Supplementary Materials

Appendix: Datasets

B. cereus ATCC-10987

Reference genome: Accession number NC_003909.8

Reads: <data from Illumina shared privately>

R. sphaeroides 2.4.1

Reference genome: Accession number AKVW01000000

Reads: Accession number SRR522246

P. stuartii ATCC 33672

Reference genome: Accession numbers CP008919.1 CP008920.1

Reads: Accession number SRR1558174

C. freundii CFN1H1

Reference genome: Accession numbers CP007557.1 CP007558.1

Reads: Accession number SRR1284629

B. cenocepacia DDS 22E-1

Reference genome: Accession numbers CP007782.1 CP007783.1 CP007784.1

Reads: Accession number SRR1618480

C.callunae DSM 20147

Reference genome: Accession number AKVW01000000

Reads: Accession number SRR522246

Acinetobacter sp. UNC434CL69Tsu2S25

Reads: http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=AcispL69Tsu2S25

Butyrivibrio sp. INlla16

Reads: http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=ButspINlla16

Lachnospiraceae bacterium NK3A20

Reads: http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=LacbacNK3A20

Luteibacter sp. UNC138MFCol5.1

Reads: http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=LutspUNC138MFC1

Prevotellaceae bacterium HUN156,

Reads: http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PrebacHUN156

Pseudoalteromonas sp. ND6B

Reads: Accession number SRR1552349

Rhodococcus sp. J21

Reads: Accession number SRR1799421

Ruminococcus flavefaciens YAD2003

Reads: http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=RumflaYAD2003

Sphingomonas sp. UNC305MFCol5.2

Reads: Accession number SRR1798208

Thermus filiformis ATT43280.

Reads: Accession number SRR1798208

Appendix: Benchmarking plasmidSPAdes on genomes with annotated plasmids

For benchmarking purposes, we introduce the notion of a *simple plasmid*, which is a circular plasmid less than a megabase in length with coverage that significantly differs from the chromosome coverage (wrt parameter *maxDeviation*). Below we focus on the problem of assembling simple plasmids and identify each putative plasmid (component in the plasmid graph) by number, where *component* 0 indicates the largest putative plasmid in the output, *component* 1 is the second largest, etc. To analyze each putative plasmid, we use QUAST to align its plasmidic contigs to the reference genome. We further use this alignment to figure how many contigs in each putative plasmid match to the reference genome and compute its plasmid and chromosome fractions. Ideally, a putative plasmid would have plasmid fraction 100% and chromosome fraction 100%.

<u>*B. cereus*</u>

plasmidSPAdes identified a single (simple) plasmid in *Bce*. The assembly consists of long contig containing 99.8% of plasmid and two short contigs of length 165 and 254 that presumably represent a variation in the plasmid.

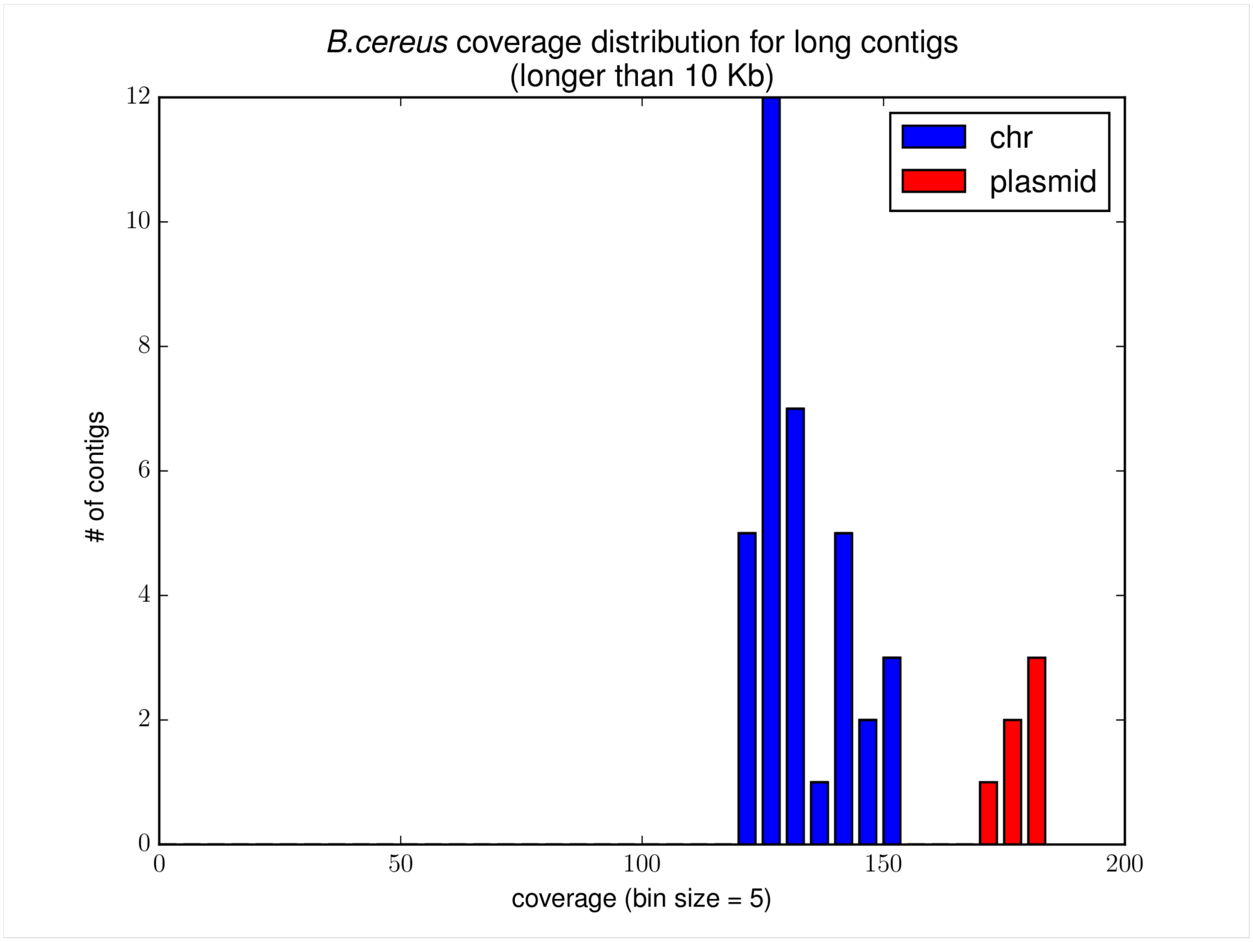

Figure 1. The distribution of *k*-mer coverage for all long contigs in *B. cereus* (*medianCoverage* = 130).

R. sphaeroides

plasmidSPAdes recovered large portions of all five annotated plasmids but combined them into a single component of the plasmid graph. The plasmid fraction is 100% for 4 of 5 plasmids. However, since these plasmids share repeats, plasmidSPAdes failed to separate them from each other in the plasmid graph. For the fifth plasmid (*Dx* plasmid), plasmidSPAdes deleted a plasmidic contig of length 4kbp which had no connections to the rest of the assembly graph. No chromosomal contigs were detected in the output of plasmidSPAdes.

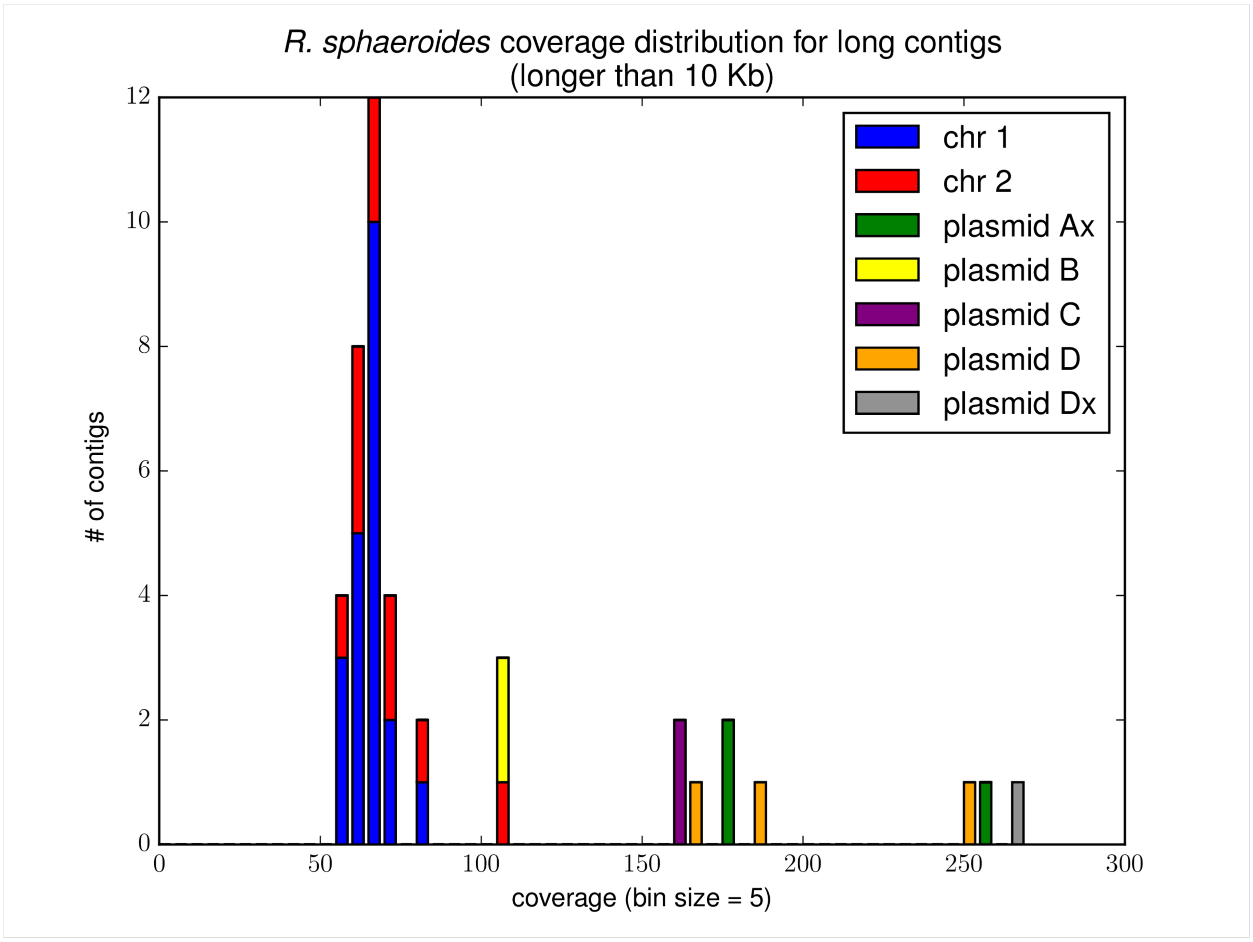

Figure 2. The distribution of *k*-mer coverage for all long contigs in *R.sphaeroides* (*medianCoverage* = 67).

P. stuartii ATCC 33672

plasmidSPAdes identified a single annotated plasmids in *Pst* dataset and assembled it into a circular plasmid (a loop-edge forming component 0). The only other putative plasmid (component 1) is a false positive artifact formed by a genomic repeat and short chromosomal edges.

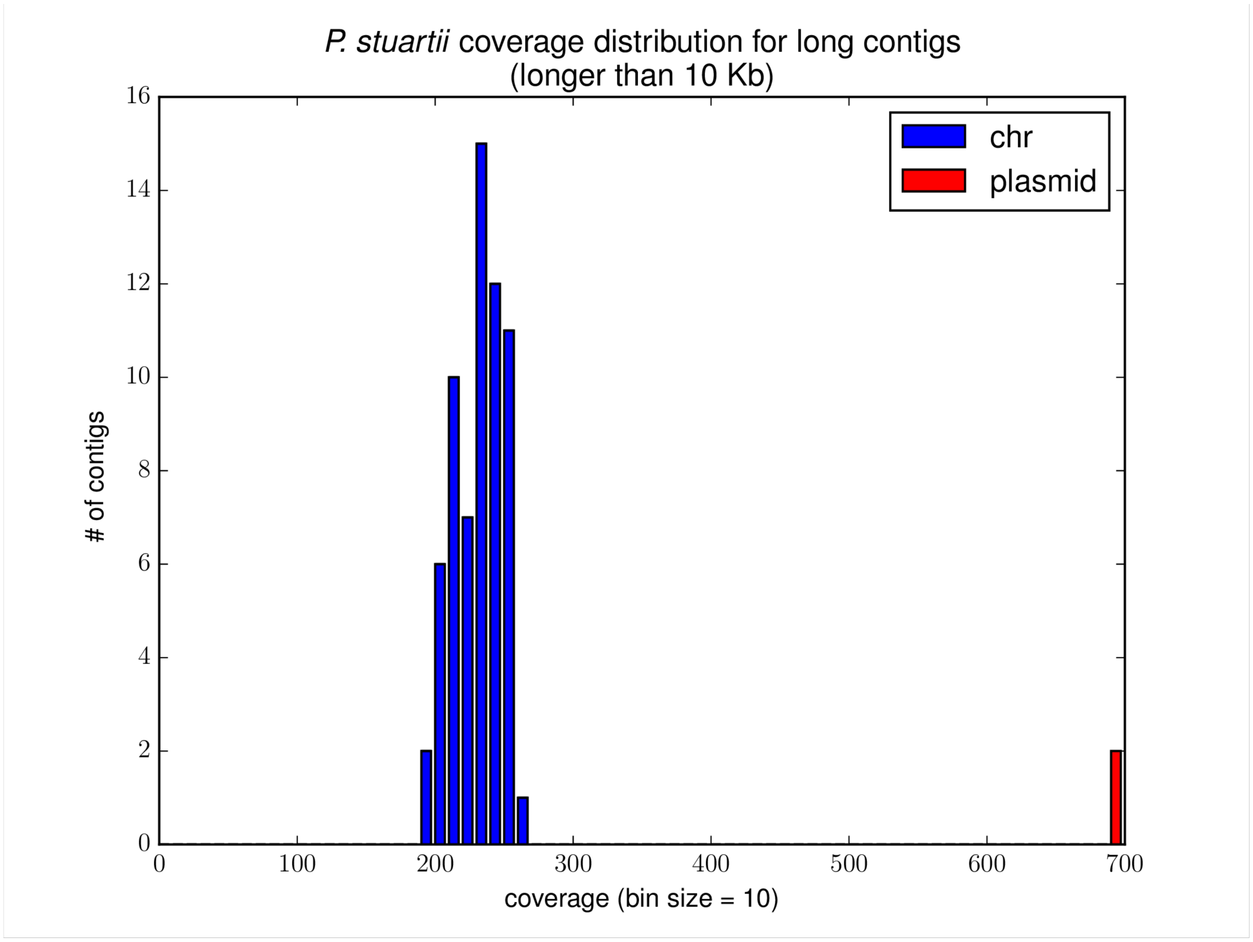

Figure 3. The distribution of *k*-mer coverage for all long contigs in *P. stuartii* (*medianCoverage* = 231).

C. freundii CFNIH1

plasmidSPAdes partially recovered a single annotated plasmids in *Cfr* dataset (plasmid coverage 99.6%) and assembled it into a complex component 0 (this plasmid has a complex repeat structure). The missed 0.4% coverage reported by QUAST is not actually missing but rather an artifact of the standard QUAST report (alignment of a contig to a single rather than multiple copies of a repeat). The component 0 also includes some short chromosomal contigs that were incorrectly retained as they are connected to the plasmid contigs in the de Bruijn graph.

Component 1 is a previously unidentified short plasmid in *C. freundii* with high copy number. It has a high-scoring BLAST hit to the plasmid pCAV1335-5410 in *Klebsiella oxytoca strain CAV1335* (alignment length 4454 and percent identity 99.9).

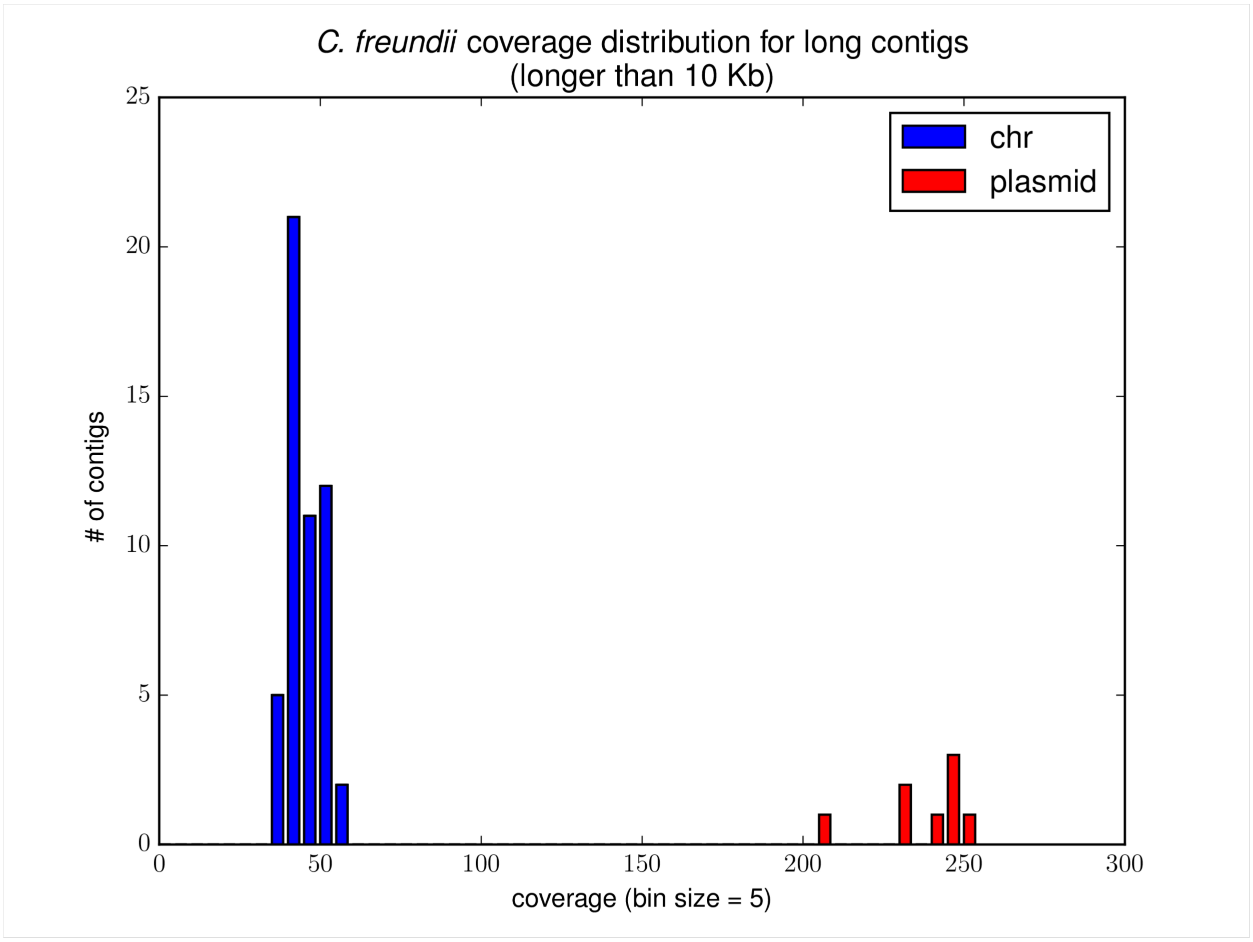

Figure 4. The distribution of *k*-mer coverage for all long contigs in *C. freundii* (*medianCoverage* = 47).

B. cenocepacia DDS 22E-1

This dataset without annotated contigs has been included in benchmarking as a negative control. plasmidSPAdes did not identify any putative plasmids in this dataset (empty plasmid graph).

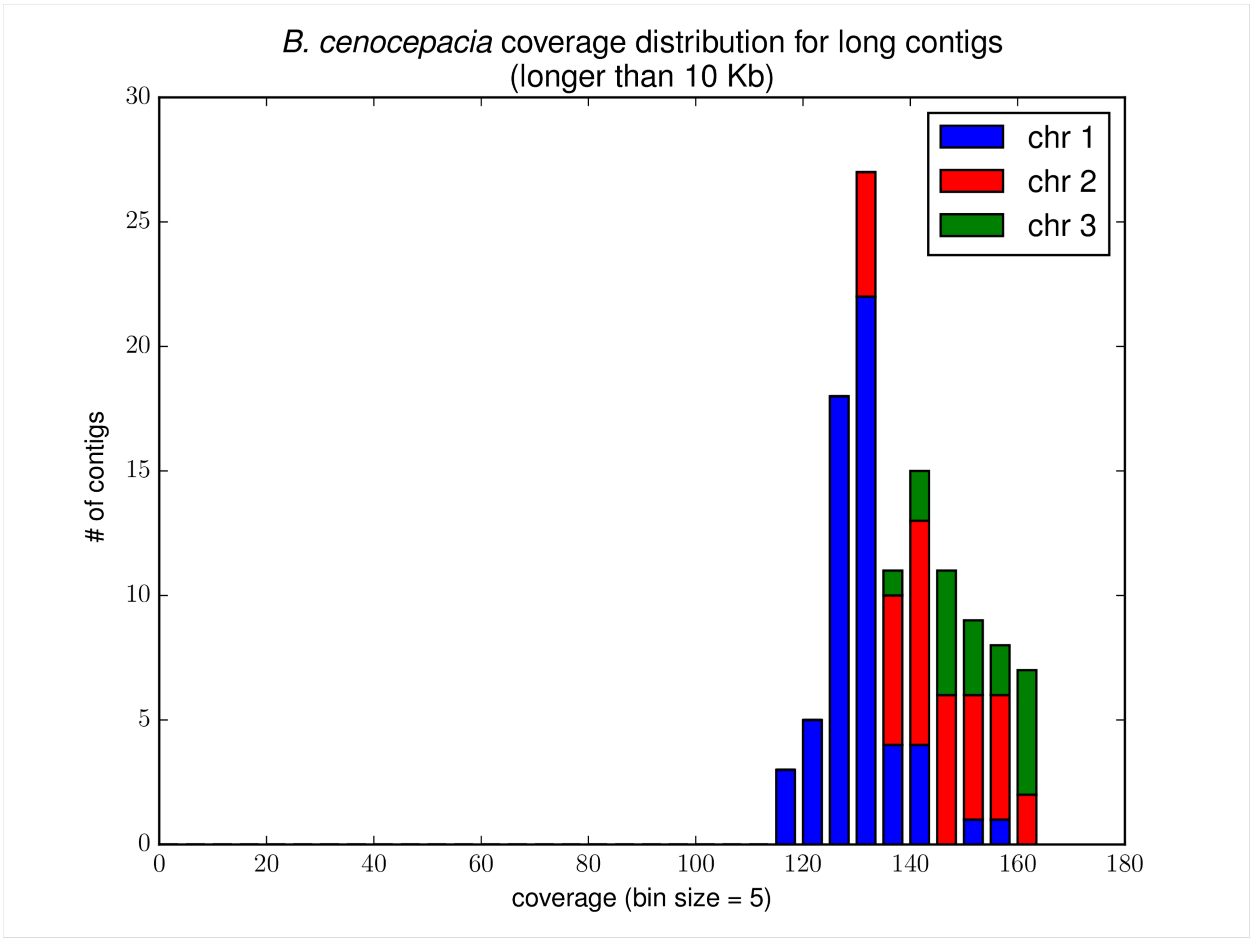

Figure 5. The distribution of *k*-mer coverage for all long contigs in *B. cenocepacia*. (*medianCoverage* = 136).

C. callunae DSM 20147

Out of 2 annotated plasmids, plasmidSPAdes assembled the short high copy number plasmid as a circular contig (component 1). The larger plasmid was not recovered since its coverage (180) is similar to the chromosome coverage (206). Component 0 contains contigs of chromosomal origin originating from a genomic repeat.

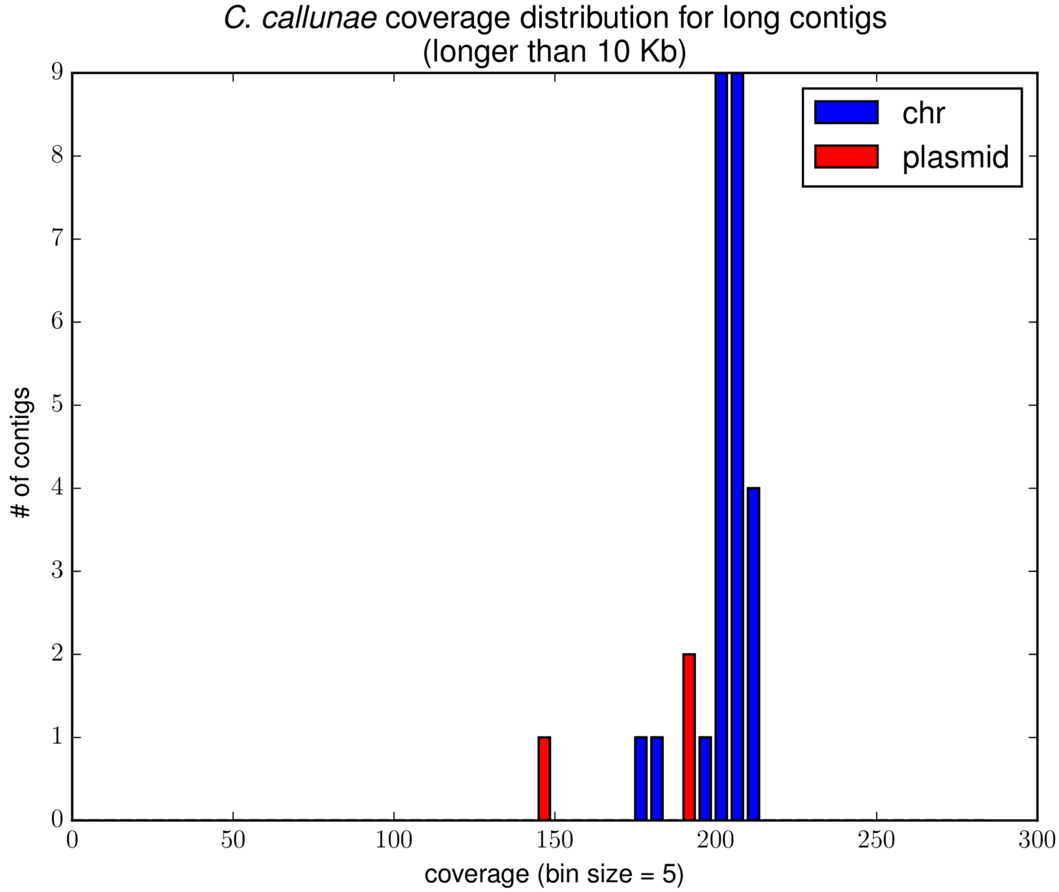

Figure 6. The distribution of *k*-mer coverage for all long contigs in *C. callunae* (*medianCoverage* = 206). The 2nd plasmid is not shown in this histogram since it does not contain long contigs.

Appendix: Benchmarking plasmidSPAdes on genomes with unannotated plasmids

Acinetobacter sp. UNC434CL69Tsu2S25

plasmidSPAdes identified four putative plasmid components in *Aci* dataset, including two circular putative plasmids (one of them has been confirmed). Component 0's best BLAST hit is to a plasmid from another *Acinetobacter* species. Component 1's best BLAST hits is to another *Acinetobacter* species chromosome. Component 2 is a circular contig that originated from a mobile element.

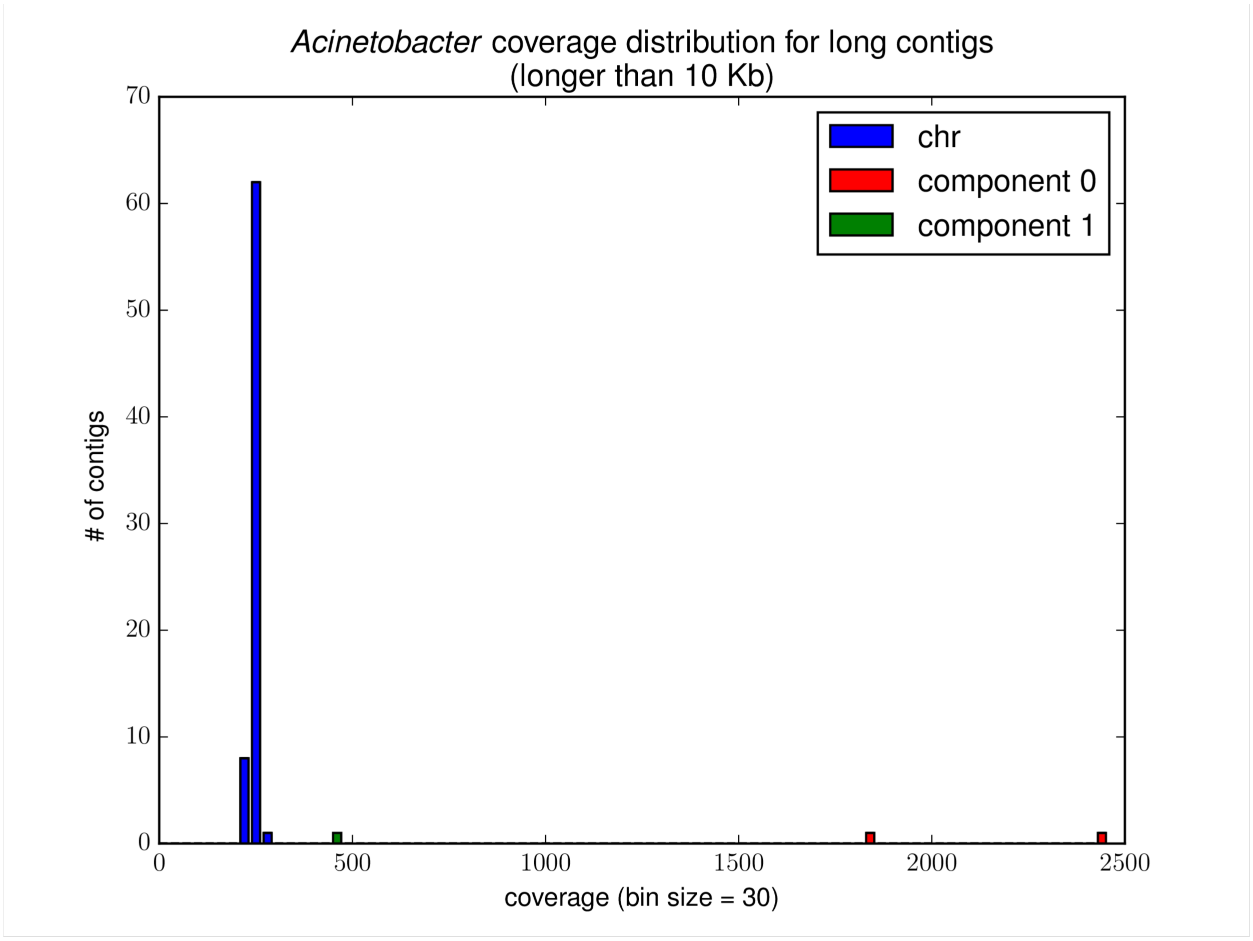

Figure 7. The distribution of *k*-mer coverage for all long contigs in *Aci* dataset (*medianCoverage* = 249). Components 2 and 3 are not shown in this histogram since they do not contain long contigs.

Butyrivibrio sp. INlla16

plasmidSPAdes identified four putative plasmid components in *But* dataset, including three circular putative plasmids (one of them is confirmed). Component 0's best BLAST hit is to another *Butyrivibrio* species plasmid. Components 1, 2 and 3 had no significant BLAST hits to any sequences in NCBI NT database. Since component 1 contains plasmid-specific genes (integrase/recombinase, MobM, transposase), we classified it as a novel plasmid in *But*.

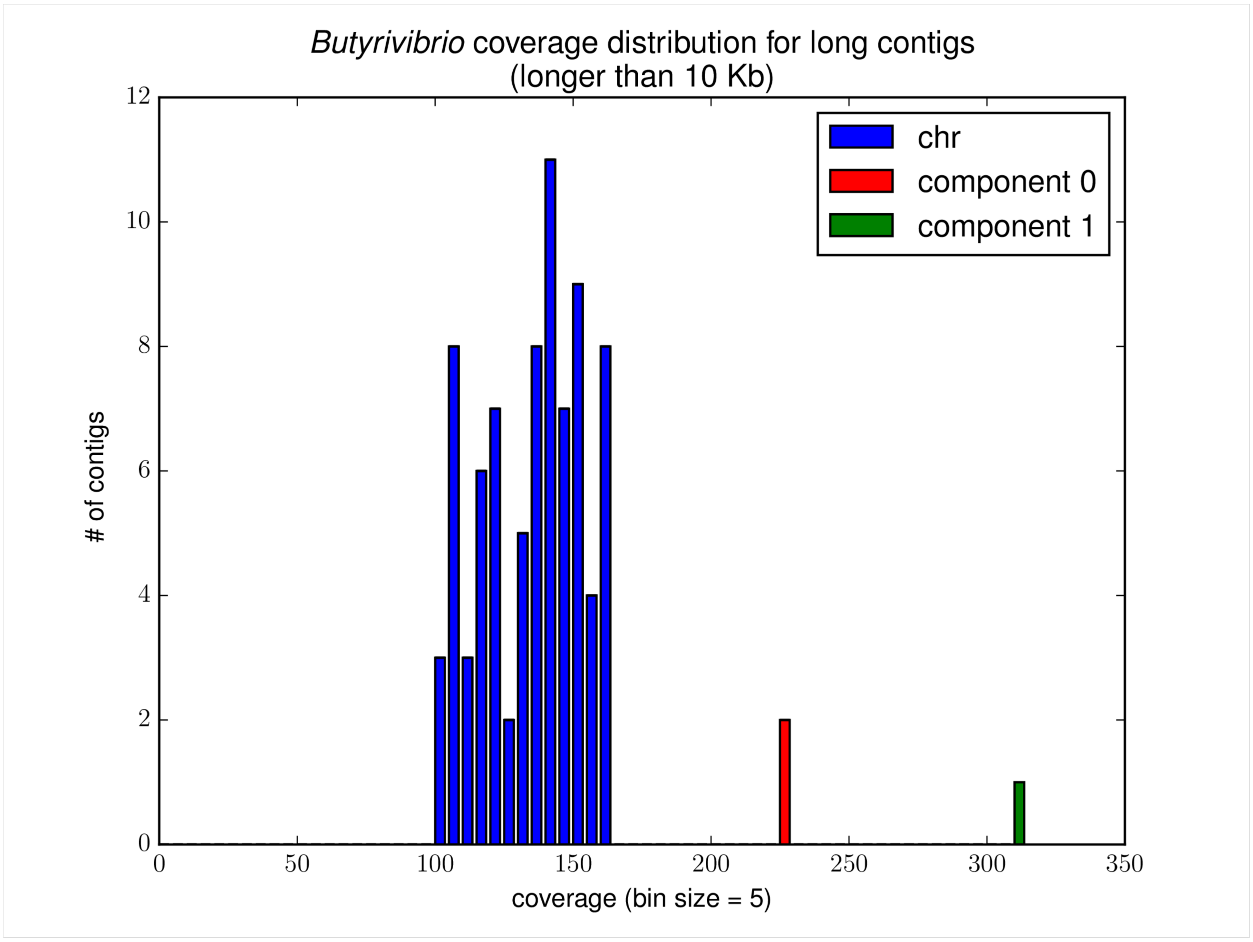

Figure 8. The distribution of *k*-mer coverage for all long contigs in *But* dataset (*medianCoverage* = 130.3). Components 2 and 3 are not shown in this histogram since they do not contain long contigs.

Lachnospiraceae NK3A20

plasmidSPAdes identified a single small putative plasmid component (with no significant match to the NCBI NT database) in *Lac* dataset. Such small putative plasmid components may represent tandem repeats in the genome (supported by the presence of a transposase gene in this component).

.

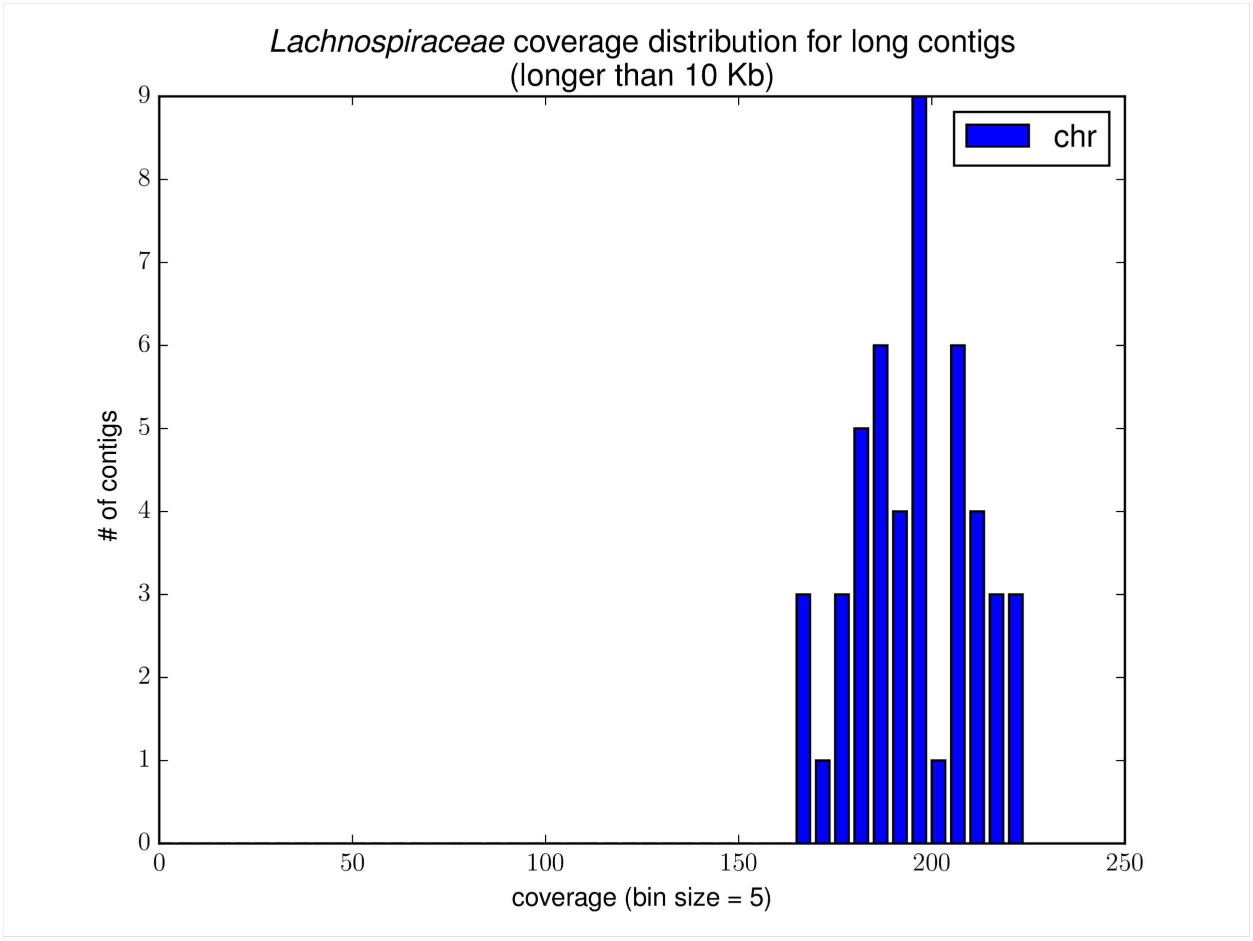

Figure 9. The distribution of *k*-mer coverage for all long contigs in *Lac* dataset (*medianCoverage* = 188.3). Component 0 is not shown in this histogram since it does not contain long contigs.

Luteibacter sp. UNC138MFCol5.1

plasmidSPAdes identified one small putative circular plasmid component in the *Lut* dataset. Since this component contains a transposase gene, it likely represents a mobile element rather than a plasmid.

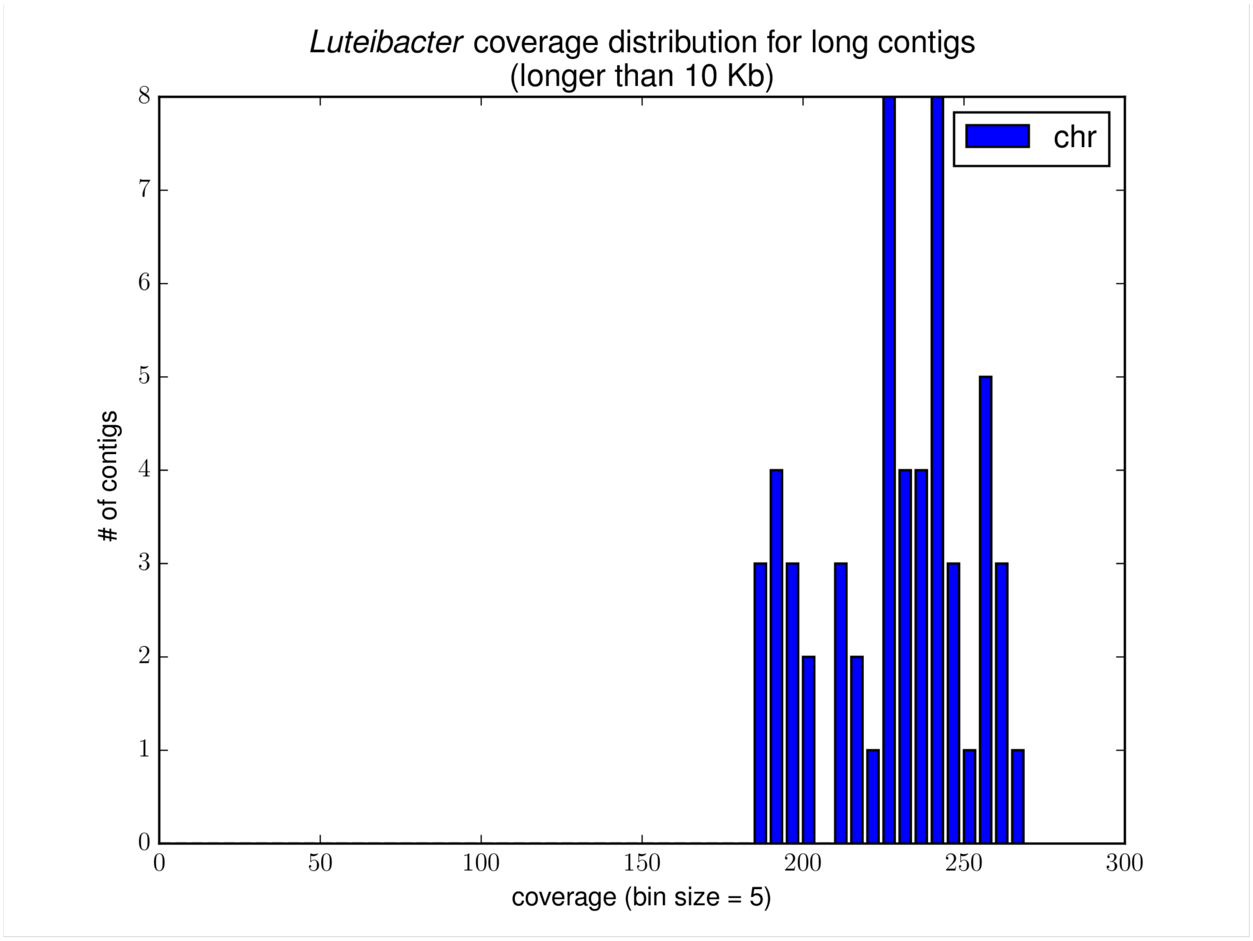

Figure 10. The distribution of *k*-mer coverage for all long contigs in *Lut* dataset (*medianCoverage* = 227). Component 0 is not shown in this histogram since it does not contain long contigs.

Prevotellaceae HUN156

plasmidSPAdes identified no putative plasmid components in *Pre* dataset.

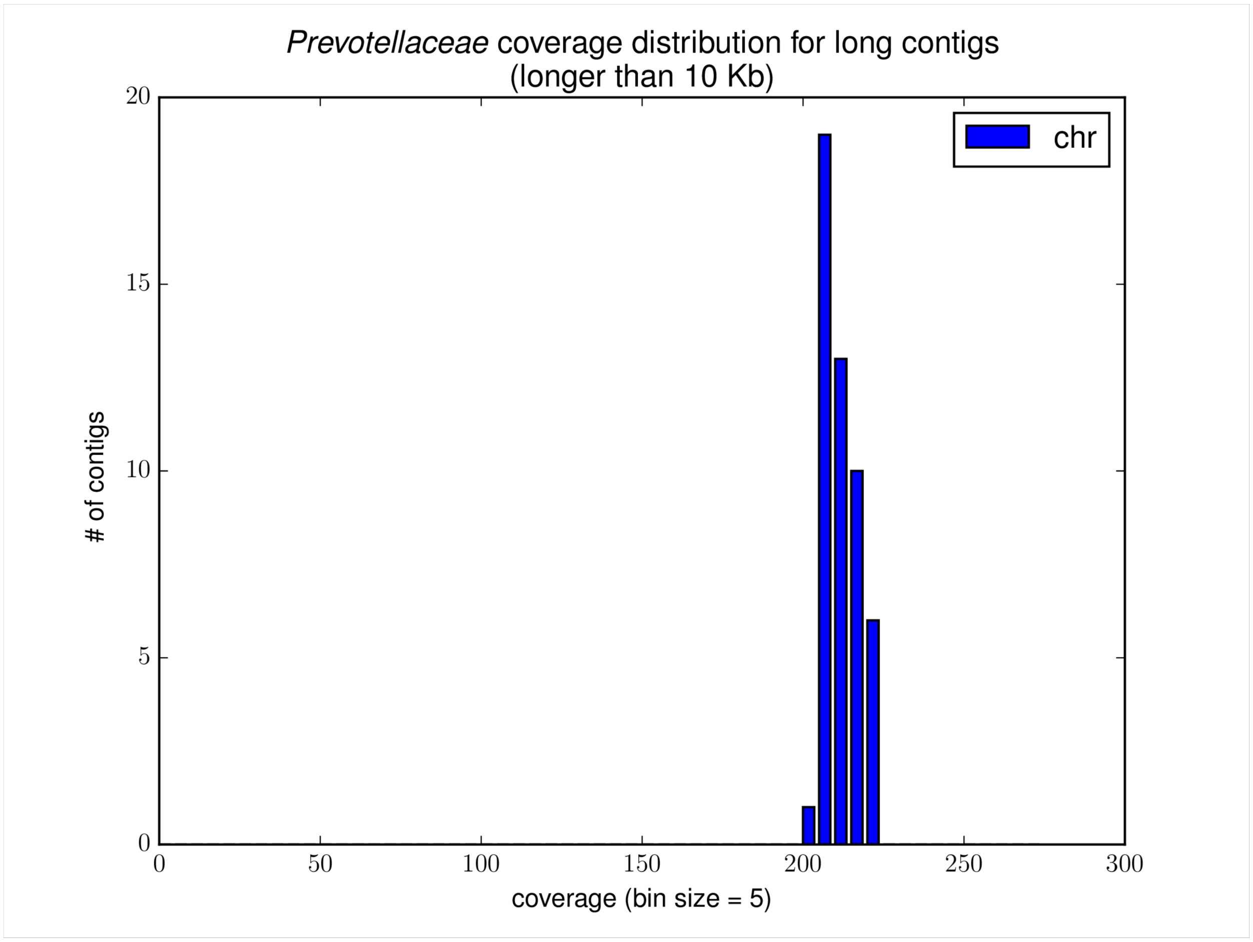

Figure 11. The distribution of *k*-mer coverage for all long contigs in *Pre* dataset (*medianCoverage* = 211).

Pseudoalteromonas Sp. ND6B

plasmidSPAdes identified a single putative plasmid component in *Pse* dataset. Its best BLAST hit is to the chromosome of *Pseudoalteromonas sp. SM9913* and represented alignment of length 22773 with 99% identity. Since this component is composed of short edges (< 10000 nt), it is not represented in the histogram below

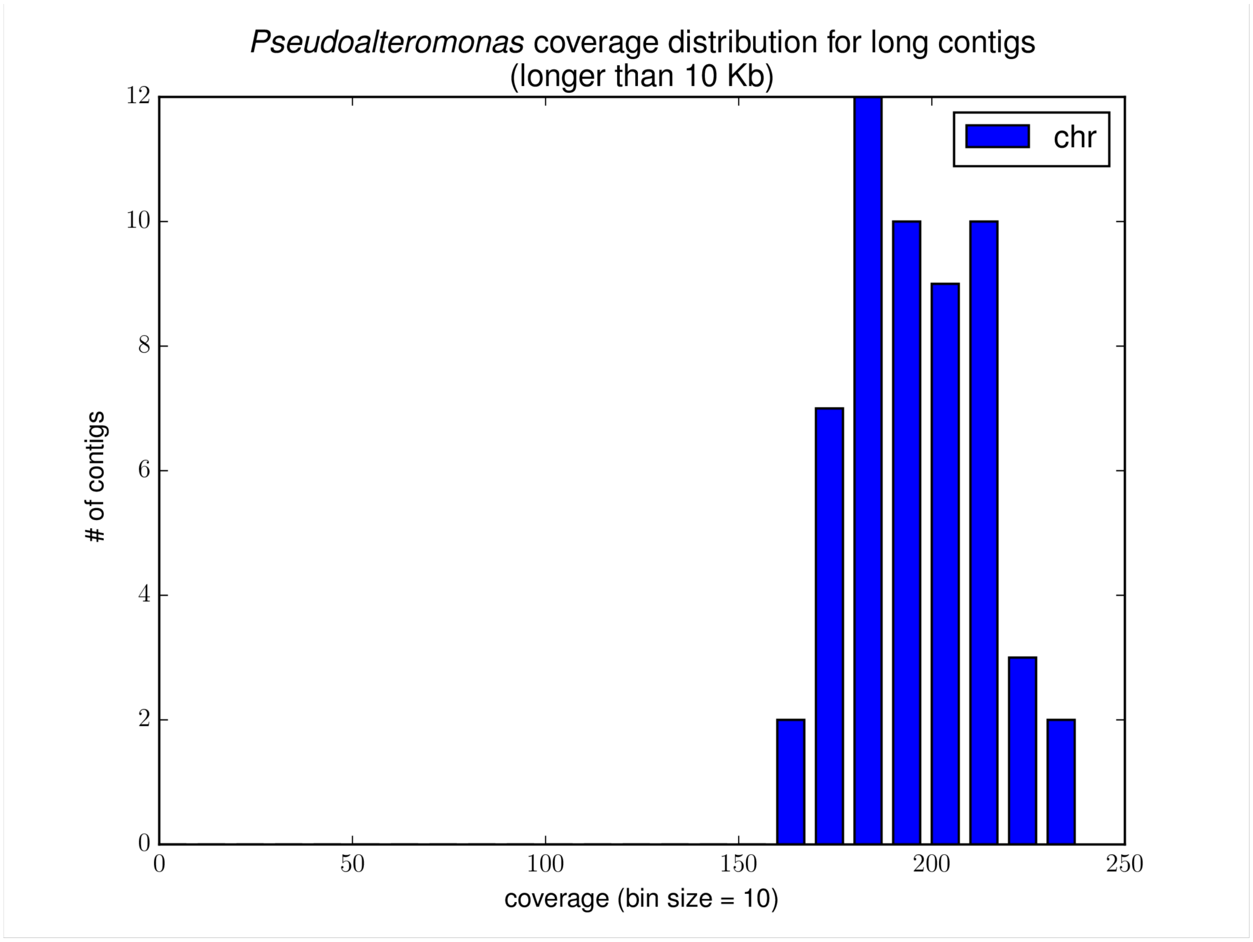

Figure 12. The distribution of *k*-mer coverage for all long contigs in *Pse* dataset (*medianCoverage* = 195.0).

Rhodococcus Sp. J21

plasmidSPAdes identified three putative plasmid components, including one short circular putative plasmid, in *Rho* dataset. Components 0 and 2 have been confirmed. The only significant match of the component 1 is to phages, most significant to *Tsukamurella phage TPA2* (with length 756 and 81% identity).

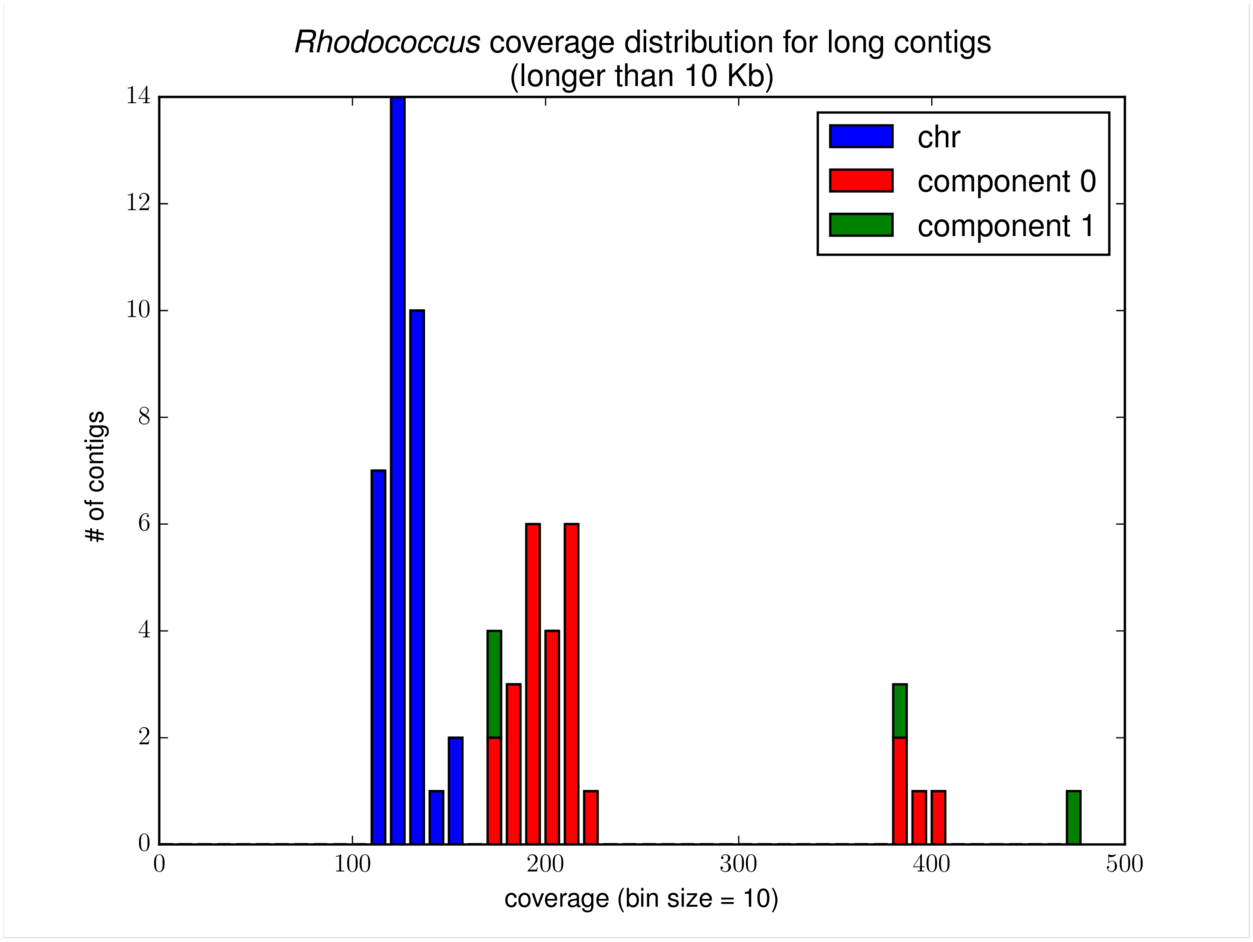

Figure 13. The distribution of *k*-mer coverage for all long contigs in *Rho* dataset (*medianCoverage* = 127). Component 2 is not shown in this histogram since it does not contain long contigs.

Ruminococcus flavefaciens YAD2003

plasmidSPAdes identified a single circular component in the plasmid graph that had no significant BLAST hits against the NCBI NT database. However, this component contains yefM-yoeB toxin-antitoxin system and plasmid replication initiation protein, indicating that it is likely a novel plasmid in *Rum*.

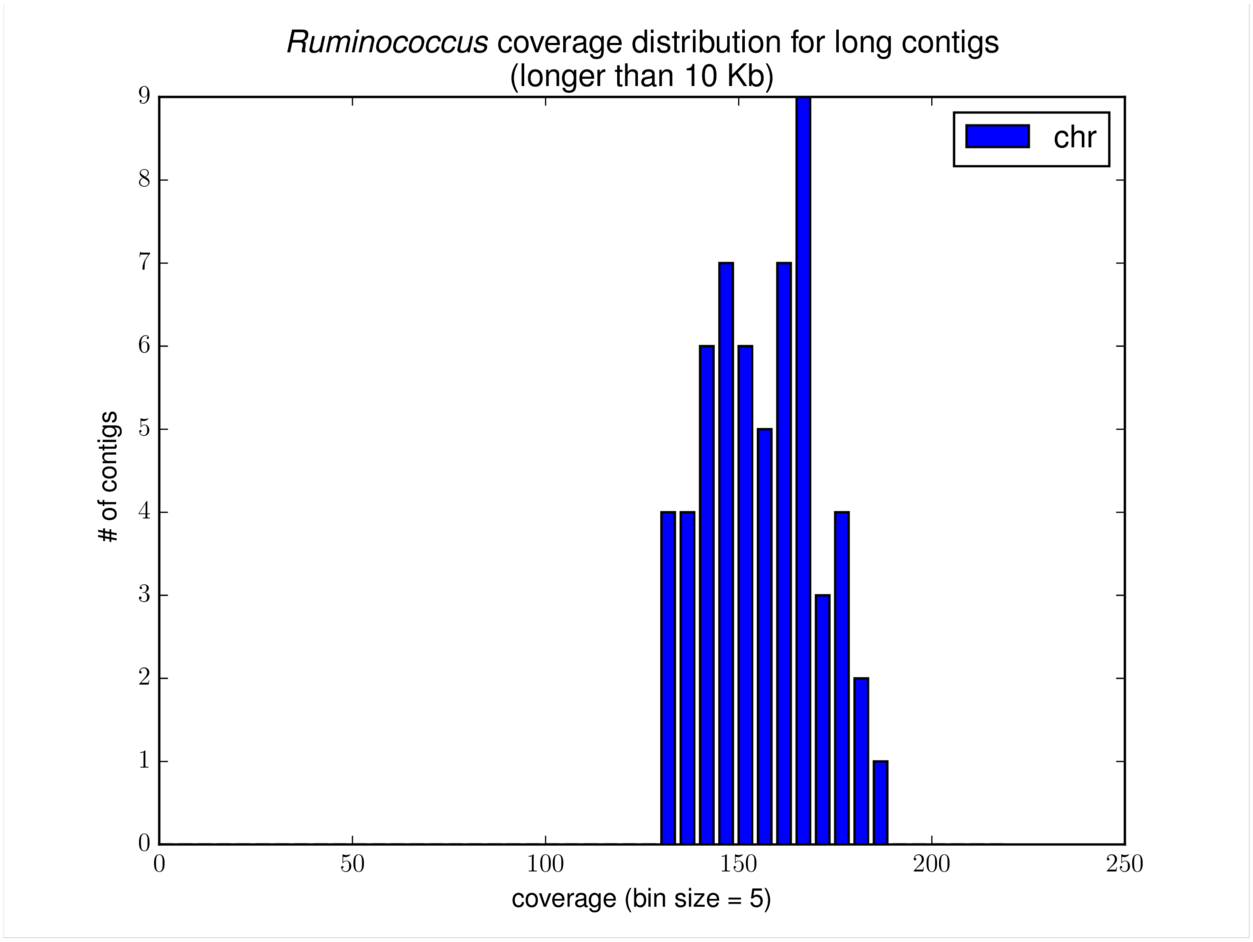

Figure 14. The distribution of *k*-mer coverage for all long contigs in *Rum* dataset (*medianCoverage* = 157.2). Component 0 is not shown in this histogram since it does not contain long contigs.

Sphingomonas Sp. UNC305MFCol5.2

plasmidSPAdes identified no putative plasmid components the *Sph* dataset.

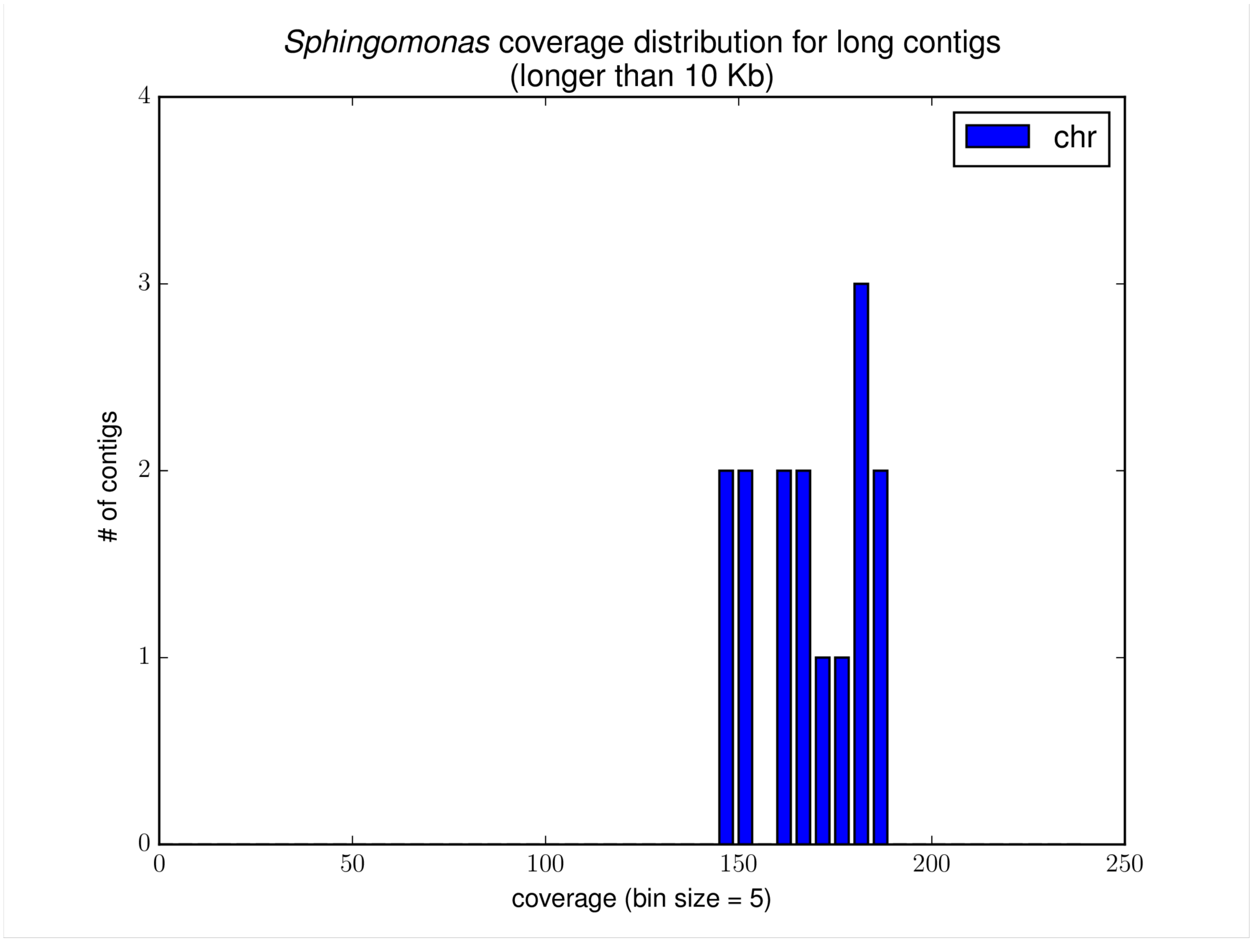

Figure 15. The distribution of *k*-mer coverage for all long contigs in *Sph* dataset (*medianCoverage* = 171).

T. filiformis ATT43280

plasmidSPAdes identified five putative plasmid components in *Tfi* dataset, including three circular putative plasmids (one of them is confirmed). Component 0's best hit was to the chromosome of *Thermus oshimai JL-2*, with alignment length 19206 and 88% identity. Component 1's best hit was to the plasmid *pTHEOS02* of *Thermus oshimai*, with alignment length 1102 and 81% identity (confirmed plasmid). Component 2's best hit was to the chromosome of *Thermus thermophilus HB27*, with alignment length 6408 and 83% identity (rRNA cluster). Blasting component 3 and 4 against the NCBI NT database resulted in no significant hits.

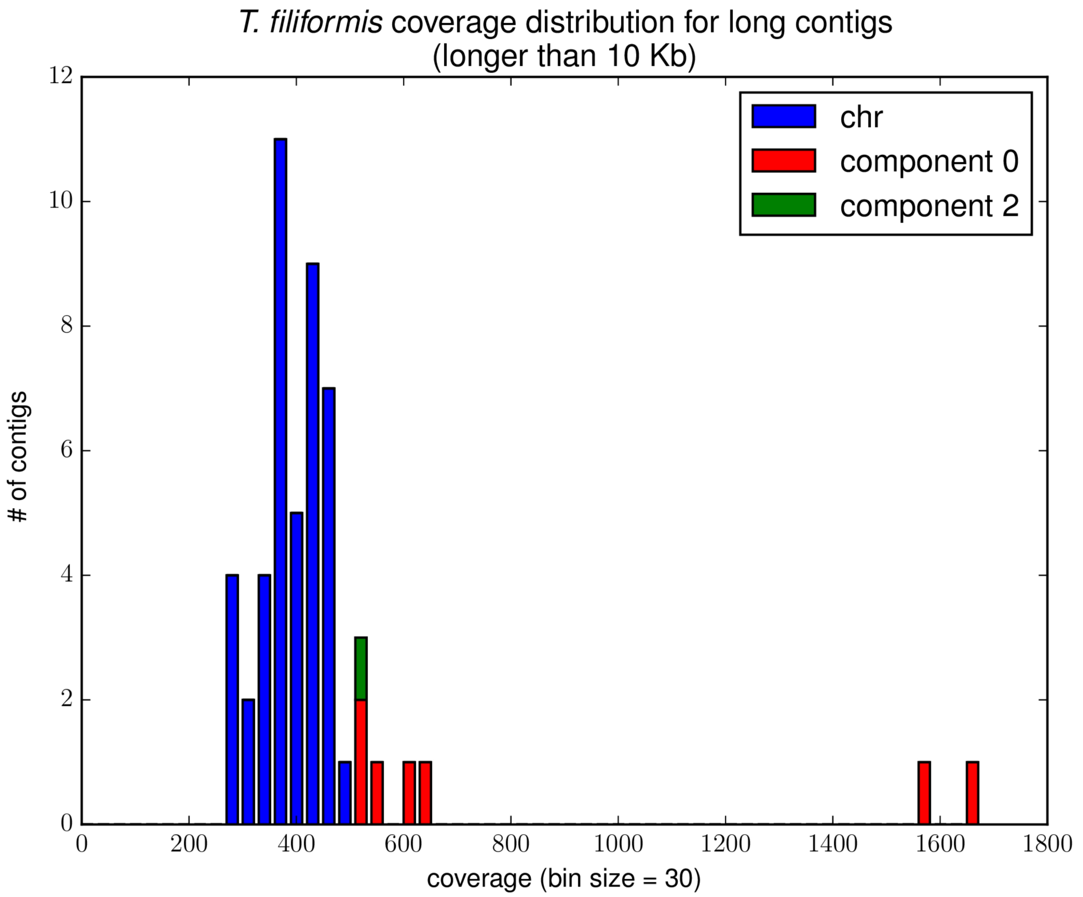

Figure 16. The distribution of *k*-mer coverage for all long contigs in *Tfi* dataset (*medianCoverage* = 388). Component 1, 3, and 4 are not shown in this histogram since they do not contain long contigs.

Appendix: Long contigs coverage distribution with respect to contigs' length

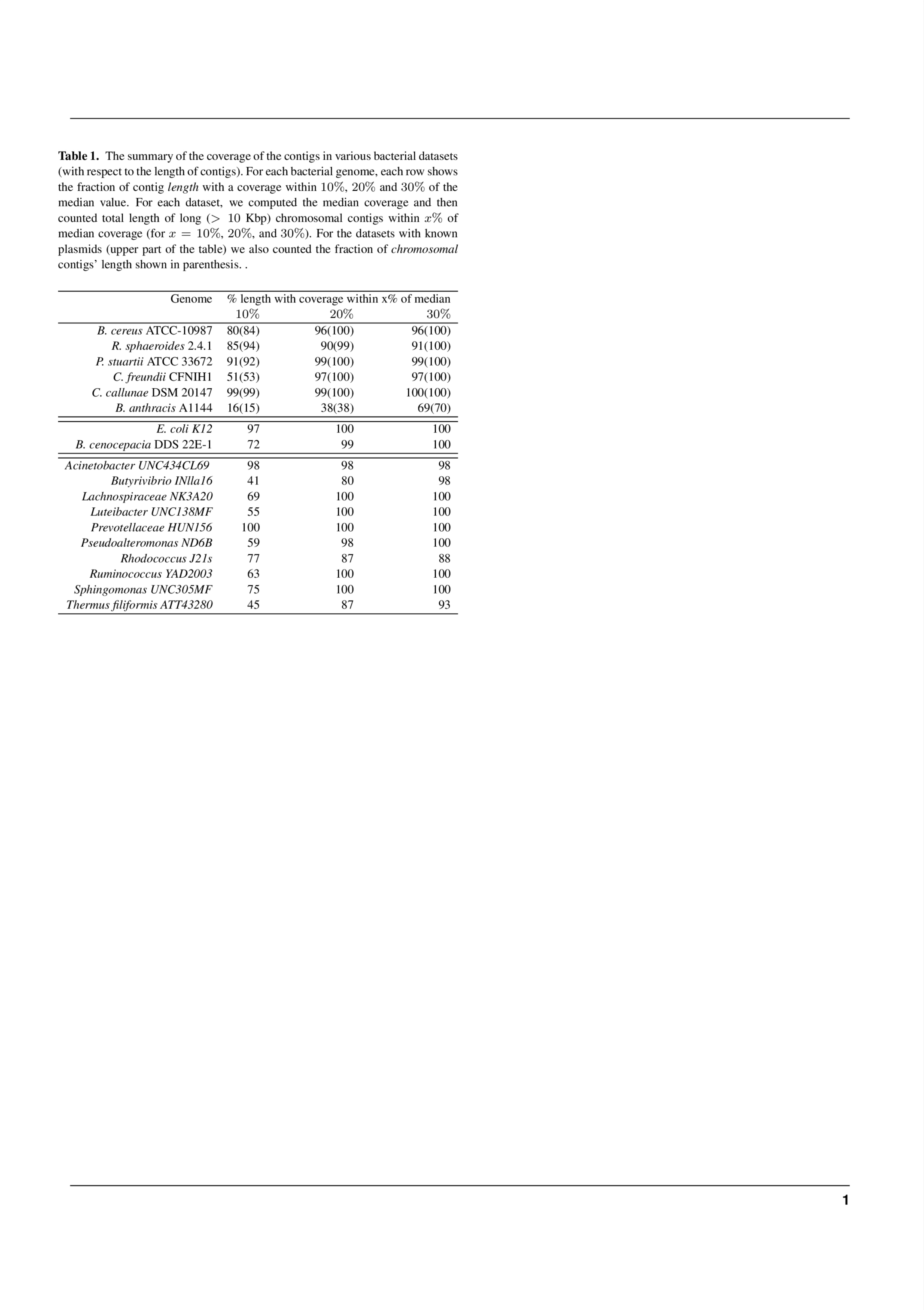

Table 1. The summary of the coverage of the contigs in various bacterial datasets (with respect to the length of contigs). For each bacterial genome, each row shows the fraction of contig *length* with a coverage within 10%, 20% and 30% of the median value. For each dataset, we computed the median coverage and then counted total length of long ($>$ 10 Kbp) chromosomal contigs within x% of median coverage (for x = 10%, 20%, and 30%). For the datasets with known plasmids (upper part of the table) we also counted the fraction of *chromosomal* contigs' length shown in parenthesis. .

Genome	% length with coverage within x% of median		
	10%	20%	30%
B. cereus ATCC-10987	80(84)	96(100)	96(100)
R. sphaeroides 2.4.1	85(94)	90(99)	91(100)
P. stuartii ATCC 33672	91(92)	99(100)	99(100)
C. freundii CFNIH1	51(53)	97(100)	97(100)
C. callunae DSM 20147	99(99)	99(100)	100(100)
B. anthracis A1144	16(15)	38(38)	69(70)
E. coli K12	97	100	100
B. cenocepacia DDS 22E-1	72	99	100
Acinetobacter UNC434CL69	98	98	98
Butyrivibrio INlla16	41	80	98
Lachnospiraceae NK3A20	69	100	100
Luteibacter UNC138MF	55	100	100
Prevotellaceae HUN156	100	100	100
Pseudoalteromonas ND6B	59	98	100
Rhodococcus J21s	77	87	88
Ruminococcus YAD2003	63	100	100
Sphingomonas UNC305MF	75	100	100
Thermus filiformis ATT43280	45	87	93