Table S1. Size and repeat structure of the *E. coli* DH1 genome sequence using *k*-mers of different size

	Total sequence (bp)							
Copy number	***k* = 15**	***k* = 17**	***k* = 19**	***k* = 21**	***k* = 23**	***k* = 25**	***k* = 27**	***k* = 29**
1x	4,343,067	4,472,277	4,489,279	4,494,886	4,498,585	4,501,634	4,504,264	4,506,621
2x	162,178	45,060	31,810	29,196	28,098	27,300	26,674	26,144
3x	28,908	21,969	21,351	20,877	20,412	19,989	19,611	19,248
4x	10,836	9,164	8,620	8,288	7,892	7,512	7,124	6,784
5x	10,080	9,770	9,585	9,370	9,250	9,130	9,080	9,025
6x	8,286	8,268	8,418	8,490	8,508	8,616	8,664	8,754
7x	37,016	36,918	36,449	35,924	35,427	34,944	34,475	33,999
8x	3,368	2,408	1,944	1,704	1,456	1,240	1,056	904
9x	297	189	234	207	198	153	135	126
10x	390	330	350	260	200	180	170	150
11x – 20x	20,469	19,820	19,152	18,757	18,535	18,372	18,224	18,078
21x – 70x	5,798	4,518	3,497	2,728	2,124	1,613	1,204	846
Total[a]	4,630,693	4,630,691	4,630,689	4,630,687	4,630,685	4,630,683	4,630,681	4,630,679

[a] The size of the sequenced *E. coli* DH1 genome is 4,630,707 bp, $k - 1$ greater than the total *k*-mers.

Table S2. Genome sizes estimated from read sets of *E. coli* strains using *k*-mers of different size

	Estimated genome size (bp)							
***E. coli* strain**	***k* = 15**	***k* = 17**	***k* = 19**	***k* = 21**	***k* = 23**	***k* = 25**	***k* = 27**	***k* = 29**
A_03_34	4,778,825	4,776,822	4,776,195	4,775,705	4,775,453	4,774,872	4,774,274	4,773,837
B_04_28	4,927,099	4,932,857	4,934,298	4,935,111	4,933,859	4,935,175	4,934,608	4,936,041
C_04_22	5,168,076	5,175,804	5,177,972	5,180,594	5,182,109	5,183,615	5,184,004	5,185,904
D_04_27	5,216,262	5,218,504	5,219,054	5,219,387	5,219,365	5,219,058	5,216,893	5,216,055
E_01_37	5,504,417	5,503,399	5,502,931	5,502,511	5,502,004	5,501,319	5,500,696	5,500,121

Table S3. I-*Ceu*I fragment lengths for *E. coli* strains

Fragment lengths (bp) estimated by PFGE						Fragment lengths (bp) based on genome sequence
Strain A_03_34	**Strain B_04_28**	**Strain C_04_22**	**Strain D_04_27**	**Strain E_01_37**	***E. coli* MG1655**	***E. coli* MG1655**
40,036	41,747	40,720	41,062	37,299	40,378	41,398
115,992	94,779	130,161	120,097	117,873	93,239	93,812
139,086	136,862	120,782	179,801	138,060	133,099	131,117
525,392	540,151	530,411	538,598	547,142	528,111	520,769
680,359	645,404	860,359	711,042	708,712	663,270	657,364
683,855	720,752	733,957	831,831	812,800	702,109	697,595
2,684,174	2,866,877	2,939,489	3,000,390	2,905,916	2,827,057	2,497,592

Table S4. Estimates of microbial genome sizes based on *k*-mer analysis of short read datasets

NCBI SRA[a] run number	Genome source	Genome size[b] (kb)	Number of replicons	Genome size estimate (kb)	Coverage estimate	Unique estimate (kb)
SRR059788 SRR059789	*Niastella koreensis* GR20-10, DSM 17620	9,033.7	1	8,392.0	53.4	8,287.2
SRR072318 SRR090709	*Burkholderia* sp. CCGE1002	7,884.9	4	8,014.3	24.3	7,676.8
SRR031266 SRR031261 SRR031262 SRR031263 SRR031264 SRR031265 SRR031266	*Burkholderia* sp. CCGE1002	7,884.9	4	7,627.3	104.6	7,410.5
SRR610299	*Cylindrospermum stagnale* PCC 7417	7,610.6	4	7,617.2	510.7	7,064.6
SRR071425	*Mycobacterium smegmatis* MC2 155	6,988.2	1	7,110.1	251.7	6,626.3
SRR610309	*Nostoc* sp. PCC 7524	6,718.9	3	6,726.8	177.2	6,162.0
SRR090599	*Planctomyces brasiliensis* IFAM 1448, DSM 5305	6,008.0	1	5,988.3	393.6	5,841.2
SRR059232 SRR059233 SRR059234 SRR059235 SRR059236	*Escherichia coli* KO11FL	5,029.3	2	5,203.4	36.8	4,710.7
SRR190843	*Owenweeksia hongkongensis* DSM 17368	4,000.0	1	3,997.3	1,065	3,925.0
SRR006332	*Acinetobacter baylyi* ADP1	3,598.6	1	3,546.2	52.6	3,480.8
SRR006330	*Acinetobacter baylyi* ADP1	3,598.6	1	3,480.5	22.6	3,373.9
SRR396647	*Listeria monocytogenes* J0161, FSL R2-499	3,000.4	1	3,353.8	33.8	2,911.6

SRR396649 SRR396650 SRR396651 SRR396653						
SRR089543 SRR089544	*Rothia dentocariosa* ATCC 17931	2,506.0	1	2,602.1	37.8	2,473.1
SRR060959 SRR060960	*Thermovirga lienii* DSM 17291	1,999.6	2	2,052.7	108.1	1,875.8
SRR006331	*Mycoplasma agalactiae* PG2	877.4	1	872.8	21.6	855.5
SRR387449	phiX174	5.4	1	5.3	171,200.0	5.284.0
SRR769601	*Escherichia coli* strain A_03_34[c]	4,908.0		4,759.9	54.6	4,642.5
SRR769603	*Escherichia coli* strain B_04_28[c]	4,980.0		4,908.5	68.5	4,821.3
SRR769600	*Escherichia coli* strain C_04_22[c]	5,039.0		5,145.4	80.3	4,585.4
SRR769602	*Escherichia coli* strain D_04_27[c]	5,278.0		5,209.8	60.5	5,004.6
SRR769599	*Escherichia coli* strain E_01_37[c]	5,196.0		5,445.7	84.5	4,801.5

[a] NCBI SRA: National Center for Biotechnology Information Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra)

[b] Genome size obtained by complete sequencing or by PFGE measurement.

[c] Novel *E. coli* isolate, reported in this study.

Table S5. Effect of quality-score-based trimming on genome size estimates

	Combined length of reads (bp)			Genome size estimate (bp)		
***E. coli* strain**	**Raw reads**	**Trimmed reads[a]**	**Percentage removed by trimming**	**Raw reads**	**Trimmed reads[a]**	**Difference in genome size estimates**
A_03_34	362,127,992	344,702,582	4.81	4,775,814	4,766,569	-0.19%
B_04_28	468,434,968	446,724,279	4.63	4,935,386	4,927,683	-0.16%
C_04_22	580,398,016	554,098,452	4.53	5,180,545	5,176,786	-0.07%
D_04_27	437,355,832	417,108,122	4.63	5,219,319	5,210,466	-0.17%
E_01_37	645,818,056	614,475,488	4.85	5,502,494	5,492,264	-0.19%

[a] Sequence reads were trimmed with *Dynamic Trim* of the **SolexaQA** vers. 2.2 package. This procedure retains the longest contiguous region for which the probability of an incorrect base call remains below 10%. Estimates of incorrect base calls are based on the quality scores assigned to each read position during sequencing.

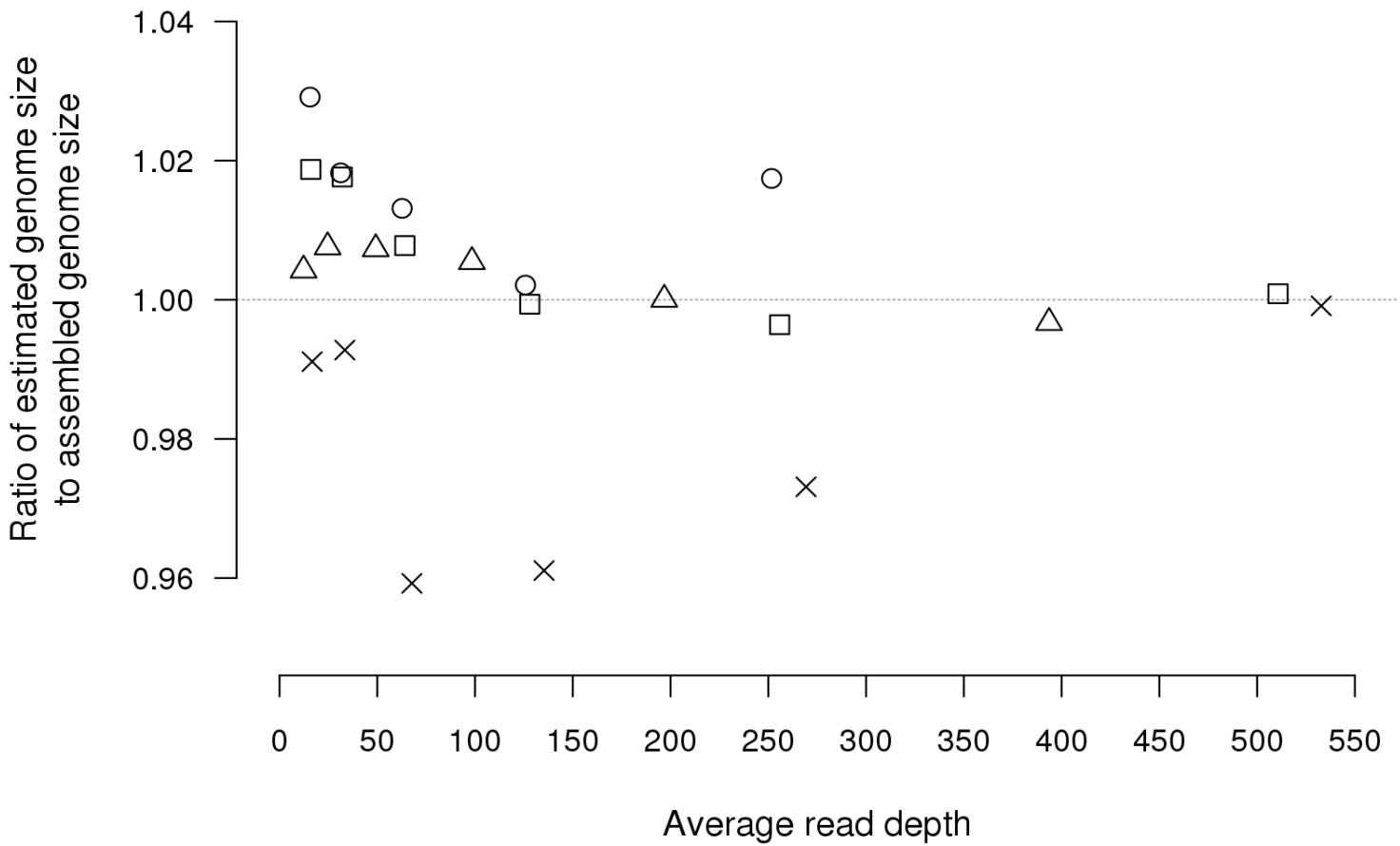

Figure S1. Ratio of assembled to estimated genome sizes at different read depths. Sample sets of different read depths were obtained by randomly selecting subsamples of short-read datasets at 0.5, 0.25, 0.125, 0.0625 and 0.03125 of their original sizes. Original dataset sizes were obtained from assembled sequences, and average read depths were estimated by 21-mer frequency analysis and correspond to the principle peak in a 21-mer spectrum. Symbols denote the organism from which the whole-genome shotgun sequence data were obtained: ○ *Cylindrospermum stagnale* PCC 7417 (SRA run number: SRR610299); □ *Mycobacterium smegmatis* MC2 155 (SRA run number: SRR071425); × *Owenweeksia hongkongensis* DSM 17368 [1] (SRA run number: SRR190843); □ *Planctomyces brasiliensis* IFAM 1448 DSM 5305 [2] (SRA run number: SRR090599).

References

1. Riedel T, Held B, Nolan M, Lucas S, Lapidus A, Tice H, Del Rio TG, Cheng JF, Han C, Tapia R, Goodwin LA, Pitluck S, Liolios K, Mavromatis K, Pagani I, Ivanova N, Mikhailova N, Pati A, Chen A, Palaniappan K, Rohde M, Tindall BJ, Detter JC, Göker M, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Klenk HP *et al.*: **Genome sequence of the orange-pigmented seawater bacterium Owenweeksia hongkongensis type strain (UST20020801(T))**. *Stand Genomic Sci* 2012, **7**:120-130.

2. Wu DY, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, Kunin V, Goodwin L, Wu M, Tindall BJ, Hooper SD, Pati A, Lykidis A, Spring S, Anderson IJ, D'haeseleer P, Zemla A, Singer M, Lapidus A, Nolan M, Copeland A, Han C, Chen F, Cheng JF, Lucas S, Kerfeld C, Lang E, Gronow S, Chain P, Bruce D *et al.*: **A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea.** *Nature,* 2009, **462**:1056-1060.