.* (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, **486**(7402), 207–214.

Gurevich, A. *et al.* (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**(8), 1072–1075.

Langmead, B. *et al.* (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**(3), R25.

Li, H. *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

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

Luo, R. *et al.* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**(1), 18.

Ondov, B. *et al.* (2011). Interactive metagenomic visualization in a Web browser. *BMC bioinformatics*, **12**(1), 385.

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.

Qin, J. *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**(7285), 59–65.

Quast, C. *et al.* (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*, **41**(D1), D590–D596.