***De novo* assembly of bacterial genomes from single cells**

Supplementary Methods

Velvet-SC: Modifications to Velvet assembly algorithm

EULER+Velvet-SC is EULER-SR's error correction [1] combined with a modification of Velvet aimed at fragment assembly for single cell short reads with highly non-uniform coverage. Velvet is an efficient yet simple and flexible implementation of a de Bruijn graph based assembler (overview in Supplementary Fig. S5; see [2] for details on the algorithm). Its flexibility allows easy modifications for specialized applications like single cell sequencing. Note that in line 8 (Supplementary Fig. S5), Velvet removes those edges that have low average coverage, a critical step that simplifies the graph by attempting to remove erroneous edges while preserving correct ones. The cutoff threshold is either automatically determined by Velvet or set as a command line option by the user. Because single cell read data has a significant number of regions with low average coverage, Velvet either eliminates a significant portion of the assembly or leaves erroneous edges, which leads to a deterioration in assembly quality.

To overcome this problem, Velvet was modified by incorporating an incremental scheme to eliminate low-coverage branches. Velvet-SC (Supplementary Fig. S5) iterates over lines 8–12 of Velvet, varying the coverage cutoff instead of using a fixed cutoff. We illustrate the effect of this iterative scheme with two examples:

Treatment of a low coverage single nucleotide error: See Fig. 1, and a continuation of it, Supplementary Fig. S3, which represents the same scenario using a *condensed de Bruijn graph* (in which nonbranching chains of vertices are merged together into a longer sequence). In this example, coverage varies between 1x-12x. There are two potential contigs to choose from in the middle, differing by a single nucleotide (C vs. T): a green contig with coverage 6.4, and a blue contig with coverage 1. With a fixed coverage threshold of 4, Velvet would delete the low coverage blue and purple contigs (at Supplementary Fig. S5a, line 8), and then merge the high coverage red and green contigs into a contig much shorter than the full genome. Velvet-SC instead starts by eliminating sequences of average coverage below 2, which only removes the blue contig (Supplementary Fig. S5b, line 9, at iteration i=2). The other contigs are combined into a single contig of average coverage 9 (illustrated in Fig. 1b, Supplementary Fig. S3b). The purple region is salvaged by Velvet-SC because it was absorbed into a higher coverage region faster than the variable coverage threshold increased; this contig will remain in the assembly as long as its average coverage is above the cutoff. Velvet-SC repeats this process with a gradually increasing low coverage threshold.

Treatment of a chimeric junction: A chimeric read may form a short, low-coverage bridge between two contigs as illustrated in Supplementary Fig. S4. As long as the support of this bridge is low coverage, it will be eliminated by both Velvet and Velvet-SC; there is no difference in its treatment.

EULER+Velvet-SC computations were done on a PC with an Intel Core i7 processor and 24 GB RAM. For each lane, EULER-SR error correction took about 30 min and Velvet-SC assembly took about 30 min.

Micromanipulation of single cells

Single cells of *Escherichia coli* (ATCC 700926) and *Staphylococcus aureus* MRSA USA300 strain FPR3757 [3] (ATCC 25923) were grown in Luria broth at 37°C. A 1 ml aliquot of log phase culture was washed 3 times with 1 ml sterile 1x phosphate buffered saline (PBS), then diluted in 1x PBS for micromanipulation. Single cells were micromanipulated as described previously [4, 5] using an Olympus IX70 inverted microscope equipped with a TransferMan NK2 and CellTram Vario (Eppendorf) and sterile glass capillaries. Single cells were rinsed in sterile TE buffer, placed in 1 µl of buffer [TE for *E. coli*; TN (50 mM TrisHCl pH 7.5, 145 mM NaCl) for *S. aureus*] in a 0.2 ml PCR tube, and stored on ice or at -20°C.

Multiple Displacement Amplification (MDA)

***E. coli* and *S. aureus* MDAs.** Cells were brought up in a final volume of 4 µl with TE buffer (TN for *S. aureus*), and lysed by addition of 1 µl of alkaline lysis solution (1075 mM KOH, 265 mM DTT, 26.5 mM EDTA pH 8.0) in a 10 min incubation on ice. After neutralization with 1 µl of 2150 mM TrisCl pH 4.5, 19 µl of GenomiPhi master mix (GE Healthcare) was added (11.25 µl of GenomiPhi Reaction Buffer, 6.5 µl GenomiPhi Sample Buffer, and 1.25 µl GenomiPhi Enzyme Mix) for a reaction volume of 25 µl. Reactions were incubated at 30°C for 4 h followed by a 10 min inactivation step at 65°C. For *S. aureus*, the initial 4 µl volume was reduced to 3 µl, and 1 µl of lysostaphin (Cell Sciences, Canton MA, 20 ng/µl solution in TN) was added, followed by a 1 h incubation at 37°C. Alkaline lysis and MDA was then performed as above. Lysostaphin was essential for successful MDA with *S. aureus* cells (data not shown). *S. aureus* MDAs were purified by standard phenol:chloroform:isoamyl alcohol extraction. Quantification of MDA yield was performed by Picogreen assay as per manufacturer's protocol in the GenomiPhi HY kit. *E. coli* and *S. aureus* MDAs with better representation of the genome were identified for sequencing by qPCR as described previously [6], and further amplification of selected MDAs for sequencing was performed as described above with 150–1500 ng of the original MDA as template.

Marine cell MDAs. MDA was performed with the GenomiPhi HY kit as described except that after lysis, a 7.5 µl mixture of 1 µl of 2150 mM TrisCl pH 4.5 and 6.5 µl GenomiPhi Sample Buffer was added, followed by a 12.5 µl mixture of GenomiPhi Reaction Buffer (11.25 µl) and GenomiPhi Enzyme Mix (1.25 µl) to make up the 25 µl reaction.

16S rRNA PCR and sequencing

16S rRNA gene (~1500 base pairs) was amplified from diluted MDA product using universal bacterial primers 27f and 1492r [7] as follows: 94°C for 3 min, 35 cycles of 94°C for 30 sec, 55°C for 30 sec, 72°C for 90 sec, and 72°C for 10 min. PCR products were treated with Exo I and shrimp alkaline phosphatase (both from Fermentas) prior to direct cycle sequencing at the Joint Technology Center (J. Craig Venter Institute, Rockville, MD). 16S tracefiles were analyzed and trimmed with the CLC Workbench software program (CLC bio, Muehltal, Germany), and taxonomy was determined using BLAST and the Ribosomal Database Project (RDP) Classifier tool [8].

APIS (Automated Phylogenetic Inference System)

APIS (Badger et al, unpublished) is a system for the automatic building and interpretation of phylogenetic trees from a set of protein or nucleotide sequences. Protein-coding regions are compared to a database of proteins from complete genomes using BLAST, and full length sequences aligned to each query using MUSCLE [9]. A bootstrapped neighbor-joining tree is inferred for each alignment and the phylogenetic position of the query is determined from the tree, a more biologicially meaningful method than phylogeny based on the top BLAST hit [10]. COG cluster identities [11] were inferred from the closest phylogenetic neighbor on the tree. If that organism was one used to generate the COG clusters, the identity of the matching cluster could be determined directly, otherwise the closest BLAST match of the neighbor to the COG database was used to determine the cluster.

Supplementary Data 3

Mapped reads, co-mapped reads, and chimeric fragments

We used Bowtie version 0.12.1 [12] with default parameters in single-end mode to map the reads to all possible locations in the reference genome. A read is called *mapped* if Bowtie outputs an alignment for it and *unmapped* otherwise. An unmapped read is called *chimeric* if it does not map to the genome but instead consists of a prefix that maps to one region of the genome and a suffix that maps to another region of the genome. These correspond to the chimeric DNA rearrangements that occur during the MDA reaction, in which a primer is partially extended on an initial template and then the 3' end is displaced and re-anneals to a second template [13]

Supplementary Table S1 presents the number of reads passing the Illumina purity filter and the number of mapped reads and co-mapped read pairs (defined below). In normal multicell data, 99% of the reads passing the purity filter map to the genome, whereas single cell data about 93% do. Illumina sequencing produces many duplicate reads; Supplementary Table S2 presents the numbers of distinct, unique, and non-unique reads and read pairs that pass the Illumina purity filter.

In the Illumina paired-end sequencing protocol, a read pair consists of two reads on opposite strands a certain distance apart. The reads should be oriented towards each other. The nominal insert length is only approximate, and the actual length varies between different read pairs. We call a read pair *co-mapped* if there is a mapping of both reads to the reference genome in the correct orientation, less than 3 times the nominal insert length apart. We call it an *abnormal* read pair otherwise (both reads map but with incorrect orientation or distance; or one read maps and the other does not). We classify deviations from these conditions as follows:

- **Anomalous pair.** A read pair in which the two reads map in the correct orientation to two positions separated by more than 3 times the nominal insert length.
- **Forward+forward/Reverse+reverse pair (FF/RR pair).** A read pair in which both reads map to the same strand. In a co-mapped pair, the reads map to opposite strands.
- **Outies.** A read pair in which the reads are oriented away from one another. (This would be the correct configuration for the Illumina "mate pair" protocol, but it is an incorrect configuration for the Illumina "paired end" protocol.)
- **Singleton+shadow pair (S pair).** A read pair in which one of the two reads (shadow) does not map to the reference genome.

An abnormal read pair is classified in the above order. Supplementary Table S3 summarizes the number and percentage of the first three types (in which both reads map); singleton+shadow pairs are shown separately in Supplementary Table S1. The rates of abnormal read pairs in single cell data is significantly higher than those in normal multicell data.

A chimeric fragment is defined from the relative orientation of aligned pairs of reads from the ends of a fragment. In the absence of chimeras, a pair of reads generated by sequencing a fragment of DNA will generate a pair of reads that map to the reference genome in a head to head orientation, i.e., 5'-3' direction on the top strand and 5'-3' direction on the bottom strand. If during an MDA reaction a synapse is formed between two non-contiguous regions of the genome, a chimera will form. In the subsequent library preparation, shearing and size selection of MDA products will generate ~250 base fragments comprising sequences from distant loci of

the reference genome sequence. The chimeric junction may fall between the two reads, or it may fall within one of the two reads. If it falls within one of the two reads, the read is chimeric; if the other read maps, it may be classified as a singleton+shadow pair, but there may be other sources of singleton+shadow pairs, such as contaminants. Altogether, anomalous pairs, FF/RR pairs, outies, and some but not all S pairs, are chimeric fragments. In the single cell *E. coli* lanes, we estimate one chimeric junction every 4-13 kb (depending on how many S pairs are chimeric fragments), which is elevated from one per 22 kb observed previously [13]. This discrepancy may reflect inaccuracies in chimera detection, as current tools for short reads are poor at detecting chimeric reads.

From the point of view of the chimera formation mechanism in MDA, abnormal read pairs fall into two categories: (1) inverted, and (2) direct rearrangements; see [13] for a detailed definition. In our terminology, inverted rearrangement is synonymous with forward+forward/reverse+reverse pairs, while direct rearrangement comprises outies and anomalous pairs.

Effect of template concentration on MDA chimera frequency

200 ng, 100 ng, 10 ng, 1 ng, 100 pg, 10 pg and 1 g of *E. coli* MG1655 DNA was amplified in 7 separate reactions using the REPLI-g Mini Kit (Qiagen). Each of the 7 amplification reactions resulted in between 37 μg and 46 μg of amplified material as quantified by a Nanodrop spectrophotometer. Short insert paired end libraries were generated from the 7 amplified DNA samples following the standard Illumina protocol [14] using 5 μg of input DNA, with the gel size selecting an average insert size of ~250 bp. A 250 bp insert paired end library was also generated from 5 μg of unamplified *E. coli* MG1655 DNA as a control. Clusters were generated from the resulting 8 libraries and sequenced on a GAIIx in a paired 35 cycle experiment. Reads were aligned to the *E. coli* MG1655 reference genome using ELAND [15].

Supplementary Table S5 records the relationship between the frequency of identifying chimeric fragments in paired-read sequencing and the quantity of DNA input to the MDA reaction. In the control sample, 734 FF/RR fragments were detected by paired end sequencing, which represents 0.02% of the total number of fragments. This background level of chimerism in the absence of MDA reflects an artifact of the ligation reaction to add sequencing adapters to fragments wherein two or more fragment ligate to form concatemers. This type of chimera can be resolved by increasing depth of coverage. The percentage of FF/RR is significantly elevated in the MDA products and increases from 0.67% for a 200 ng input to 3.04% for a 1 pg input. As the FF/RR percentage increases, the ratio of directed to inverted MDA fragments decreases. Note that the percentages in this table are not directly comparable with those in Supplementary Table S3 because the reads there are 100 bp instead of 35 bp, so junctions between 36 bp and 100 bp from either end result may render one of the reads unmappable and result in a different error classification.

Supplementary Table S1. The number of reads that pass the Illumina purity filter; the number of mapped reads; and the number of fragments where both reads (or just one read) map, in the *E. coli* and *S. aureus* datasets. All lanes (1-4, 6-8, and normal) are *E. coli* except the one marked *S. aureus.* Percentages are out of reads that pass the Illumina purity filter.

Dataset	**Reads**	**Mapped reads**	**Fragments with both reads mapped**	**Fragments with one read mapped (S pairs)**
lane 1	29124078	27064409 (92.93%)	13879571 (95.31%)	694733 (4.77%)
lane 6	27573794	25510891 (92.52%)	13067928 (94.79%)	624965 (4.53%)
normal	28428648	28157555 (99.05%)	14204847 (99.93%)	252139 (1.77%)
S. aureus	66845058	64307038 (96.20%)	31849494 (95.29%)	1397726 (4.18%)
Replicates of *E. coli* lane 1				
lane 2	31885042	29773731 (93.38%)	15239293 (95.59%)	704855 (4.42%)
lane 3	32743056	20638018 (63.03%)	15693271 (95.86%)	748524 (4.57%)
lane 4	32323444	30202564 (93.44%)	15466756 (95.70%)	730948 (4.52%)
Replicates of *E. coli* lane 6				
lane 7	26695478	24673652 (92.43%)	12642138 (94.71%)	610624 (4.57%)
lane 8	24631296	22767781 (92.43%)	11677221 (94.82%)	586661 (4.76%)

Supplementary Table S2. The number of distinct, unique, and non-unique reads and pairs (representative of fragments) that pass the Illumina purity filter in the *E. coli* and *S. aureus* datasets. "Distinct" counts each distinct read or pair once, no matter how many times it's repeated in the data. "Unique" counts reads or read pairs that occur exactly once in the data. "Non-unique" counts reads or read pairs that occur multiple times in the data, including their multiplicity. All lanes (1-4, 6-8, and normal) are *E. coli* except the one marked *S. aureus.* Percentages are out of reads that pass the Illumina purity filter.

Dataset	**Distinct Reads**	**Unique Reads**	**Non-unique Reads**	**Distinct Pairs**	**Unique Pairs**	**Non-unique Pairs**
lane 1	17198215 (59.05%)	13992267 (48.04%)	15131811 (51.96%)	14562036 (100.00%)	14562033 (100.00%)	6 (0.00%)
lane 6	14937838 (54.17%)	11713716 (42.48%)	15860078 (57.52%)	13786892 (100.00%)	13786887 (100.00%)	10 (0.00%)
normal	18887690 (66.4%)	14333028 (50.42%)	14095620 (49.58%)	14144998 (99.51%)	14076858 (99.03%)	137465 (0.97%)
S. aureus	11218204 (16.78%)	6554774 (9.81%)	60290284 (90.19%)	29860912 (89.34%)	27253986 (81.54%)	6168543 (18.46%)
Replicates of *E. coli* lane 1						
lane 2	17625800 (55.28%)	14182796 (44.48%)	17702246 (55.52%)	15942520 (100.00%)	15942519 (100.00%)	2 (0.00%)
lane 3	18171243 (55.50%)	14702209 (44.90%)	18040847 (55.10%)	16371523 (100.00%)	16371518 (100.00%)	10 (0.00%)
lane 4	17847840 (55.22%)	14383188 (44.50%)	17940255 (55.50%)	16161719 (100.00%)	16161717 (100.00%)	4 (0.00%)
Replicates of *E. coli* lane 6						
lane 7	14351161 (53.76%)	11167250 (41.83%)	15528228 (58.17%)	13347735 (100.00%)	13347731 (100.00%)	8 (0.00%)
lane 8	14260615 (57.90%)	11392974 (46.25%)	13238322 (53.75%)	12315647 (100.00%)	12315646 (100.00%)	2 (0.00%)

Supplementary Table S3. The number and percentage of forward+forward/reverse+reverse pairs, outies, anomalous pairs, and subtotal of these; co-mapped pairs (correct size and orientation); and the ratio of direct to inverted rearrangements in the *E. coli* and *S. aureus* datasets. All lanes are *E. coli* except the one marked *S. aureus*. Percentages are out of the number of fragments with both reads mapped (in Supplementary Table S1).

Dataset	**FF/RR**	**Outies**	**Anomalous**	**Subtotal**	**Co-mapped**	**Dir/Inv ratio**
lane 1	152434 (1.16%)	50493 (0.38%)	59692 (0.45%)	262619 (1.99%)	12922219 (98.01%)	0.72
lane 6	164504 (1.32%)	48337 (0.39%)	63932 (0.51%)	276773 (2.22%)	12166190 (97.78%)	0.68
normal	732 (0.01%)	328 (0%)	744 (0.01%)	1804 (0.01%)	13950904 (99.99%)	1.46
S. aureus	109598 (0.36%)	22365 (0.07%)	26991 (0.09%)	158954 (0.52%)	30292814 (99.48%)	0.45
Replicates of *E. coli* lane 1						
lane 2	175122 (1.2%)	58437 (0.4%)	69836 (0.48%)	303395 (2.09%)	14231043 (97.91%)	0.73
lane 3	183342 (1.23%)	61410 (0.41%)	72419 (0.48%)	317171 (2.12%)	14627576 (97.88%)	0.73
lane 4	180423 (1.22%)	59873 (0.41%)	70892 (0.48%)	311188 (2.11%)	14424620 (97.89%)	0.72
Replicates of *E. coli* lane 6						
lane 7	157648 (1.31%)	46675 (0.39%)	61563 (0.51%)	265886 (2.21%)	11765628 (97.79%)	0.69
lane 8	147962 (1.33%)	44221 (0.4%)	57984 (0.52%)	250167 (2.26%)	10840393 (97.74%)	0.69

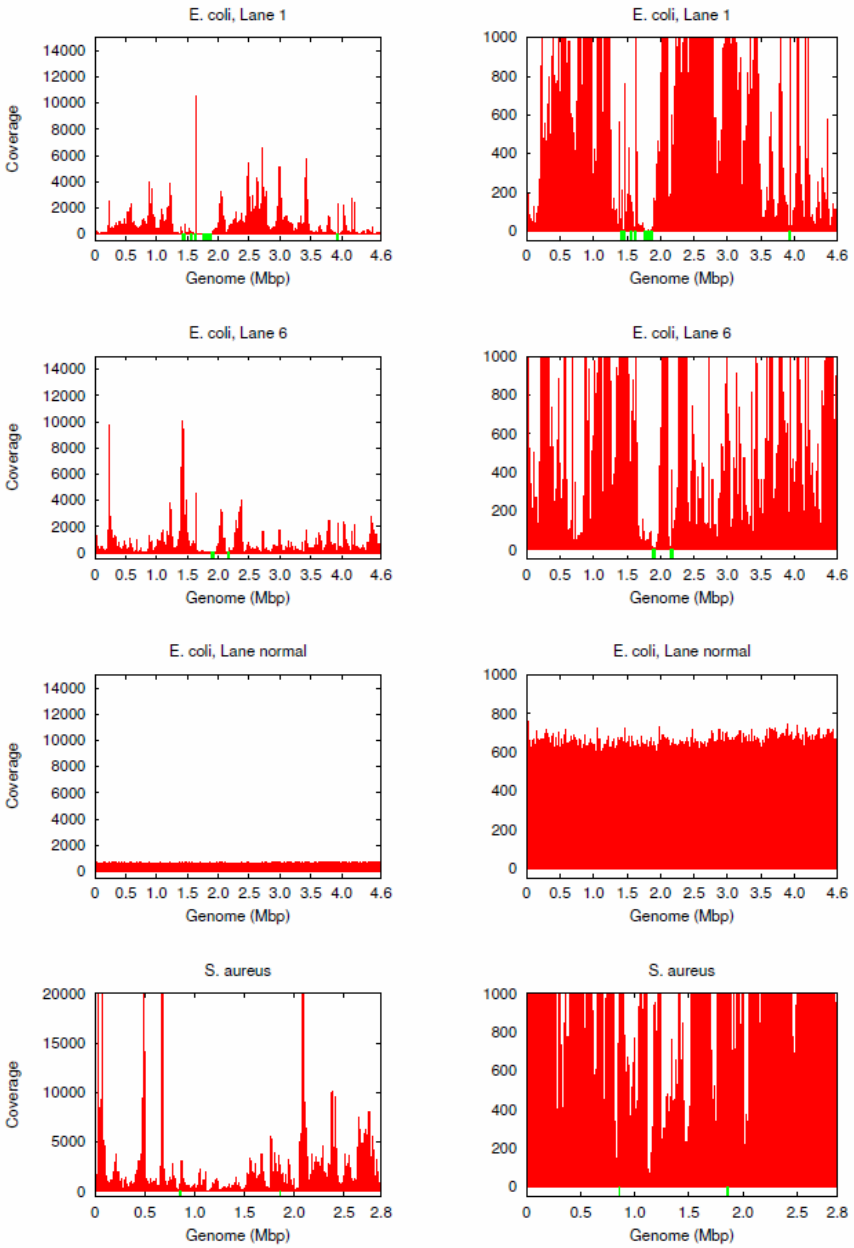
Supplementary Table S4. The number and percentage of bases with low coverage (from 0 to 5) and the number of blackout regions in the *E. coli* and *S. aureus* datasets. The reads were mapped using Bowtie with default parameters that allow a maximum of 2 mismatches in the first 28 bases of the read and a maximum of 70 for the sum of quality values at all mismatched read positions throughout the entire alignment. All lanes are *E. coli* except the one marked *S. aureus.*

Dataset	**Number of blackout regions**	**Mean length**	**Standard deviation**	**N50**	**Number (%) of bases with coverage**					
					0	1	2	3	4	5
lane 1	94	1220	3692.2	5558	84K (1.8%)	32K (0.7%)	19K (0.4%)	14K (0.3%)	15K (0.3%)	15K (0.3%)
lane 6	50	193	269.7	518	5K (0.1%)	8K (0.2%)	10K (0.2%)	10K (0.2%)	9K (0.2%)	9K (0.2%)
lanes 1 and 6	0	0	0	0	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	9 (0.0%)	43 (0.0%)
normal	0	0	0	0	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.0%)	1 (0.0%)
S. aureus	2	95	25.4	83	143 (0.0%)	4 (0.0%)	0 (0.0%)	4 (0.0%)	2 (0.0%)	3 (0.0%)
Replicates of *E. coli* lane 1:										
lane 2	91	1183	2937.8	4700	77K (1.7%)	32K (0.7%)	19K (0.4%)	14K (0.3%)	13K (0.3%)	14K (0.3%)
lane 3	92	1159	3359.5	5842	77K (1.7%)	30K (0.7%)	20K (0.4%)	14K (0.3%)	14K (0.3%)	14K (0.3%)
lane 4	88	1225	3049.9	6156	76K (0.0%)	33K (0.7%)	19K (0.4%)	13K (0.3%)	13K (0.3%)	14K (0.3%)
Replicates of *E. coli* lane 6:										
lane 7	63	153	250.5	456	5K (0.1%)	8K (0.2%)	10K (0.2%)	10K (0.2%)	9K (0.2%)	10K (0.2%)
lane 8	61	185	262.5	573	6K (0.1%)	9K (0.2%)	10K (0.2%)	10K (0.2%)	10K (0.2%)	11K (0.2%)

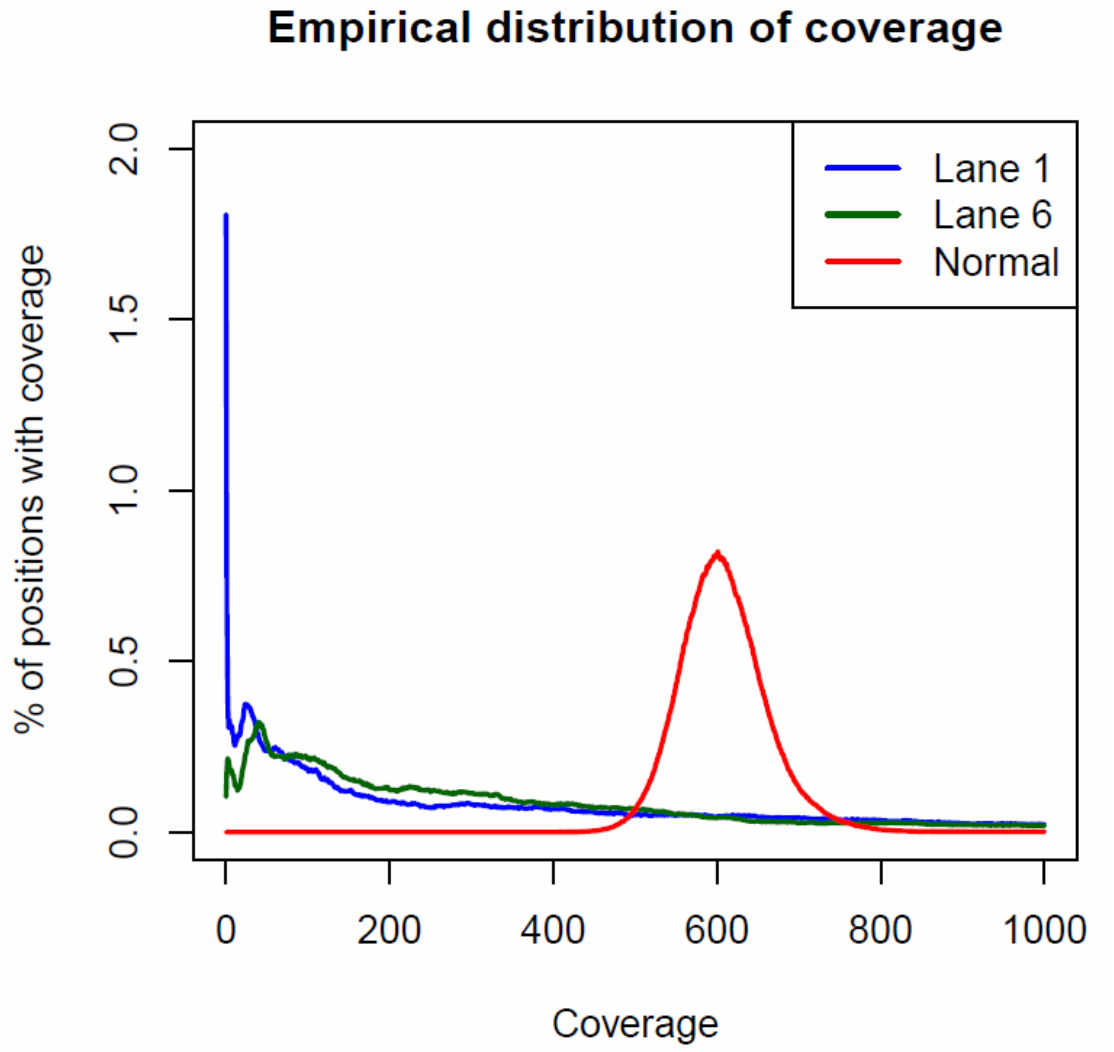
Supplementary Table S5. The number and percentage of various types of fragments for libraries that had different quantities of *E. coli* DNA input to the amplification ranging from 1 pg to 200 ng. The control sample CT1658 is a library generated from 5 μg of unamplified *E. coli* DNA. Counts are given for forward+forward/reverse+reverse pairs, outies, and correct orientation but wrong size pairs; the subtotal of these types of chimeric fragments; co-mapped pairs (correct size and orientation); and the ratio of direct to inverted rearrangements.

Library ID	Input to MDA	FF/RR	Outies	Wrong size	Subtotal	Co-mapped	Dir/Inv ratio
CT1658	control	734 (0.03%)	604 (0.02%)	3484 (0.13%)	4822 (0.18%)	2710643 (99.82%)	5.57
CT1655	200 ng	25127 (0.67%)	6192 (0.17%)	206725 (5.55%)	238044 (6.39%)	3486115 (93.61%)	8.47
CT1656	100 ng	22277 (0.71%)	3833 (0.12%)	184704 (5.86%)	210814 (6.69%)	2939627 (93.31%)	8.46
CT1657	10 ng	38426 (1.00%)	5100 (0.13%)	293430 (7.67%)	336956 (8.81%)	3489329 (91.19%)	7.77
CT1517	1 ng	52793 (1.47%)	4835 (0.13%)	245579 (6.82%)	303207 (8.42%)	3296039 (91.58%)	4.74
CT1659	100 pg	52421 (1.90%)	3436 (0.12%)	163335 (5.92%)	219192 (7.95%)	2538438 (92.05%)	3.18
CT1660	10 pg	72991 (2.42%)	4274 (0.14%)	129148 (4.29%)	206413 (6.86%)	2803741 (93.14%)	1.83
CT1661	1 pg	104947 (3.04%)	5806 (0.17%)	107389 (3.11%)	218142 (6.32%)	3234809 (93.68%)	1.08

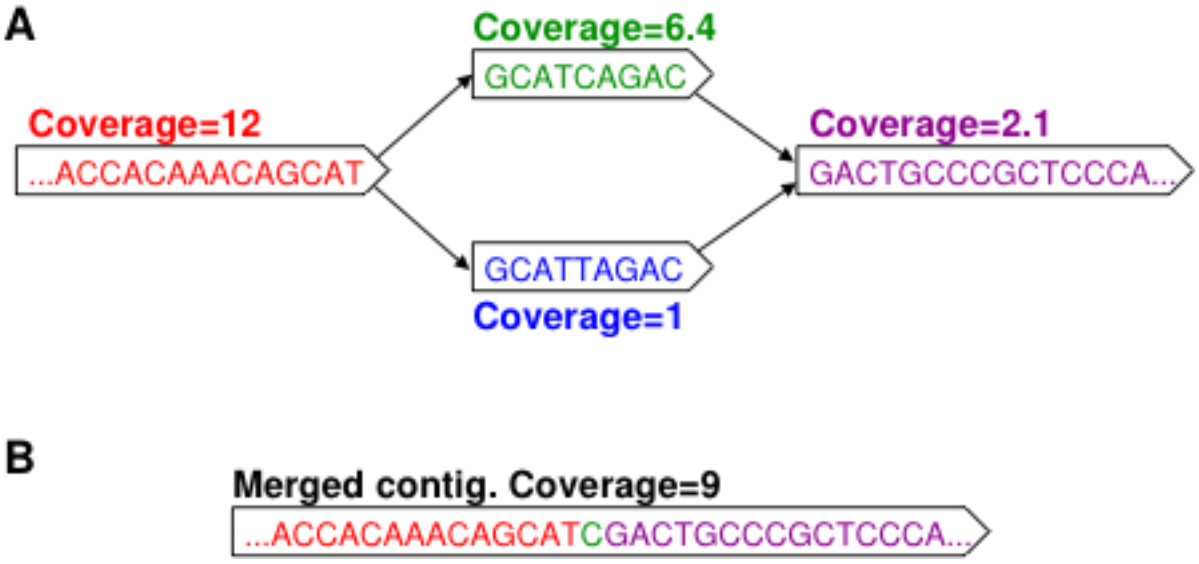
Supplementary Figure S1. Coverage per genome position in the *E. coli* datasets for lane 1 (top), lane 6 (middle), normal (multicell) lane (middle) and *S. aureus* dataset (bottom). The *y*-axis shows the number of reads that contain position *x* of the genome in red with the genome binned into 1000 bp windows. The green track is coverage at the base level, without binning, indicating regions with no coverage (obscured by binning into 1000 bp windows). Bowtie version 0.12.1 [12] was used with default parameters in single-end mode to map the reads to all possible locations in the reference genome. The *coverage* of a position in the genome is a *weighted* count of the mapped reads containing that position (rather than a simple count that would boost the coverage in repeat regions): a read that maps to *n* different locations in the genome contributes coverage 1/*n* to each nucleotide in each of the *n* locations. On the right, coverage plots restricted to coverage from 0 to 1000.



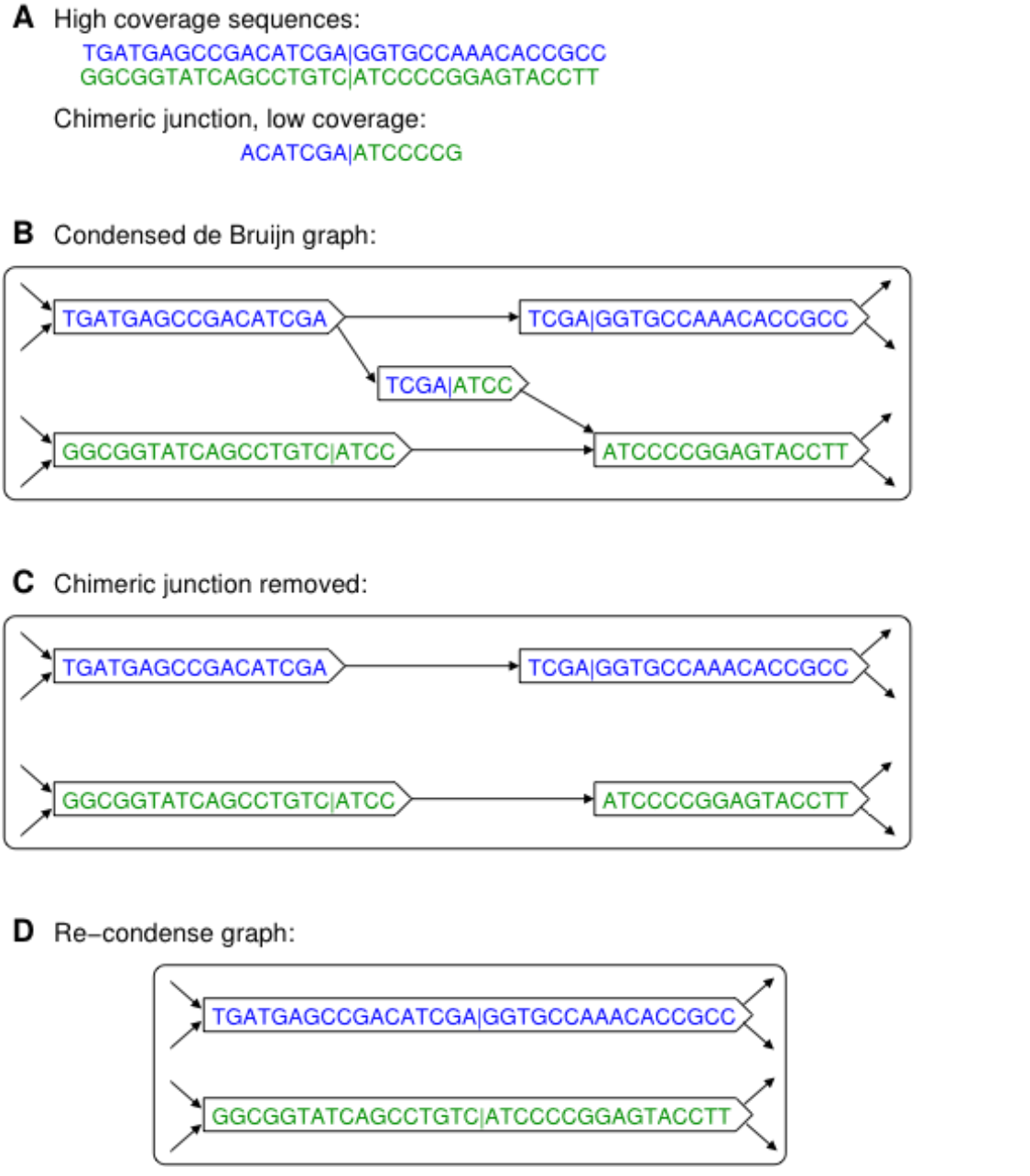
Supplementary Figure S2. Histogram showing fraction of positions having given coverage (using the read mapping data described in Supplementary Table S4), for *E. coli* lane 1 (blue), lane 6 (green), and normal multicell (red). The multicell sample has nearly uniform coverage along the genome: the coverage distribution is roughly a normal distribution with a peak near 600x coverage, and most positions have between 450-800x coverage. In the single cell samples, small coverage has very high probability, and the distribution decays as coverage increases. 1.8% of positions in lane 1, and 0.1% of positions in lane 6, have no coverage. A long thin tail on the x-axis has been clipped: there are positions with coverage well above 1000x, yielding average coverages above 600x in all three cases.



Supplementary Figure S3. Continuation of Figure 1 from the main text. For efficiency, the de Bruijn graph is *condensed* by merging "1-in 1-out" chains (a series of successive vertices with one edge in, one edge out) together. (a) In Figure 1c), the reads are broken into successive 5-mers (as the vertices), and successive vertices have an overlap of size *k*-1=4. Here we represent the same data in a condensed de Bruijn graph, in which the sequences at the vertices have variable lengths, and successive vertices still overlap by *k*-1=4. (b) After eliminating the blue contig, the red, green, and purple contigs form a 1-in 1-out chain of vertices, which we further condense into a single contig.



Supplementary Figure S4. Assembly with a chimeric read junction. (a) Two high coverage sequences are shown (top sequence in blue, bottom sequence in green). There is a chimeric read with part of the blue sequence and part of the green sequence. A bar "|" is shown in all sequences to indicate the junction. (b) Shown is the condensed de Bruijn graph for k=5. This yields a low coverage bridge vertex in the center, consisting of the last k-1=4 nucleotides from before the junction and the first k-1=4 nucleotides from after the junction. (c) If the bridge vertex has low support, such as one read, it will be eliminated whether we use a fixed cutoff (Velvet) or a variable cutoff (Velvet-SC). (d) After removing the chimeric bridge, we recondense the graph to obtain two contigs with high coverage support.



Supplementary Figure S5. Velvet and Velvet-SC assembly algorithms.

(a) ***Velvet*** **assembly algorithm**

Inputs: An odd integer $k > 0$ (k-mer size),
an integer coverage cutoff > 1,
and a set of reads *R*.

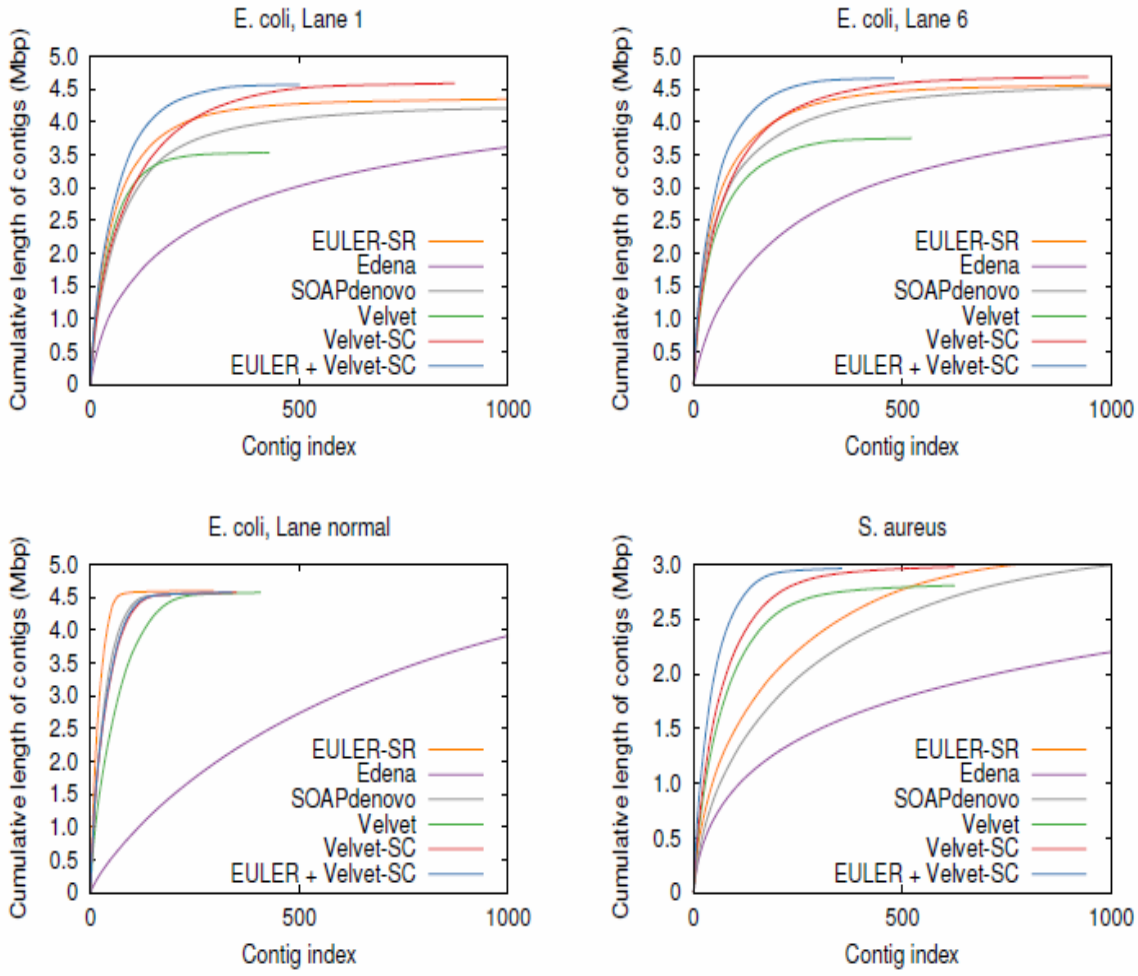
Output: Assembly contigs.

1: Build a roadmap *rdmap* from *R* by indexing all *k*-mers.
2: Build a de Bruijn pregraph *pg* from *rdmap.*
3: Clip tips of *pg.*
4: Build a *graph* from *pg* by threading *R*.
5: Condense *graph* by merging 1-in 1-out vertices.
6: Clip tips of *graph*.
7: Correct *graph* by the Tour Bus algorithm.
8: Remove edges of *graph* with average coverage < cutoff.
9: Clip tips of *graph*.
10: Correct *graph* by the Tour Bus algorithm.
11: Resolve repeats using read pairing information in *R*.
12: Condense *graph* by merging 1-in 1-out vertices.
13: Return edges of *graph* as contigs.

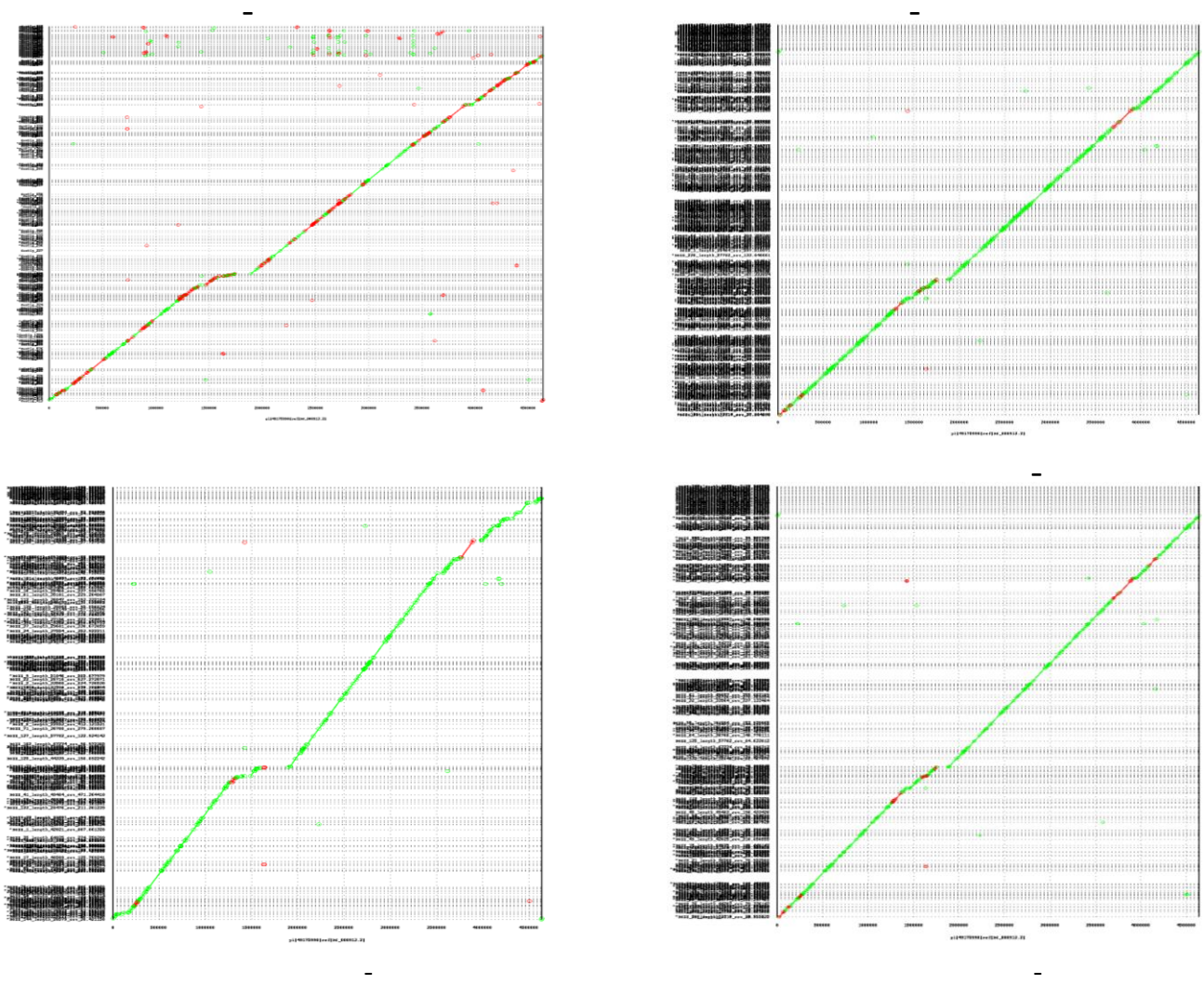
(b) ***Velvet-SC*** **assembly algorithm**

1-7: Same as *Velvet* assembly algorithm.
8: **for** *i* =2 to cutoff **do**
9: Remove edges of *graph* with average coverage < *i*.
10: Clip tips of *graph*.
11: Correct *graph* by the Tour Bus algorithm.
12: Resolve repeats using read pairing information in *R*.
13: Condense *graph* by merging 1-in 1-out vertices.
14: **end for**
15: Return edges of *graph* as contigs.

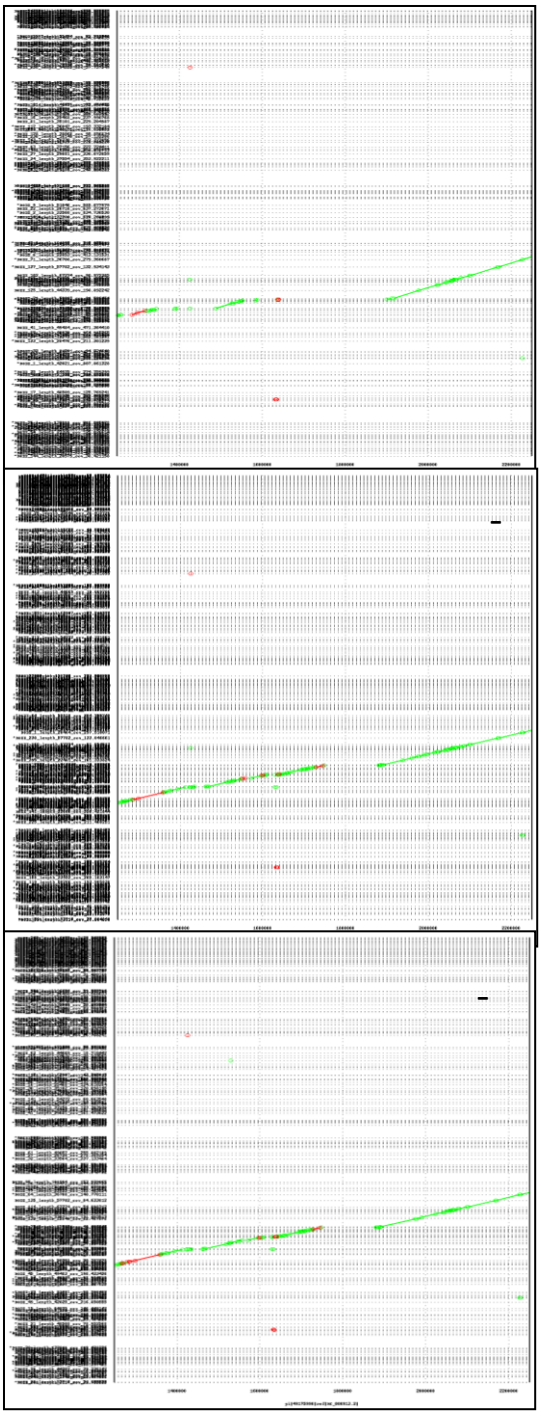
Supplementary Figure S6. Comparison of different assemblers. The contigs from Table 1 are sorted in descending order of sizes, and the *y*-axis is the cumulative size of *x* longest contigs.



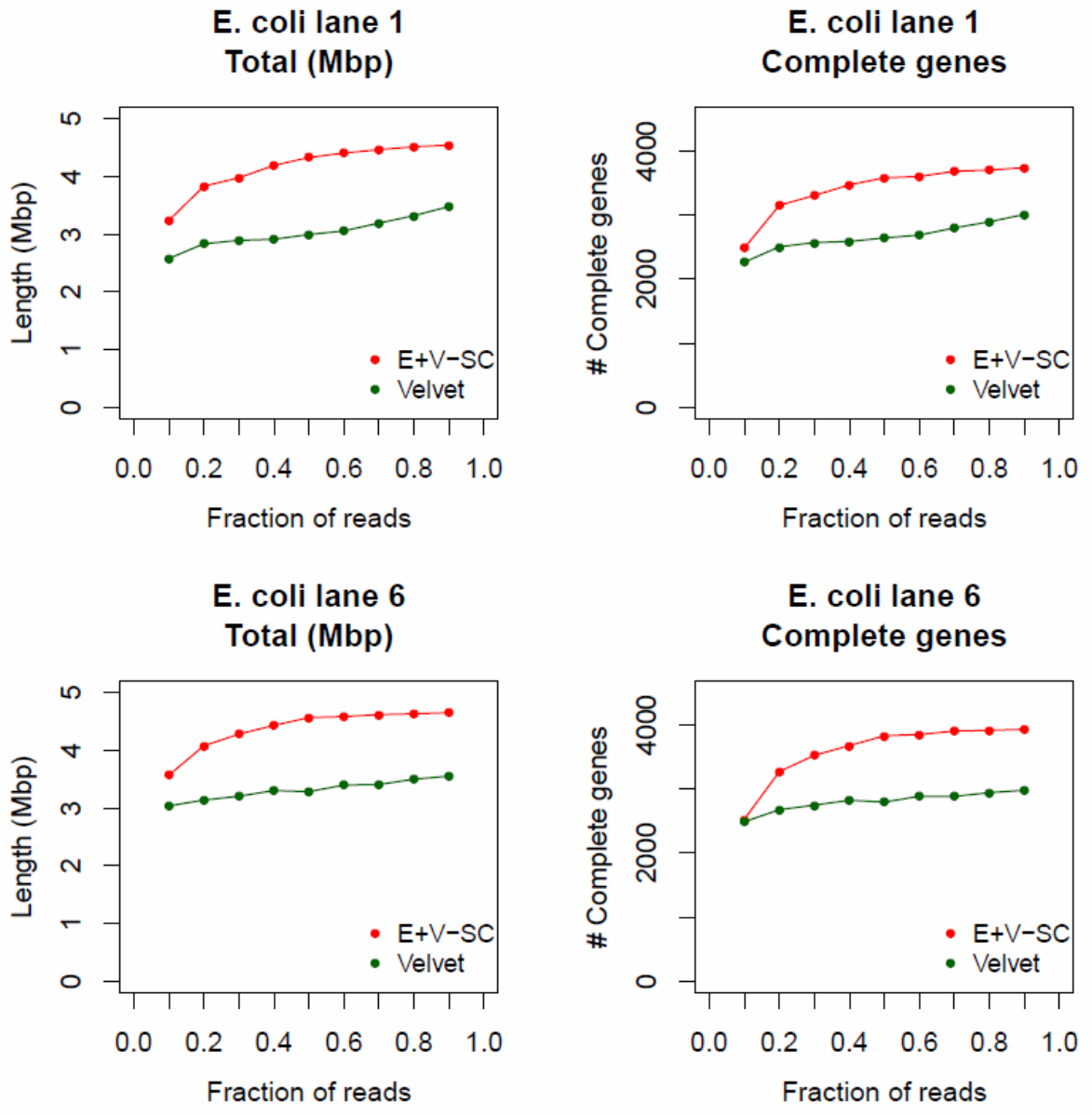
Supplementary Figure S7. For the lane 1 *E. coli* single cell dataset, contigs from the EULER-SR, Velvet, Velvet-SC and EULER+Velvet-SC assemblers are each aligned against the *E. coli* reference genome. The X axis indicates 500,000 bp increments of the reference genome, and assembly contigs are on the Y axis. Nucmer [16] was run (default parameters except for –l 80), and results were visualized using mummerplot to identify synteny. Scaffolds with more than 20 bps unaligned on either end were highlighted in red.. Assemblies using Velvet show a marked decrease in off-diagonal alignments relative to the EULER-SR assembly indicating that short potentially duplicate contigs are more completely incorporated in Velvet assemblies. There is no evidence of mis-assembly within contigs that would be evidenced by cross-diagonal alignments.



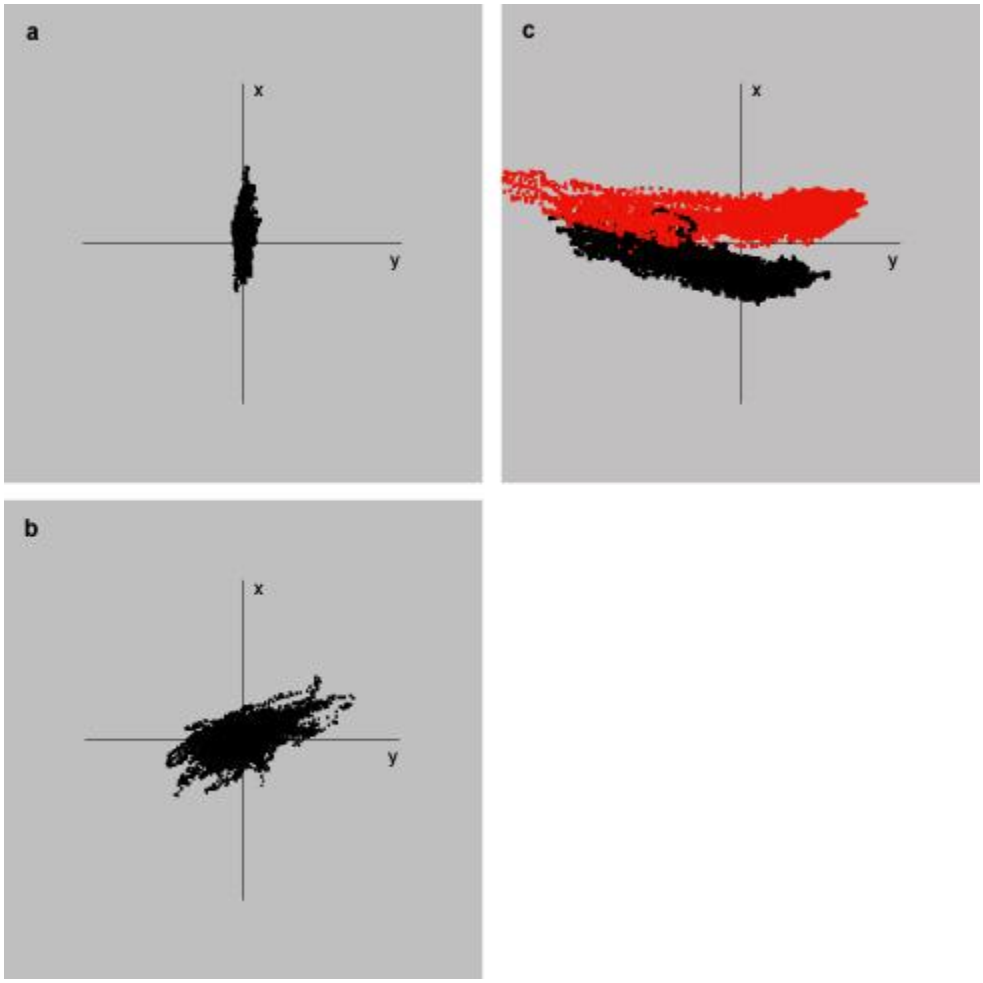
Supplementary Figure S8. For the lane 1 dataset, the 1.25-2.25 Mbp region of the *E. coli* reference genome (on each X axis) that includes regions with no or low coverage (see Supplementary Fig. S7) is compared to the 3 different assembies (on the Y axes) using Nucmer [16] as in Supplementary Figure S7. Velvet-SC enhances assembly within this region of reduced coverage.



Supplementary Figure S9. Comparison of single cell *E. coli* assembly results for various fractions of the original read data sets of lane 1 and lane 6. Total nucleotides in the assembly and number of complete genes present in the assembly are presented. Velvet and EULER+Velvet-SC were run with *k*-mer size 55 on fractions 0.1, 0.2, ...,, 0.9 of the original read data set. Both Velvet and EULER+Velvet-SC deteriorate significantly and show similar results when the fraction of used reads is reduced to 0.1 (corresponding to coverage around 60). A similar deterioration effect was observed for multicell assembly[1] when the coverage is reduced below 30 but the deterioration is more pronounced in the single cell case with non-uniform coverage. Since only contigs without gaps in coverage are assembled by the de Bruijn approach, the coverage in the single cell projects should be relatively high (preferably exceeding 300) as compared to multicell projects. Gene annotations are from http://www.ecogene.org.

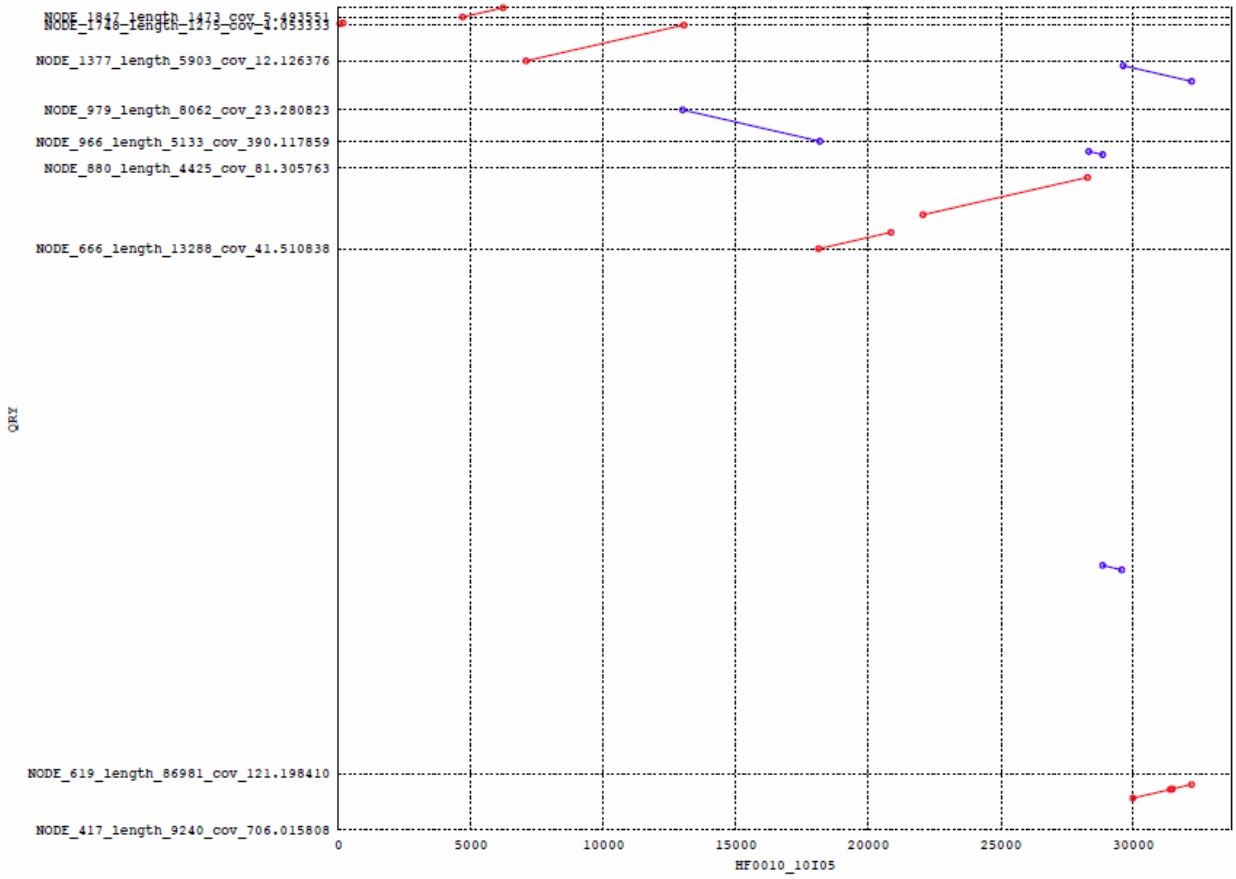


Supplementary Figure S10. Principal component analysis. Principal component analysis of 5-mers from the E+V-SC SAR324_MDA assembly (from 10 kb pieces) was performed using a Multi-dimensional Scatter Plot Viewer representation (http://gos.jcvi.org/openAccess/scatterPlotViewer2.html), with the following changes from default settings: word size, 5; chop sequences at size 10,000, overlap 9700. **b** is a 90 degree rotation of **a**. The planar, ovoid shape is typical of genomes from reference strains. **c** is an example from the website of two reference genomes plotted together.



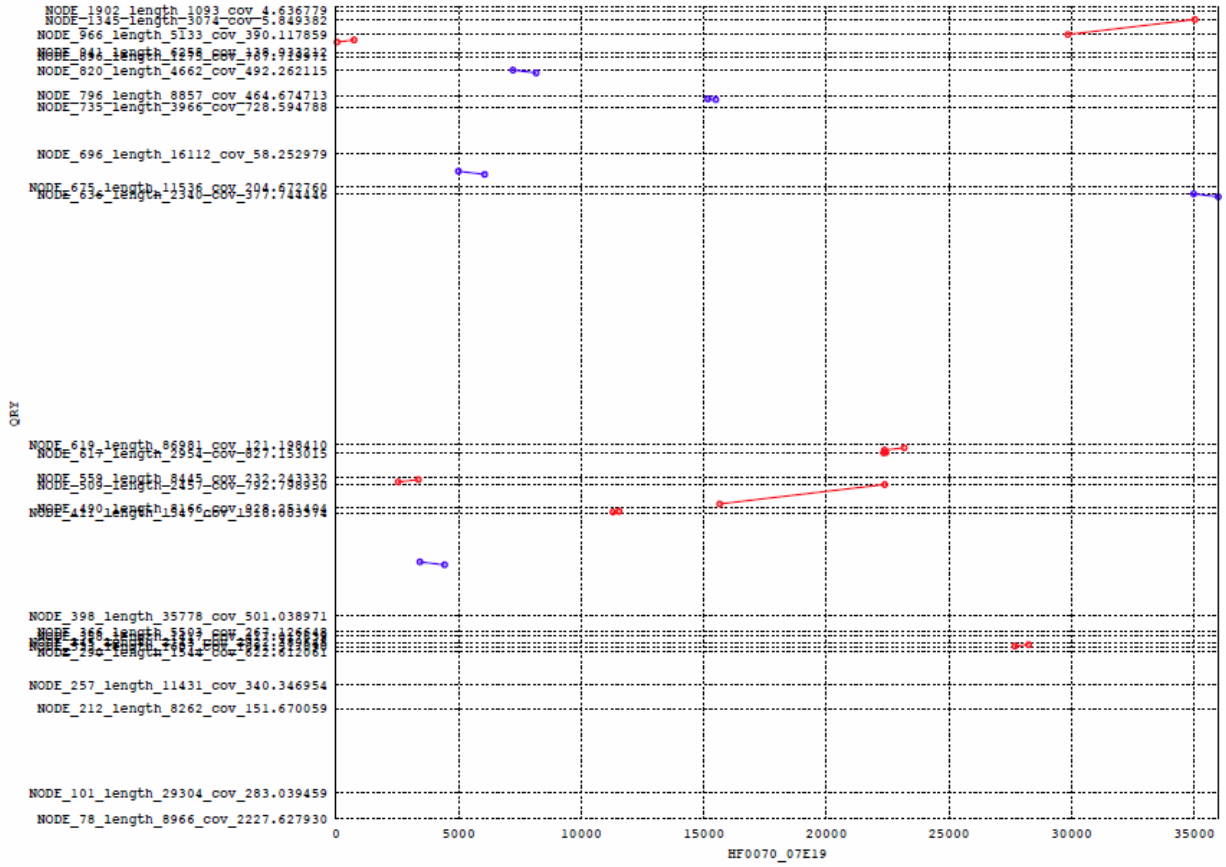
Supplementary Figure S11. Nucmer plots of nucleotide similarity between SAR324 fosmids and the SAR324_MDA assembly contigs (QRY). The plots were generated with a default setting of 20 bp minimum identity for extension. HF0010_10I05 (**a**) and HF0070_07E19 (**b**) are 36 kb and 34 kb, respectively [17]. Node_966 of the assembly includes tRNA and 16S and 23S genes, and has 98% ID to each fosmid over a 5.2 kb region. The % similarity for alignments between the fosmids and the other contigs (represented as "nodes" or "vertices") ranges from 82%-100%.

(**a**)

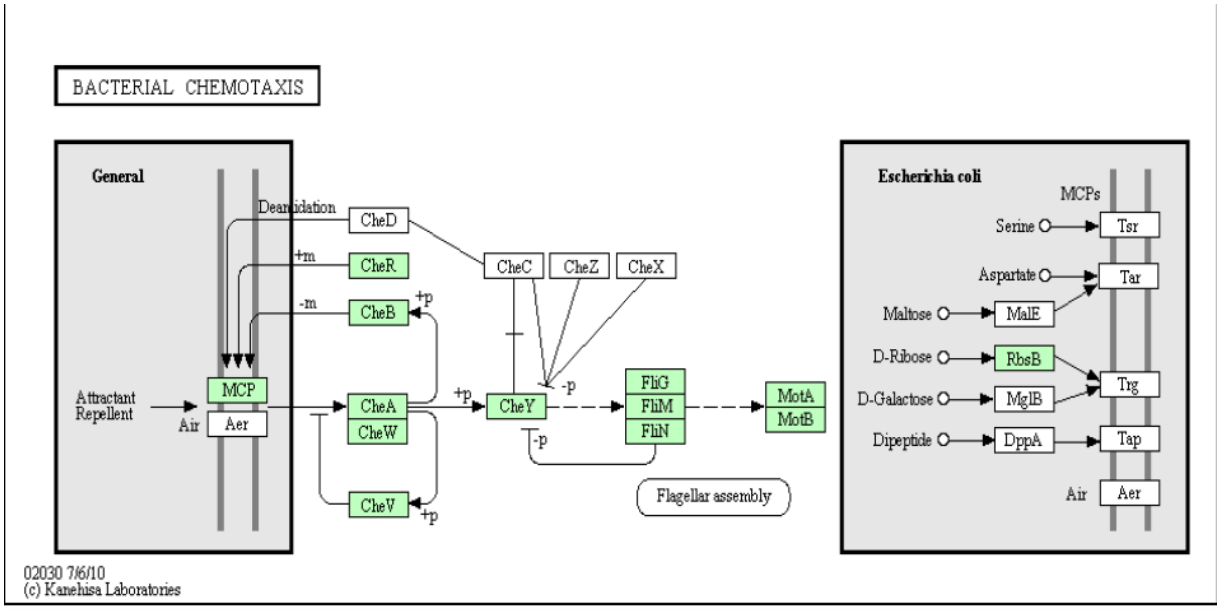


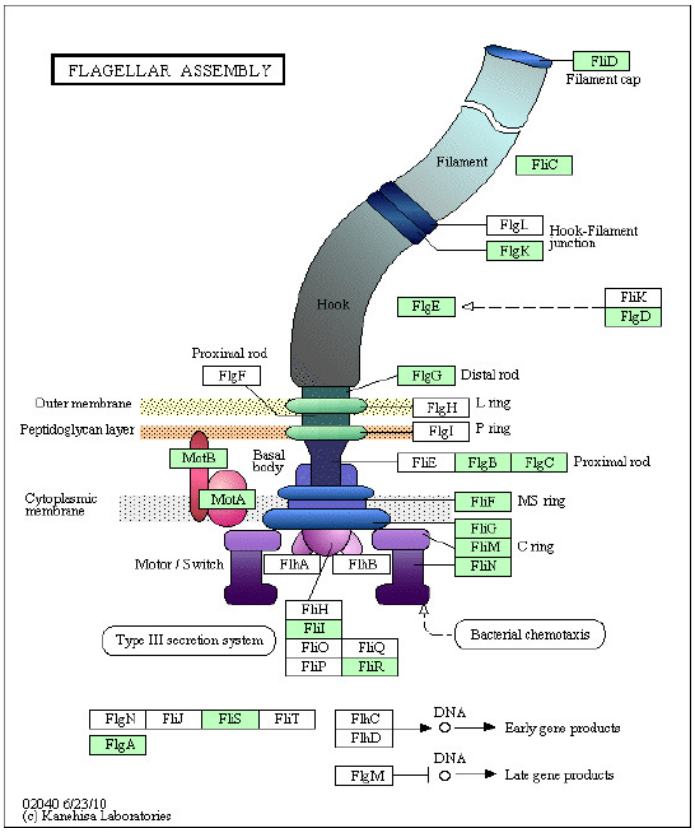
Supplementary Figure S11. Continued.

(**b**)

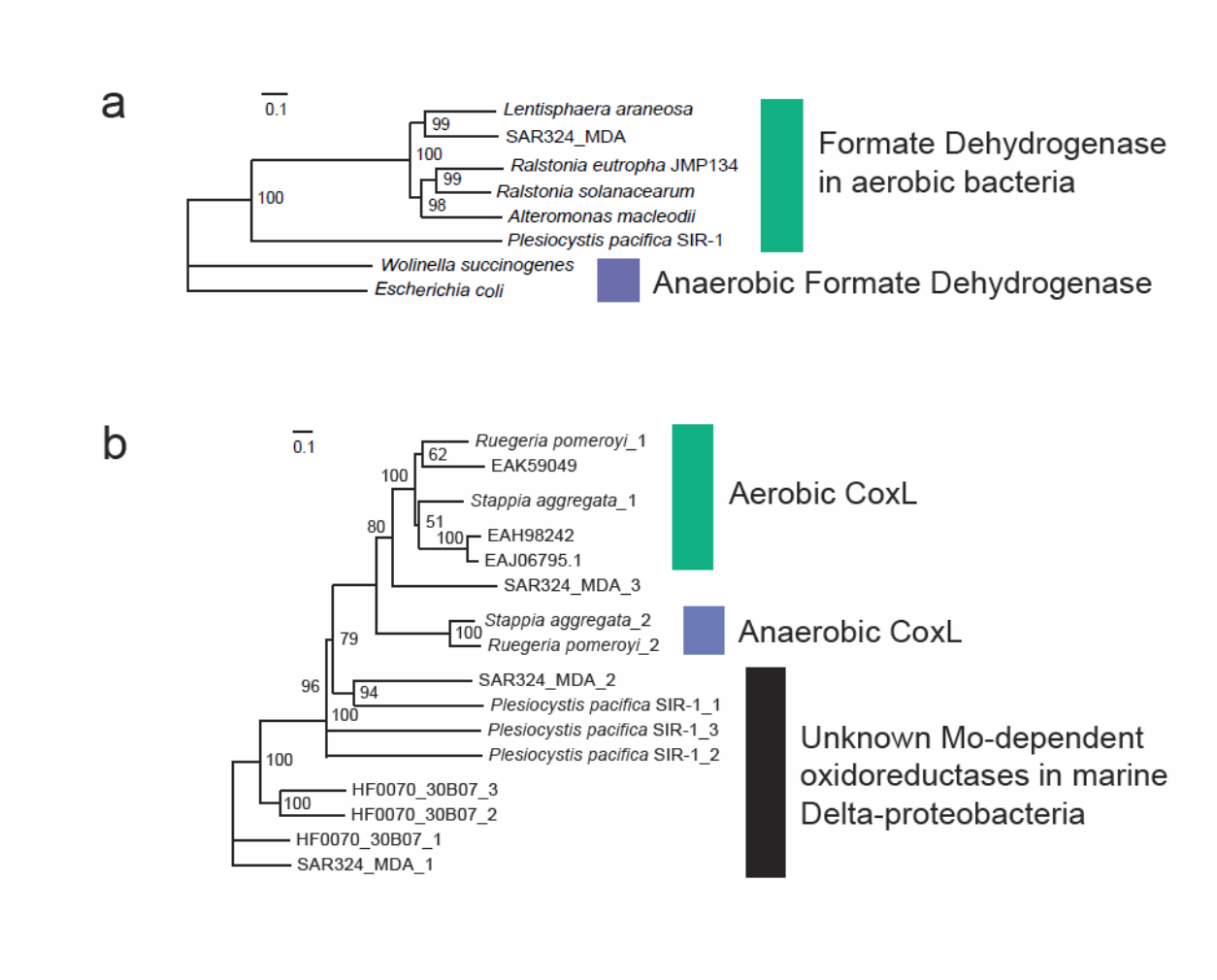


Supplementary Figure S12. SAR324_MDA-encoded proteins of the bacterial chemotaxis and flagellar assembly pathways. Proteins (rectangles) present in the assembly are green, undetected components are white. Proteins for KEGG pathway mapping were identified amongst the metagenomic ORFs using the KAAS annotator (see Methods).





Supplementary Figure S13. Phylogeny of MoCo-binding proteins in SAR324_MDA. Maximum likelihood phylogenies of putative Formate Dehydrogenases (**a**) and CO Dehydrogenases (**b**) found in the assembly and other Bacteria. Bootstrap support of greater than 50% for 100 replicates is shown. HF0070_30B07 is a SAR324 clone, also present on the 16S phylogenetic tree in Figure 3. Numbered suffixes indicate different homologs from the source sequence.



References

1. Chaisson, M.J. & Pevzner, P.A. Short read fragment assembly of bacterial genomes. *Genome Res.* **18**, 324-330 (2008).
2. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821-829 (2008).
3. Diep, B.A. et al. Complete genome sequence of USA300, an epidemic clone of community-acquired meticillin-resistant Staphylococcus aureus. *Lancet* **367**, 731-739 (2006).
4. Ishoey, T., Woyke, T., Stepanauskas, R., Novotny, M. & Lasken, R.S. Genomic sequencing of single microbial cells from environmental samples. *Curr. Opin. Microbiol.* **11** 198-204 (2008).
5. Marcy, Y. et al. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci U S A* **104**, 11889-11894 (2007).
6. Raghunathan, A. et al. Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* **71**, 3342-3347 (2005).
7. Lane, D.J. in Nucleic acid techniques in bacterial systematics. (eds. S. E & G. M) p. 115-175 (John Wiley and Sons, Ltd, New York, NY; 1991).
8. Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261-5267 (2007).
9. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797 (2004).
10. Koski, L.B. & Golding, G.B. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **52**, 540-542 (2001).
11. Tatusov, R.L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
12. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
13. Lasken, R.S. & Stockwell, T.B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol.*, 19 (2007).
14. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).
15. Cox, A.J. European Patent No. EP1704506: Multiple inexact pattern matching. (2006).
16. Delcher, A.L., Phillippy, A., Carlton, J. & Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478-2483 (2002).
17. Rich, V.I., Pham, V.D., Eppley, J., Shi, Y. & Delong, E.F. Time-series analyses of Monterey Bay coastal microbial picoplankton using a 'genome proxy' microarray. *Environ. Microbiol.* (2010).