

MaxBin Supplementary Materials – Wu et al., 2014

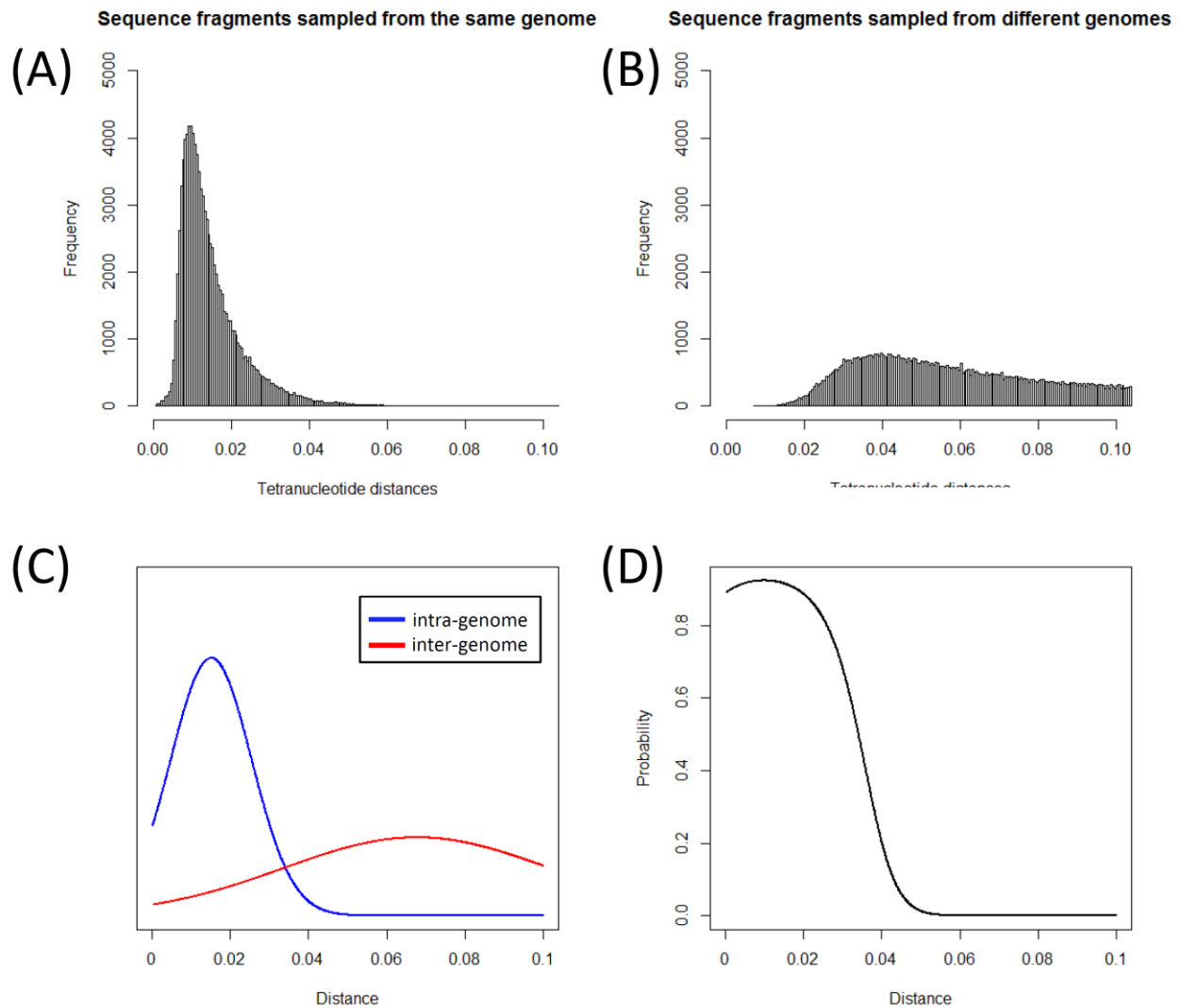


Figure S1. Euclidean distance distributions between sequences extracted from the same genome (intra-genome) and from two different genomes (inter-genome). (A) Histogram of distances between sequences sampled from the same genome. (B) Histogram of distances between sequences sampled from different genomes. (C) Approximate distance distributions of intra- and inter-genome sequences. (D) Probability distribution of two sequences belong to the same genome, which was obtained by

calculating $\frac{P(\text{distance} | \text{intra_genome})}{P(\text{distance} | \text{intra_genome}) + P(\text{distance} | \text{inter_genome})}$. See also Methods for details.

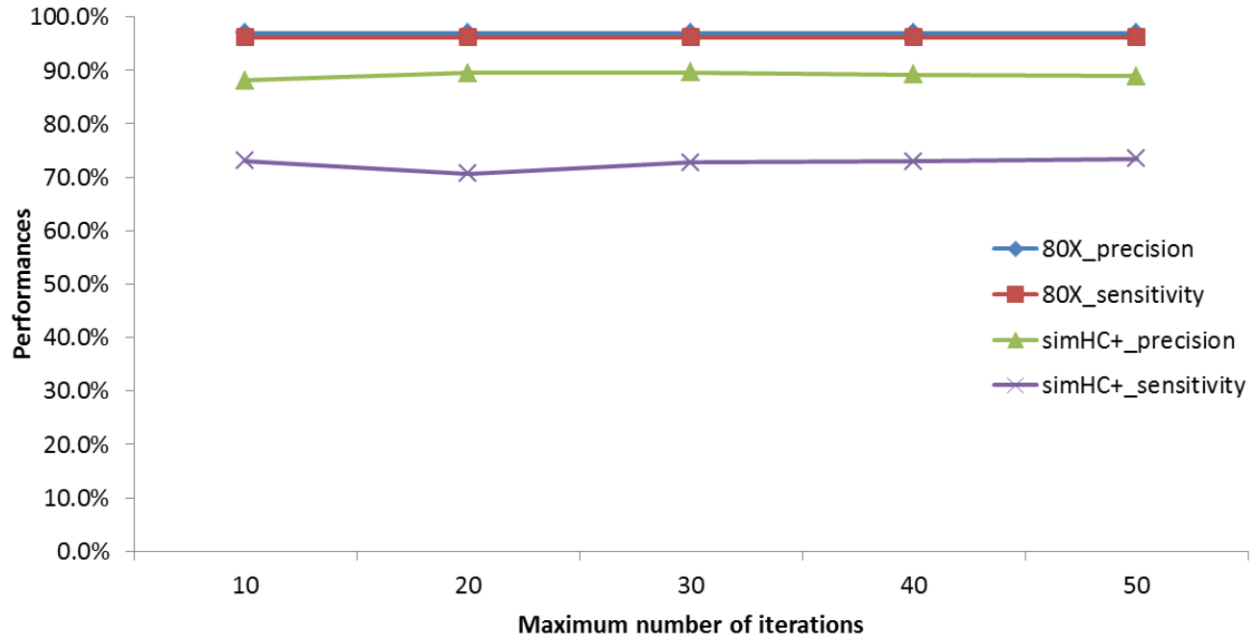


Figure S2. Binning performances (in terms of precision and sensitivity) for different settings of maximum number of iterations to run Expectation-Maximization algorithm.

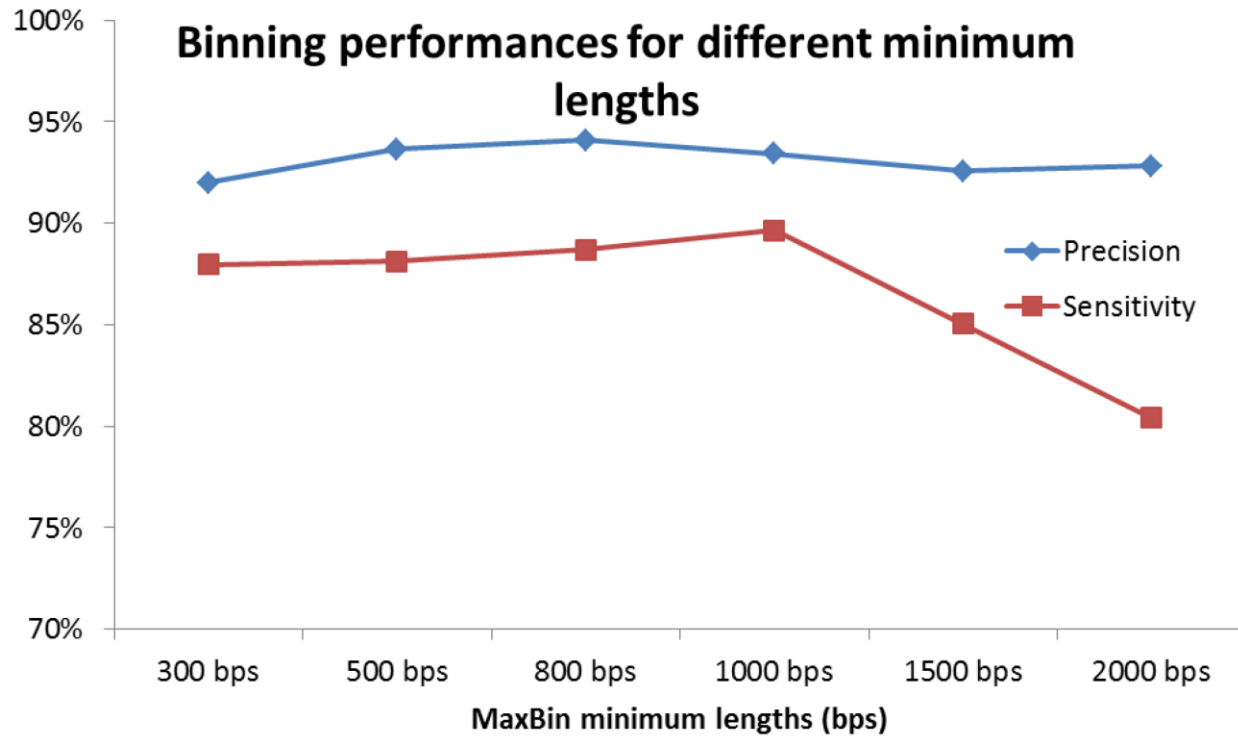
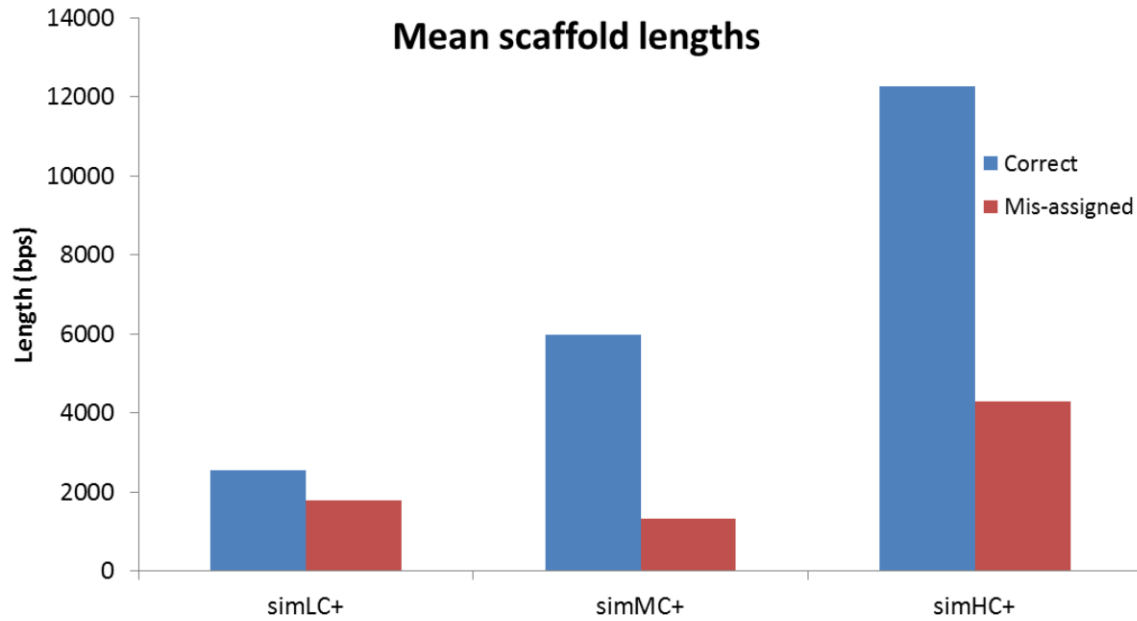


Figure S3. Precision and Sensitivity measured for different MaxBin minimum length cutoff values. The precision and sensitivity were estimated on all simulated datasets (including 80X, 20X, simLC+, simMC+, and simHC+) and were averaged to yield the values.



Dataset		Mean ¹	Std. dev. ¹	Max ¹	Min ¹
simLC+	Correct	2540.12	5662.25	245807	1000
	Mis-assigned	1785.92	1048.93	40810	1000
simMC+	Correct	5988.06	13587.51	180068	1000
	Mis-assigned	1332.06	732.65	39625	1000
simHC+	Correct	12275.15	16735.6	389523	1000
	Mis-assigned	4277.26	5716.08	99124	1000

¹ Length units are all base pairs (bps)

Figure S4. Correctly-classified and mis-assigned scaffolds for the 100-genome simulated datasets, simLC+, simMC+, and simHC+. The distribution of the sequence lengths, including mean, standard deviation, maximum and minimum length values are provided in the table.

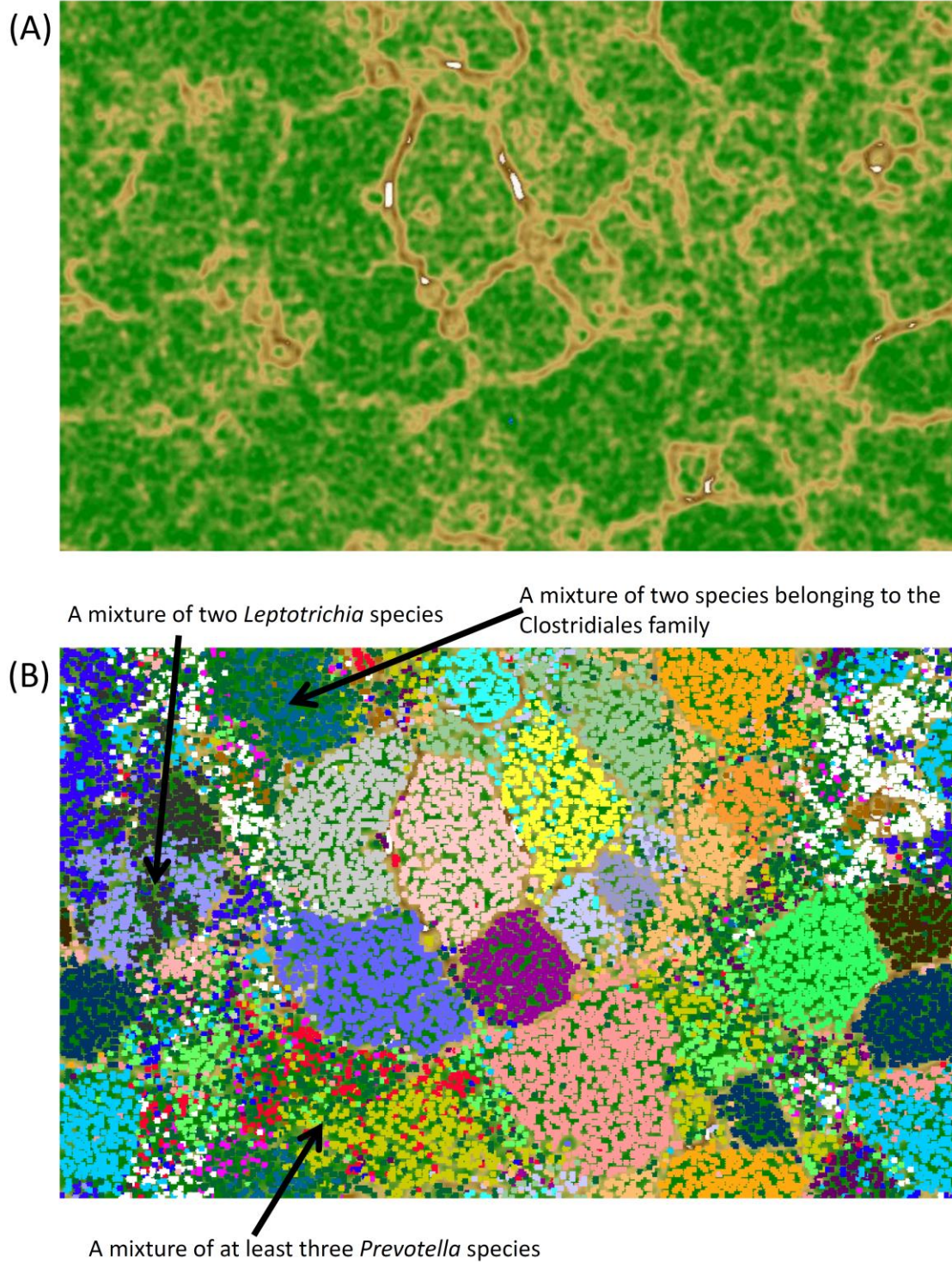


Figure S5. The emergent self-organizing maps (ESOM) binning results of the Human Microbiome Project (HMP) sample SRS013705. (A) Original ESOM plot. (B) Points colored based on MaxBin binning results. Different colors indicate different bins.

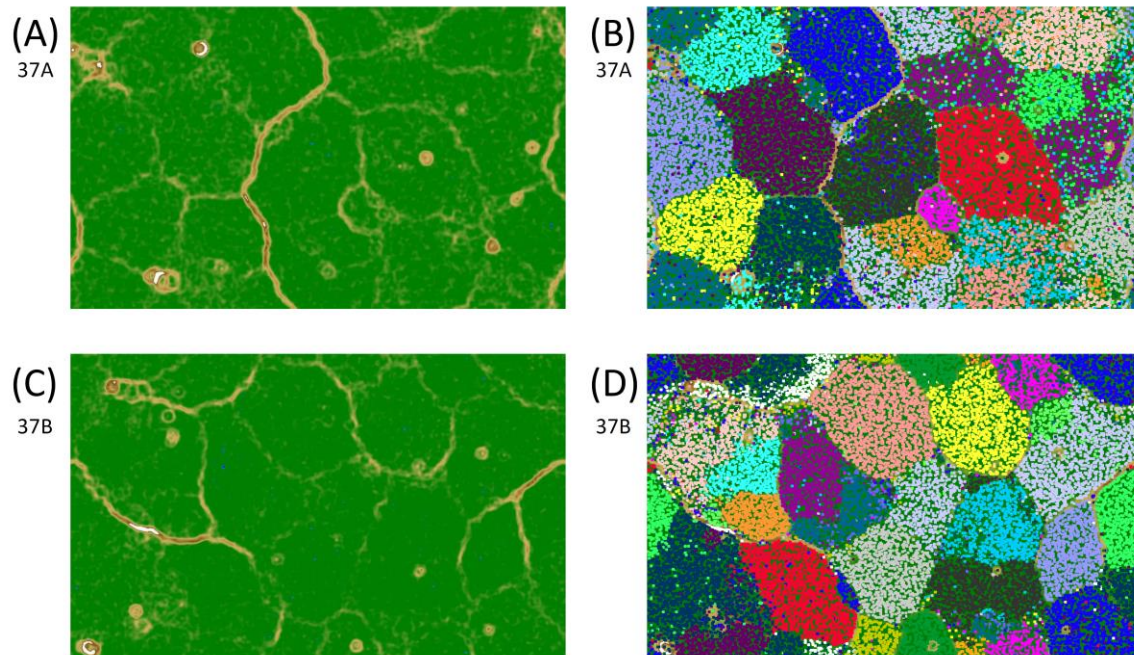
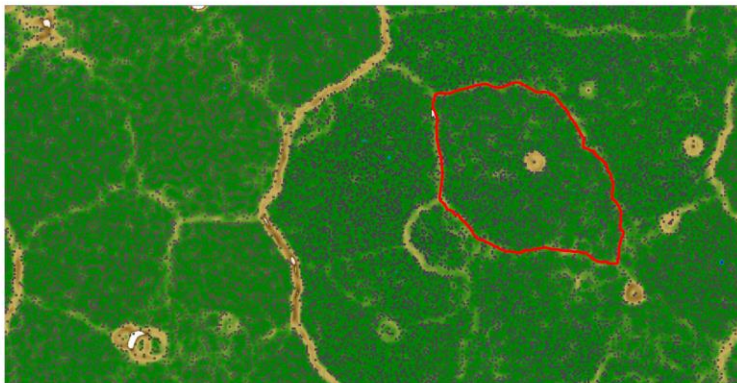


Figure S6. The emergent self-organizing maps (ESOM) binning results of 37A and 37B datasets. (A) Original ESOM plot of 37A. (B) 37A ESOM plot colored by MaxBin binning result. (C) Original ESOM plot of 37B. (D) 37B ESOM plot colored by MaxBin binning result.

(A)



(B)

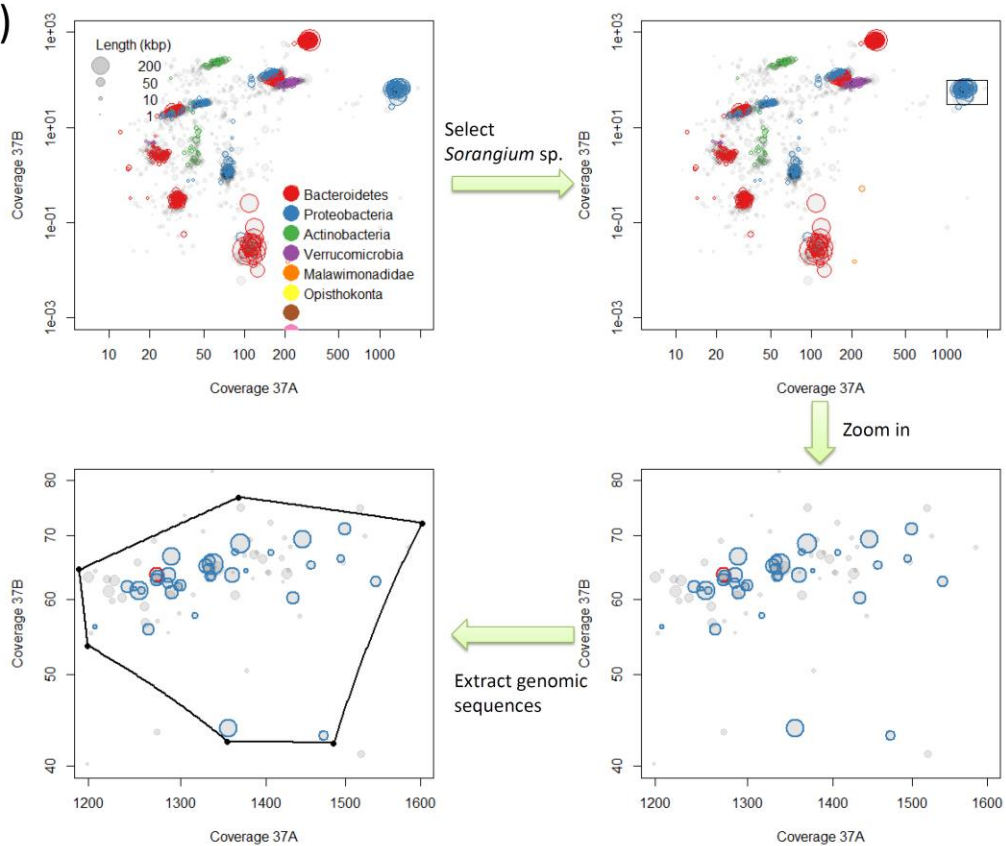


Figure S7. Steps of applying emergent self-organizing maps (ESOM) and differential coverage binning to extract *Sorangium* sp. genomic sequences. (A) Manually select the ESOM region belong to *Sorangium* sp. (B) Steps to draw differential coverage plot, select and zoom in to the region that belongs to *Sorangium* sp., and select the boundaries that include *Sorangium* sp. scaffolds.

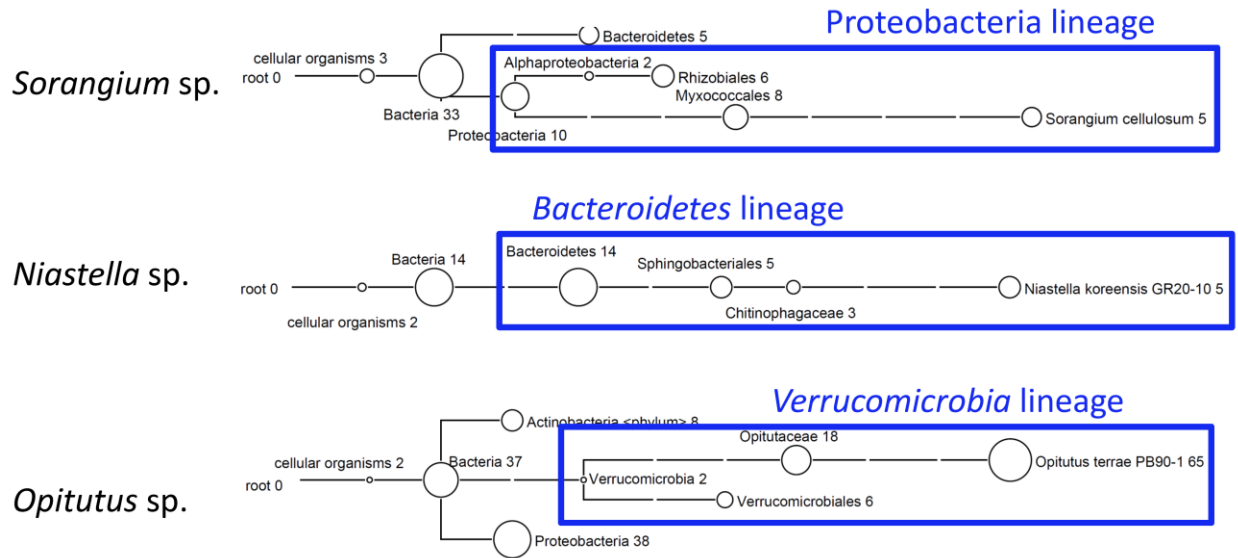
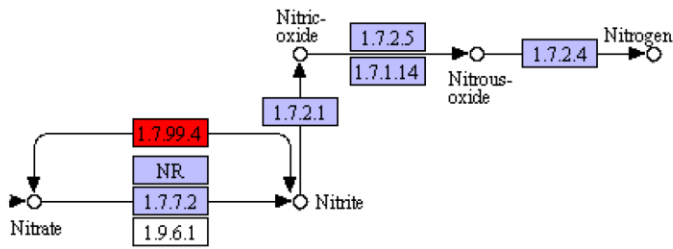
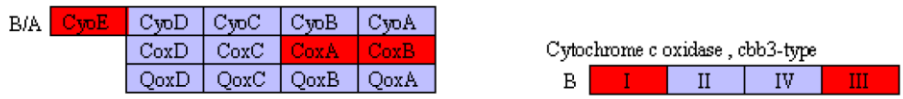


Figure S8. Using MEGAN to find the taxonomy of the scaffold sequences that were binned by MaxBin but missed by differential coverage binning approach. Numbers indicate the number of scaffolds that are assigned to corresponding taxonomy. Blue blocks indicate the lineages that most likely belong to the target species, which are *Sorangium sp.*, *Niastella sp.*, and *Opatutus sp.*

(A) Denitrification



(B) Cytochrome C Oxidase



(C) Sulfate reduction

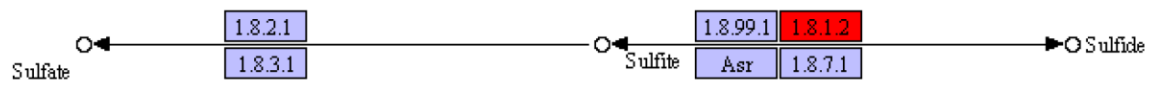


Figure S9. Pathway maps for denitrification, sulfate reduction, and cytochrome c oxidase in *Sorangium* sp. Red blocks indicate that the gene is present in the genome, and blue otherwise.

Table S1. Species and their relative abundance ratios used in 10-genome simulation

NCBI Genbank ID	Species name	Relative Abundance Ratios
256653503	<i>Acetobacter pasteurianus</i>	37.9%
347534971	<i>Flavobacterium branchiophilum</i>	18.9%
261854630	<i>Halothiobacillus neapolitanus</i>	15.2%
327384027	<i>Lactobacillus casei</i>	7.6%
300112745	<i>Nitrosococcus watsonii</i>	3.8%
325278757	<i>Odoribacter splanchnicus</i>	3.8%
111017022	<i>Rhodococcus jostii</i>	3.4%
182682970	<i>Streptococcus pneumoniae</i>	3.4%
307723218	<i>Thermoanaerobacter sp</i>	3.0%
154243958	<i>Xanthobacter autotrophicus</i>	3.0%

Table S2. Species and their relative abundance ratios used in 100-genome simulation.

NCBI Genbank ID	Species name	Relative Abundance Ratios		
		simLC+	simMC+	simHC+
256653503	<i>Acetobacter pasteurianus</i>	29.1%	14.5%	2.3%
330827700	<i>Aeromonas veronii</i>	6.2%	14.5%	1.9%
398314590	<i>Amycolatopsis mediterranei</i>	5.2%	11.6%	1.3%
308175814	<i>Arthrobacter arilaitensis</i>	2.5%	7.0%	1.2%
158421624	<i>Azorhizobium caulinodans</i>	1.4%	4.7%	1.2%
217957581	<i>Bacillus cereus</i>	1.1%	4.3%	1.1%
296500838	<i>Bacillus thuringiensis</i>	0.8%	1.2%	1.1%
42521650	<i>Bdellovibrio bacteriovorus</i>	0.8%	0.6%	1.1%
119025018	<i>Bifidobacterium adolescentis</i>	0.8%	0.6%	1.1%
295793053	<i>Bifidobacterium animalis</i>	0.7%	0.6%	1.1%
343385146	<i>Brachyspira intermedia</i>	0.7%	0.5%	1.1%
15791399	<i>Campylobacter jejuni</i>	0.7%	0.5%	1.1%
71082709	<i>Candidatus Pelagibacter ubique</i>	0.7%	0.5%	1.1%
194246403	<i>Candidatus Phytoplasma mali</i>	0.7%	0.5%	1.1%
256370581	<i>Candidatus Sulcia muelleri</i>	0.7%	0.5%	1.1%
297749010	<i>Chlamydia trachomatis</i>	0.7%	0.5%	1.1%
334694771	<i>Chlamydophila psittaci</i>	0.7%	0.5%	1.1%
325507407	<i>Clostridium acetobutylicum</i>	0.7%	0.5%	1.1%
331268188	<i>Clostridium botulinum</i>	0.7%	0.5%	1.1%
28209834	<i>Clostridium tetani</i>	0.7%	0.5%	1.1%

125972525	<i>Clostridium thermocellum</i>	0.7%	0.5%	1.1%
376247367	<i>Corynebacterium diphtheriae</i>	0.7%	0.5%	1.1%
385806437	<i>Corynebacterium pseudotuberculosis</i>	0.7%	0.5%	1.1%
334695745	<i>Corynebacterium ulcerans</i>	0.7%	0.5%	1.1%
284928601	<i>Cyanobacterium UCYN</i>	0.7%	0.5%	1.1%
307149945	<i>Cyanothece sp</i>	0.7%	0.5%	1.1%
46562128	<i>Desulfovibrio vulgaris</i>	0.7%	0.5%	1.1%
58616727	<i>Ehrlichia ruminantium</i>	0.7%	0.5%	1.1%
378937014	<i>Enterococcus faecium</i>	0.7%	0.5%	1.1%
336065242	<i>Erysipelothrix rhusiopathiae</i>	0.7%	0.5%	1.0%
209917191	<i>Escherichia coli</i>	0.7%	0.5%	1.0%
385805051	<i>Fervidicoccus fontis</i>	0.7%	0.5%	1.0%
302325342	<i>Fibrobacter succinogenes</i>	0.7%	0.5%	1.0%
347534971	<i>Flavobacterium branchiophilum</i>	0.7%	0.5%	1.0%
118496615	<i>Francisella novicida</i>	0.7%	0.5%	1.0%
156501369	<i>Francisella tularensis</i>	0.6%	0.5%	1.0%
19703352	<i>Fusobacterium nucleatum</i>	0.6%	0.5%	1.0%
333392846	<i>Gardnerella vaginalis</i>	0.6%	0.5%	1.0%
322433659	<i>Granulicella tundricola</i>	0.6%	0.5%	1.0%
148826757	<i>Haemophilus influenzae</i>	0.6%	0.5%	1.0%
301154649	<i>Haemophilus parainfluenzae</i>	0.6%	0.5%	1.0%
170717206	<i>Haemophilus somnus</i>	0.6%	0.5%	1.0%
12057215	<i>Halobacterium sp</i>	0.6%	0.5%	1.0%
261854630	<i>Halothiobacillus neapolitanus</i>	0.6%	0.5%	1.0%

261838873	<i>Helicobacter pylori</i>	0.6%	0.5%	1.0%
338736863	<i>Hyphomicrobium sp</i>	0.6%	0.5%	1.0%
385808586	<i>Ignavibacterium album</i>	0.6%	0.5%	1.0%
375256816	<i>Klebsiella oxytoca</i>	0.6%	0.5%	1.0%
332290650	<i>Krokinobacter sp</i>	0.6%	0.5%	1.0%
116332681	<i>Lactobacillus brevis</i>	0.6%	0.5%	1.0%
327384027	<i>Lactobacillus casei</i>	0.6%	0.5%	1.0%
104773257	<i>Lactobacillus delbrueckii</i>	0.6%	0.5%	1.0%
94986445	<i>Lawsonia intracellularis</i>	0.6%	0.5%	1.0%
296105497	<i>Legionella pneumophila</i>	0.6%	0.5%	1.0%
330833867	<i>Metallosphaera cuprina</i>	0.6%	0.5%	0.9%
124484829	<i>Methanocorpusculum labreanum</i>	0.6%	0.5%	0.9%
19918815	<i>Methanosarcina acetivorans</i>	0.6%	0.5%	0.9%
73667559	<i>Methanosarcina barkeri</i>	0.6%	0.5%	0.9%
239916571	<i>Micrococcus luteus</i>	0.6%	0.5%	0.9%
356592064	<i>Mycobacterium bovis</i>	0.6%	0.5%	0.9%
108796981	<i>Mycobacterium sp</i>	0.6%	0.5%	0.9%
400276727	<i>Mycoplasma gallisepticum</i>	0.6%	0.5%	0.9%
330723203	<i>Mycoplasma hyorhinis</i>	0.6%	0.5%	0.9%
308388224	<i>Neisseria meningitidis</i>	0.6%	0.5%	0.9%
300112745	<i>Nitrosococcus watsonii</i>	0.6%	0.5%	0.9%
325980881	<i>Nitrosomonas sp</i>	0.6%	0.5%	0.9%
54021964	<i>Nocardia farcinica</i>	0.6%	0.5%	0.9%
325278757	<i>Odoribacter splanchnicus</i>	0.6%	0.5%	0.9%

386720569	<i>Paenibacillus mucilaginosus</i>	0.6%	0.4%	0.9%
261403876	<i>Paenibacillus sp</i>	0.6%	0.4%	0.9%
54307237	<i>Photobacterium profundum</i>	0.5%	0.4%	0.9%
126695337	<i>Prochlorococcus marinus</i>	0.5%	0.4%	0.9%
347537839	<i>Pseudogulbenkiania sp</i>	0.5%	0.4%	0.9%
313496345	<i>Pseudomonas putida</i>	0.5%	0.4%	0.9%
116249766	<i>Rhizobium leguminosarum</i>	0.5%	0.4%	0.9%
111017022	<i>Rhodococcus jostii</i>	0.5%	0.4%	0.9%
380760311	<i>Rickettsia prowazekii</i>	0.5%	0.4%	0.9%
378722019	<i>Rickettsia rickettsii</i>	0.5%	0.4%	0.9%
374318767	<i>Rickettsia slovaca</i>	0.5%	0.4%	0.9%
99079841	<i>Ruegeria sp</i>	0.5%	0.4%	0.9%
194447306	<i>Salmonella enterica</i>	0.5%	0.4%	0.9%
269118642	<i>Sebaldella termitidis</i>	0.5%	0.4%	0.9%
114045513	<i>Shewanella sp</i>	0.5%	0.4%	0.9%
30061571	<i>Shigella flexneri</i>	0.5%	0.4%	0.9%
85057978	<i>Sodalis glossinidius</i>	0.4%	0.4%	0.9%
311222926	<i>Staphylococcus aureus</i>	0.4%	0.4%	0.9%
182682970	<i>Streptococcus pneumoniae</i>	0.4%	0.4%	0.9%
28894912	<i>Streptococcus pyogenes</i>	0.4%	0.4%	0.9%
354984442	<i>Streptococcus suis</i>	0.4%	0.4%	0.9%
116626972	<i>Streptococcus thermophilus</i>	0.4%	0.3%	0.9%
290954631	<i>Streptomyces scabiei</i>	0.4%	0.3%	0.9%
51891138	<i>Symbiobacterium thermophilum</i>	0.4%	0.3%	0.9%

320114857	<i>Thermoanaerobacter brockii</i>	0.4%	0.3%	0.9%
307723218	<i>Thermoanaerobacter sp</i>	0.4%	0.3%	0.9%
242397997	<i>Thermococcus sibiricus</i>	0.4%	0.3%	0.8%
239819985	<i>Variovorax paradoxus</i>	0.4%	0.3%	0.8%
323436265	<i>Weeksella virosa</i>	0.4%	0.3%	0.8%
225629872	<i>Wolbachia sp</i>	0.4%	0.3%	0.8%
154243958	<i>Xanthobacter autotrophicus</i>	0.4%	0.3%	0.7%
162418099	<i>Yersinia pestis</i>	0.3%	0.3%	0.4%

Table S3. Closest reference species of the binned genome from three Human Microbiom Project (HMP) samples

Bin	Closest reference species	Genome coverage	Completeness	Genome size	GC%
SRS013705					
001	<i>Prevotella melaninogenica</i>	313.04	40.20%	1806112	40
002	<i>Streptococcus salivarius</i>	163.75	50.50%	1700440	36
003	<i>Neisseria flavescens</i>	123.35	52.30%	1583928	49
004	<i>Porphyromonas</i> sp. oral taxon 279	117.61	62.60%	1619277	60
005	<i>Actinomyces graevenitzi</i>	95.36	72.90%	1949837	58
006	<i>Streptococcus parasanguinis</i>	84.76	33.60%	2162969	41
007	<i>Veillonella atypica</i>	53.88	85.00%	3759355	36
008	<i>Leptotrichia buccalis</i>	49.1	68.20%	2550570	33
009	<i>Stomatobaculum longum</i>	46.03	67.30%	1773751	55
010	<i>Prevotella pallens</i>	42.53	89.70%	3702189	45
011	<i>Actinomyces</i> sp. ICM47	38.78	21.50%	2214274	65
012	<i>Leptotrichia hofstadii</i>	31.53	36.40%	2300441	29
013	<i>Treponema denticola</i>	31.37	98.10%	2148366	41

014	<i>Capnocytophaga</i> sp. CM59	26.3	86.90%	2727613	43
015	<i>Mogibacterium</i> sp. CM50	24.72	27.10%	819989	39
016	<i>Oribacterium</i> sp. ACB1	23.39	10.30%	1608463	43
017	<i>Johnsonella ignava</i>	22.28	69.20%	1882961	40
018	[<i>Eubacterium</i>] <i>infirmum</i>	21.37	67.30%	944042	42
019	<i>Gemella sanguinis</i>	21.32	61.70%	3475425	30
020	<i>Prevotella</i> sp. oral taxon 306	18.51	56.10%	3845362	45
021	<i>Lachnospiraceae</i> bacterium ICM7	17.17	31.80%	4836083	35
022	<i>Megasphaera micronuciformis</i>	16.16	65.40%	773971	46
023	<i>Capnocytophaga</i> sp. oral taxon 329	14.83	77.60%	3082884	41
024	<i>Prevotella salivae</i>	13.41	68.20%	3902144	41
025	<i>Lachnospiraceae</i> bacterium ICM7	12.91	57.00%	1955251	36
026	<i>Treponema vincentii</i>	12.86	92.50%	2614737	45
027	uncultured bacterium (gcode 4)	11.81	27.10%	1196134	36
028	<i>Peptostreptococcus stomatis</i>	10.99	32.70%	899206	39
029	<i>Prevotella</i> sp. oral taxon 473	10.78	65.40%	3294481	57
030	<i>Selenomonas sputigena</i>	10.75	45.80%	1189325	58

031	<i>Butyrivibrio proteoclasticus</i>	10.27	55.10%	3209210	49
<hr/>					
SRS014477					
001	<i>Bacteroidetes</i> oral taxon 274	15.33	86.90%	4365854	44
002	<i>Treponema denticola</i>	14.03	15.90%	1560979	32
003	<i>Corynebacterium matruchotii</i>	12.82	80.40%	3673816	55
004	<i>Treponema vincentii</i>	11.77	59.80%	4473939	46
<hr/>					
SRS018705					
001	<i>Dialister invisus</i>	198.47	59.80%	4521428	57
002	<i>Haloplasma contractile</i>	114.01	88.80%	1406648	29
003	<i>Ruminococcus bromii</i>	54.1	97.20%	2746565	41
004	<i>Subdoligranulum variabile</i>	48.21	77.60%	4978007	58
005	<i>Parabacteroides distasonis</i>	23.3	77.60%	3960717	46
006	<i>Bacteroides cellulosilyticus (fragilis)</i>	18.64	52.30%	9401347	46
007	<i>Bacteroides nordii</i>	16.11	40.20%	7278618	39
008	<i>Akkermansia muciniphila</i>	14.73	87.90%	2714789	57
009	<i>Alistipes shahii</i>	14.56	49.50%	4516450	60

010	<i>Clostridium leptum</i>	9.88	68.20%	1445238	48
-----	---------------------------	------	--------	---------	----

Table S4. Complete list of species binned for 37A and 37B.

37A							
Bin#	Species	AA identity to closest species	Coverage	Relative abundance ratio (%)	Completeness	Genome size (bps)	GC content
001	<i>Sorangium</i> sp.	60.7% to <i>Sorangium cellulosum</i>	1144.24	48.1%	95.30%	5001615	64
002	<i>Niastella</i> sp.	73.1% to <i>Niastella koreensis</i>	270.36	11.4%	95.30%	4340643	48
003	<i>Opitutus</i> sp.	74.6% to <i>Opitutus terrae</i>	178.73	7.5%	84.10%	4971920	64
004	<i>Chitinophaga</i> sp.	64.9% to <i>Chitinophaga pinensis</i>	153.87	6.5%	98.10%	2915301	38
005	<i>Rhodanobacter</i> sp.	69.9% to <i>Rhodanobacter</i> sp. 115	130.42	5.5%	88.80%	3116221	69
006	<i>Cytophaga</i> sp.	62.9% to <i>Cytophaga hutchinsonii</i>	102.83	4.3%	98.10%	4813206	43
007	<i>Opitutus</i> sp.	74.5% to <i>Opitutus terrae</i>	63.95	2.7%	60.70%	2860184	65
008	<i>Oceanibaculum</i> sp.	65.6% to <i>Oceanibaculum indicum</i>	61.58	2.6%	86.00%	4800981	70
009	<i>Pelagibacterium</i> sp.	68.2% to <i>Pelagibacterium halotolerans</i>	38.24	1.6%	66.40%	3835464	66
010	<i>Cellulomonas</i> sp.	65.2% to <i>Cellulomonas fimi</i>	31.1	1.3%	89.70%	2906911	66

011	<i>Niastella</i> sp.	67.2% to <i>Niastella koreensis</i>	29.35	1.2%	92.50%	3269393	41
012	<i>Niastella</i> sp.	67.8% to <i>Niastella koreensis</i>	29.32	1.2%	94.40%	4579505	49
013	<i>Weeksella</i> sp.	66.2% to <i>Weeksella virosa</i>	23.89	1.0%	92.50%	3132025	39
014	<i>Parvibaculum</i> sp.	80.4% to <i>Parvibaculum lavamentivorans</i>	22.8	1.0%	80.40%	4114944	60
015	<i>Cellulomonas</i> sp.	73.6% to <i>Cellulomonas fimi</i>	21.75	0.9%	48.60%	1341252	72
016	<i>Mesorhizobium</i> sp.	75.7% to <i>Mesorhizobium alhagi</i>	21.51	0.9%	48.60%	2579242	68
017	<i>Opitutus</i> sp.	71.8% to <i>Opitutus terrae</i>	19.01	0.8%	38.30%	3499670	62
018	<i>Dyadobacter</i> sp.	67.5% to <i>Dyadobacter fermentans</i>	18.56	0.8%	76.60%	4364829	44
019	<i>Pseudoxanthomonas</i> sp.	87.4% to <i>Pseudoxanthomonas suwonensis</i>	18.32	0.8%	21.50%	1032550	70

37B

Bin#	Species	AA identity to closest species	Coverage	Relative abundance ratio	Completeness	Genome size (bps)	GC content
001	<i>Niastella</i> sp.	73.2% to <i>Niastella koreensis</i>	622.31	26.6%	94.40%	4338602	48
002	<i>Teredinibacter</i> sp.	67.4% to <i>Teredinibacter</i>	296.81	12.7%	99.10%	4268546	51

		<i>turnerae</i>					
003	<i>Sphingomonas</i> sp.	71.8% to <i>Sphingomonas wittichii</i>	233.2	10.0%	96.30%	2928274	68
004	<i>Cellulomonas</i> sp.	64.9% to <i>Cellulomonas fimi</i>	184.87	7.9%	83.20%	1907226	71
005	<i>Cellulomonas</i> sp.	65.8% to <i>Cellulomonas fimi</i>	181.24	7.8%	92.50%	2549936	66
006	<i>Chitinophaga</i> sp.	65.0 % to <i>Chitinophaga pinensis</i>	112.51	4.8%	98.10%	3590649	41
007	<i>Rhodanobacter</i> sp.	69.2% to <i>Rhodanobacter</i> sp. 115	107.18	4.6%	46.70%	2825683	69
008	<i>Pseudoxanthomonas</i> sp.	71.4% to <i>Pseudoxanthomonas suwonensis</i>	105.12	4.5%	95.30%	3591875	66
009	<i>Opiritatus</i> sp.	74.6% to <i>Opiritatus terrae</i>	69.14	3.0%	92.50%	5888819	64
010	<i>Sorangium</i> sp.	60.0% to <i>Sorangium cellulosum</i>	58.69	2.5%	93.50%	5113973	64
011	<i>Fluviicola</i> sp.	60.7% to <i>Fluviicola taffensis</i>	51.75	2.2%	98.10%	3781707	64
012	<i>Micavibrio</i> sp.	61.5% to <i>Micavibrio aeruginosavorus</i>	33.68	1.4%	99.10%	2453873	59
013	<i>Cytophaga</i> sp.	64.2% to <i>Cytophaga hutchinsonii</i>	33.17	1.4%	94.40%	5812044	39

014	<i>Opitutus</i> sp.	69.7% to <i>Opitutus terrae</i>	28.08	1.2%	93.50%	3702696	59
015	<i>Pelagibacterium</i> sp.	68.5% to <i>Pelagibacterium halotolerans</i>	26.42	1.1%	51.40%	2575205	66
016	<i>Cellulomonas</i> sp.	82.4% to [<i>Cellvibrio</i>] <i>gilvus</i>	23.09	1.0%	47.70%	2578944	74
017	<i>Niastella</i> sp.	67.5% to <i>Niastella koreensis</i>	21.06	0.9%	94.40%	3176136	41
018	<i>Rhodopseudomonas</i> sp.	77.7% to <i>Rhodopseudomonas palustris</i>	18.66	0.8%	57.00%	3646122	68
019	<i>Pelagibacterium</i> sp.	76.6% to <i>Pelagibacterium halotolerans</i>	18.61	0.8%	29.90%	484089	64
020	<i>Parvibaculum</i> sp.	81.2% to <i>Parvibaculum lavamentivorans</i>	16.67	0.7%	58.90%	1815727	60
021	<i>Sphingobacterium</i> sp.	71.1% to <i>Sphingobacterium</i> sp. 21	16.33	0.7%	22.40%	2941390	51
022	<i>Sphingobacterium</i> sp.	73.4% to <i>Sphingobacterium spiritivorum</i>	15.95	0.7%	62.60%	2980773	45
023	<i>Mesorhizobium</i> sp.	75.7% to <i>Mesorhizobium alhagi</i>	15.94	0.7%	53.30%	1305609	64
024	<i>Opitutus</i> sp.	69.6% to <i>Opitutus terrae</i>	15.51	0.7%	67.30%	2297225	68
025	<i>Phycisphaera</i> sp.	62.0% to <i>Phycisphaera</i>	15.29	0.7%	58.90%	3008188	59

		<i>mikurensis</i>					
026	<i>Pseudoxanthomonas</i> sp.	86.2% to	14.36	0.6%	15.00%	1305733	69
		<i>Pseudoxanthomonas</i>					
		<i>suwonensis</i>					

Table S5. Comparison between the binning results of *Niastella* sp. and *Opitutus* sp. produced by MaxBin and differential coverage binning approaches. The minimum length threshold is set to 1000 bps.

	<i>Niastella</i> sp.		<i>Opitutus</i> sp.	
	MaxBin	Diff. Cov.	MaxBin	Diff. Cov.
Total length	4340643	4174821	4972625	4249012
Scaffold count	134	91	488	281
Mean length (bps)	32639.3	45877.2	10210.7	15121
Maximum length (bps)	265826	265826	74880	74880
%GC	48	48	64	64
Total marker gene count	105	106	104	100
Unique marker gene count	102	105	90	99

Table S6. Numbers of genes for each clusters of orthologous groups (COG) category for the three *Sorangium* species.

	<i>Sorangium</i>	<i>S. Cellulosum</i>	<i>S. Cellulosum</i>
	sp.	So0157-2	So Ce 56
A - RNA processing and modification	1	4	4
B - Chromatin Structure and dynamics	0	2	2
C - Energy production and conversion	149	372	383
D - Cell cycle control, cell division, chromosome partitioning	29	52	50
E - Amino acid transport and metabolism	189	455	412
F - Nucleotide transport and metabolism	56	103	98
G - Carbohydrate transport and metabolism	158	432	379
H - Coenzyme transport and metabolism	139	276	259
I - Lipid transport and metabolism	95	224	232
J - Translation, ribosomal structure and biogenesis	173	212	214
K - Transcription	201	682	588
L - Replication, recombination and repair	135	264	281
M - Cell wall/membrane/envelop biogenesis	227	372	353
N - Cell motility	113	117	124
O - Posttranslational modification, protein turnover, chaperones	149	305	288
P - Inorganic ion transport and metabolism	143	310	307

Q - Secondary metabolites biosynthesis, transport and catabolism	56	247	220
R - General function prediction only	371	1128	1062
S - Function unknown	301	866	850
T - Signal transduction	266	802	852
U - Intracellular trafficking, secretion, and vesicular transport	152	163	180
V - Defense mechanism	47	134	105
Z - Cytoskeleton	5	13	12
<hr/> Total	<hr/> 3155	<hr/> 7535	<hr/> 7255
<hr/>			

Table S7. Numbers of extracted glycoside hydrolase (GH) and auxiliary activity (AA) genes in the three *Sorangium* species.

<i>Sorangium</i> sp.		<i>Sorangium cellulosum</i>	<i>Sorangium cellulosum</i>
		So0157	So Ce 56
GH1	12	7	5
GH2	3	2	1
GH3	10	6	4
GH5	21	11	10
GH6	1	1	
GH8	6	3	3
GH9	4	2	2
GH10	11	5	6
GH11	8	5	3
GH12	2	1	1
GH13	20	10	10
GH15	9	3	6
GH16	8	4	4
GH18	4	1	3
GH20	1	1	
GH23	11	6	5
GH25	1		1
GH26	3	1	2
GH28	3		3

GH30	4	2	2
GH31	6	4	2
GH32	3	2	1
GH33	6	2	4
GH35	2	1	1
GH39	2	1	1
GH42	2	1	1
GH43	31	18	13
GH44	3	2	1
GH52	1	1	
GH53	2	1	1
GH55	1	1	
GH57	2	1	1
GH62	7	4	3
GH63	2	1	1
GH65	2	1	1
GH67	2	1	1
GH73	1	1	
GH74	2		2
GH76	2	1	1
GH77	3	1	2
GH78	5	2	3
GH81	2	1	1
GH94	2	1	1

GH97	1	1	
GH105	2	1	1
GH109	13	7	6
GH113	3	2	1
GH114	3	1	2
GH115	3	2	1
GH116	1		1
GH117	1		1
GH120	1	1	
GH127	1	1	
GH128	1		1
GH131	1	1	
AA3	12	6	6
AA4	4	2	2
AA5	2	1	1
AA6	8	4	4
AA7	8	4	4

Table S8. Running time for all tested datasets.

Dataset	Total sizes of scaffolds (Mbps)	Assembly quality (N50)	Running time (hour:min:sec) ¹
10 genome simulation (80X)	39.06 Mbps	4574	00:11:58
10 genome simulation (20X)	18.51 Mbps	691	00:02:50
100 genome simulation (simLC+)	326.71 Mbps	1293	03:10:11
100 genome simulation (simMC+)	288.77 Mbps	383	00:19:17
100 genome simulation (simHC+)	319.26 Mbps	17169	02:39:15
HMP – SRS013705	157.46 Mbps	1078	01:03:22
HMP – SRS014477	39.47 Mbps	694	00:05:06
HMP – SRS018656	70.82 Mbps	1990	00:13:25
37A	109.52 Mbps	2907	00:33:25
37B	138.10 Mbps	1994	00:45:09

¹The running time calculation assumes that the scaffold coverage information is readily available. It does not include the amount of time to map reads against scaffolds to obtain scaffold coverage information using Bowtie2.