nature biotechnology

Efficient *de novo* assembly of single-cell bacterial genomes from short-read data sets

Hamidreza Chitsaz[1,6], Joyclyn L Yee-Greenbaum[2,6], Glenn Tesler[3], Mary-Jane Lombardo[2], Christopher L Dupont[2], Jonathan H Badger[2], Mark Novotny[2], Douglas B Rusch[4], Louise J Fraser[5], Niall A Gormley[5], Ole Schulz-Trieglaff[5], Geoffrey P Smith[5], Dirk J Evers[5], Pavel A Pevzner[1] & Roger S Lasken[2]

Whole genome amplification by the multiple displacement amplification (MDA) method allows sequencing of DNA from single cells of bacteria that cannot be cultured. Assembling a genome is challenging, however, because MDA generates highly nonuniform coverage of the genome. Here we describe an algorithm tailored for short-read data from single cells that improves assembly through the use of a progressively increasing coverage cutoff. Assembly of reads from single *Escherichia coli* and *Staphylococcus aureus* cells captures >91% of genes within contigs, approaching the 95% captured from an assembly based on many *E. coli* cells. We apply this method to assemble a genome from a single cell of an uncultivated SAR324 clade of Deltaproteobacteria, a cosmopolitan bacterial lineage in the global ocean. Metabolic reconstruction suggests that SAR324 is aerobic, motile and chemotaxic. Our approach enables acquisition of genome assemblies for individual uncultivated bacteria using only short reads, providing cell-specific genetic information absent from metagenomic studies.

A myriad of uncultivated bacteria are found in environments ranging from the surface ocean[1] to the human body[2]. Although metagenomics provides a method to study these bacterial communities, the resulting data are gene-centric, as it is difficult to conclude which genes are found together in a single organism. Methods based on amplifying DNA directly from individual cells without requiring growth in culture have enabled genome-centric sequencing studies, which are a powerful complement to metagenomics. In these methods, the femtograms of DNA present in a single cell are amplified into the micrograms of DNA necessary for existing sequencing technologies. Genomic sequencing from single bacterial genomes was first demonstrated[3] with cells isolated by flow cytometry, using MDA[4–6] to prepare the template. MDA is now the preferred method for whole genome amplification from single cells[7,8]. In the first attempt to assemble a complete bacterial genome from one cell[9] researchers further explored the challenges of assembly from DNA amplified using MDA, including amplification bias and the formation of chimeras in which noncontiguous sequences have been joined. Amplification bias results in orders-of-magnitude difference in coverage[3], and absence of coverage in some regions. Chimera formation occurs during the DNA branching process by which the phi29 DNA polymerase used in MDA amplifies the DNA[10], but greater sequencing coverage helps to alleviate this problem. Despite their limitations, even these first single-cell sequencing methods have enabled investigation of novel uncultured microbes[11–13].

Recent studies have continued to improve assemblies from single cells[14–18], but the full potential of single-cell sequencing has not yet been realized. The challenges facing single-cell genomics are increasingly computational rather than experimental[17]. All previous single-cell studies used standard fragment-assembly tools[19,20], developed for data models characteristic of standard (rather than single-cell) sequencing. These algorithms are not ideal for use with nonuniform read coverage because most existing fragment assembly tools implicitly assume nearly uniform coverage, and most produce erroneous contigs (that is, linking noncontiguous genomic fragments) when the rate of chimeric reads or chimeric read pairs exceeds a certain threshold. Thus, there is a need to adapt existing assembly tools for single-cell sequencing.

We developed a specialized software tool for assembling sequencing reads from single-cell MDAs. Applying it to assemble single-cell data sets from two known genomes and an unknown marine genome yielded valuable assemblies that identified the majority of genes, with no efforts to close gaps and resolve repeats.

RESULTS

Velvet-SC improves assembly of short reads with highly nonuniform coverage

There are two key algorithmic paradigms in fragment assembly: the overlap-layout consensus approach (which dominated assembly projects using long reads from Sanger sequencing) and the de Bruijn graph approach (which dominates next-generation sequencing assembly projects based on technologies that generate short reads). Most existing next-generation sequencing assemblers use a two-stage procedure[21] that first involves correcting errors in reads followed by

[1]Department of Computer Science, University of California, San Diego, La Jolla, California, USA. [2]J. Craig Venter Institute, San Diego, California, USA. [3]Department of Mathematics, University of California, San Diego, La Jolla, California, USA. [4]J. Craig Venter Institute, Rockville, Maryland, USA. [5]Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterfield, Nr Saffron Walden, Essex, UK. [6]These authors contributed equally to this work. Correspondence should be addressed to R.S.L. (rlasken@jcvi.org).

Received 21 December 2010; accepted 9 August 2011; published online 18 September 2011; doi:10.1038/nbt.1966

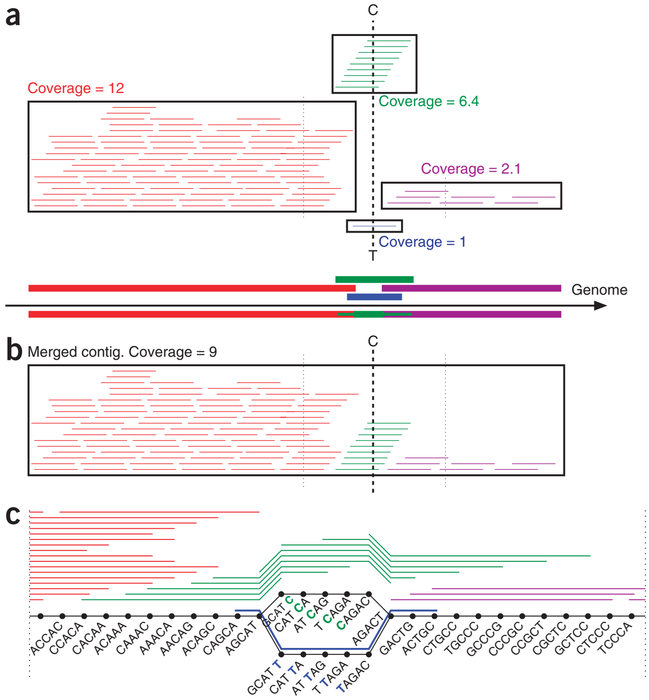

Figure 1 Assembling single-cell reads using Velvet-SC. (**a**) Coverage varies widely along the genome, between 1× and 12×. Reads (short thin colored lines) and potential contigs (thick lines) are positioned along the genome, with a box around the reads supporting each contig. There are two potential contigs to choose from in the middle, differing by a single nucleotide (C versus 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, 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 1×, which only removes the blue contig. (**b**) The other contigs are combined into a single contig of average coverage 9×. The purple region is salvaged by Velvet-SC because it was absorbed into a higher coverage region as coverage threshold increased. Velvet-SC repeats this process with a gradually increasing low-coverage threshold. (**c**) A portion of the de Bruijn graph for the contigs described in **a**. The black circles are the 'vertices' and represent 5-mer strings derived from the reads, which are indicated by colored lines alongside the chains of vertices, including a blue read with an erroneous T. The lines between the vertices are termed 'edges' and represent the overlaps between the 5-mers. The edges are directed from left to right in this example. The read with the C/T mismatch results in two alternative paths for assembly, both with five intermediate vertices. The lower of the two paths arises from the erroneous blue read and has coverage 1×; it is the only part of the graph eliminated by Velvet-SC, leaving a single chain of vertices that gives a single contig for the entire genome. See **Supplementary Figure 3** for an example of the condensing of contigs. An example of Velvet-SC handling of a chimeric read is presented in **Supplementary Figure 4**.

assembly of the error-corrected reads using a de Bruijn graph, a conceptual tool from computer science. A de Bruijn graph represents each read as a series of *k*-mers (*k* consecutive bases) by examining successive fragments *k*-bases long along the length of the read. The *k*-mers are represented as 'vertices' (shown as small circular points in **Fig. 1**), and consecutively overlapping *k*-mers are represented by 'edges' (shown as short lines). Reads will share vertices and edges when they have at least *k* consecutive bases in common. This construction reveals reads with shared *k*-mers and builds contigs in an efficient manner.

Nonuniform coverage poses a serious problem for existing *de novo* assembly algorithms. Of de Bruijn–based assemblers[21], for example, Velvet[19] and ABySS[22] use an average coverage cutoff threshold for contigs to prune out low-coverage regions, which tend to include more errors, whereas EULER-SR[23] uses a *k*-mer coverage cutoff. This pruning step reduces the complexity of the underlying de Bruijn graph substantially and makes the algorithms practical. Data sets of single-cell reads (detailed below) have highly variable coverage, and a single coverage cutoff prevents assembly of a substantial portion of the data (**Table 1**, **Fig. 2** and **Supplementary Fig. 1**). **Supplementary Figure 2** plots the percentage of positions with given coverage. In our multicell *E. coli* data set, most positions in the genome have coverage of 450–800×, and pruning by a coverage threshold helps eliminate erroneous reads; only 0.1% of positions have coverage <450×. In contrast, in one of our single-cell *E. coli* data sets (labeled 'lane 1' in **Table 1**), 5% of positions have <10× coverage and 11% of positions have <30× coverage. Assembly of reads from a standard multicell sample by current next-generation sequencing assemblers usually requires at least 30× coverage of a region for a successful assembly without gaps.

We developed the EULER+Velvet-SC algorithm specifically for single-cell assembly. Velvet-SC (Velvet single cell; source code available in **Supplementary Data 1** and at http://bix.ucsd.edu/singlecell/) is a modification of the popular open-source assembly program Velvet that incorporates lower coverage sequences that most existing assemblers discard. Briefly, instead of using a fixed cutoff to prune contigs from the de Bruijn graph that are covered on average by few reads, Velvet-SC uses a variable cutoff that starts at 1 and gradually increases. After the lowest coverage contigs are removed based on the current cutoff, some contigs may merge into a larger contig, whose average coverage is recomputed. This tends to incorporate low-coverage contigs into higher coverage regions, sparing them from deletion. This process is iterated with a gradually increasing cutoff (Online Methods, **Fig. 1** legend, **Supplementary Methods** and **Supplementary Figs. 3**–**5** for further explanation of Velvet and Velvet-SC). EULER+Velvet-SC combines the error correction from EULER-SR—source code available in **Supplementary Data 2** and at http://bix.ucsd.edu/singlecell/—with Velvet-SC to further improve the assembly.

Characteristics of single-cell sequences

DNA amplified from single cells displays a range of genome representation due to amplification bias, template quality and presence of contaminating DNA, as discussed previously[3,17]. For this study, sets of DNA amplified from single *E. coli* and *S. aureus* cells prepared in parallel from clonal populations were evaluated for genome representation using quantitative PCR for ten loci, as described previously[3]. In some amplified DNA, a few loci were detected and some in all were detected (data not shown). We chose amplified DNA from two *E. coli* cells and one *S. aureus* cell for which all ten loci were detected for this study. An average of 93% of reads from the three cells mapped to the respective reference genomes (**Supplementary Table 1**) versus 99% of the reads in the multicell *E. coli* data set. Nonmapping reads in MDA data sets can often be attributed to minor contaminating sequences[17] (**Supplementary Tables 2** and **3** and **Supplementary Data 3**). Chimeric fragments (where the ends map to different regions of the genome) were 2% of the *E. coli* read pairs and 0.5% of the *S. aureus* read pairs (**Supplementary Table 3** and **Supplementary Data 3**). These data are consistent with previous data regarding chimeras in MDA sequence data sets[9,10].

The single-cell data sets (reads available at http://bix.ucsd.edu/singlecell/) display highly nonuniform coverage typical of single-cell amplification[3] (**Supplementary Figs. 1** and **2** and **Supplementary Table 4**), including blackout regions, which are contiguous regions of

Table 1 Comparison of assemblies of known genomes (for contigs >110 bp)

Data set	Assembler	No. contigs	N_{50} (bp)	Largest (bp)	Total (bp)	Subs. error (per 100 kbp)	Known genes	Complete genes	Partial genes	Predicted operons	Complete operons	Partial operons
E. coli lane 1	EULER-SR	1,344	26,662	**140,518**	4,369,634	16.1	4,324	3,178	627	884	553	248
	Edena	1,592	3,919	44,031	3,996,911	**2.6**		2,425	1,112		317	444
	SOAPdenovo	1,240	18,468	87,533	4,237,595	98.3		3,021	612		520	248
	Velvet	**428**	22,648	132,865	3,533,351	3.0		3,055	**170**		584	**106**
	Velvet-SC	872	19,791	121,367	**4,589,603**	4.4		3,617	325		643	184
	E+V-SC	501	**32,051**	132,865	4,570,583	2.7		**3,753**	185		**713**	109
E. coli lane 6	EULER-SR	1,820	29,551	170,385	4,469,152	16.6		3,339	734		561	283
	Edena	1,536	4,899	42,342	4,147,566	3.2		2,705	1,075		368	428
	SOAPdenovo	1,397	20,319	**204,730**	4,576,388	48.3		3,353	646		586	244
	Velvet	522	18,410	168,533	3,753,818	4.1		3,131	253		566	145
	Velvet-SC	945	27,113	144,462	**4,688,759**	3.8		3,779	409		694	161
	E+V-SC	**481**	**36,581**	173,901	4,668,135	**1.7**		**3,943**	**158**		**749**	**101**
E. coli lane normal	EULER-SR	295	**110,153**	**221,409**	4,598,020	3.5		4,119	115		788	80
	Edena	1,673	3,814	20,470	**4,611,645**	3.8		3,019	1,189		317	538
	SOAPdenovo	**192**	62,512	172,567	4,529,677	26.8		**4,128**	81		802	**53**
	Velvet	408	31,503	129,378	4,569,225	1.6		4,061	139		760	108
	Velvet-SC	350	52,522	166,115	4,571,760	**1.1**		4,121	157		804	58
	E+V-SC	339	54,856	166,115	4,571,406	1.5		4,124	**66**		**808**	57
S. aureus	EULER-SR	4,398	7,247	66,549	**3,376,776**	53.1	2,622	1,958	640	—	nd	nd
	Edena	1,288	1,881	37,770	2,358,911	**3.0**		1,222	925			
	SOAPdenovo	2,470	5,385	37,397	3,273,188	42.9		482	1,740			
	Velvet	625	15,800	67,677	2,807,042	6.2		2,244	268			
	Velvet-SC	1,084	20,163	76,884	3,001,635	4.2		2,100	458			
	E+V-SC	**355**	**32,296**	**107,657**	2,962,136	4.7		**2,408**	**173**			

Number of contigs, genome N_{50}, the length of the largest contig, total nucleotides in the assembly, substitution error rate in the assembled contigs (per 100 kbp), number of genes completely or partially present in the assembly, and number of operons completely or partially present in the assembly. Partial means that a gene and a contig (or an operon and a contig) have an overlap of at least 100 nucleotides. Best by each criteria is indicated in bold. EULER-SR 2.0.1, Velvet 0.7.60, Velvet-SC and EULER+Velvet-SC were run with k-mer size equal to 55. Edena 2.1.1 (ref. 40) was run with a minimum overlap of 55. SOAPdenovo 1.0.4 (ref. 41) was run with k = 27–31. E+V-SC stands for EULER+Velvet-SC. Gene annotations were from http://www.ecogene.org/ (*E. coli*) and http://cmr.jcvi.org/cgi-bin/CMR/GenomePage.cgi?org=ntsa10 (*S. aureus*). Operon annotations (*E. coli*) were from http://csbl1.bmb.uga.edu/OperonDB/displayNC.php?id=215 (ref. 42). Some of the contigs in the single-cell assemblies represent contaminants. nd, not determined.

the genome to which no reads aligned (coverage 0). Single-cell sequencing at ~600× depth results in 94 and 50 blackout regions for our two *E. coli* data sets 'lane 1' and 'lane 6', respectively, whereas sequencing of unamplified DNA results in no blackout. Genome regions with coverage 0 or 1 comprise ~116 kbp in *E. coli* lane 1 and ~13 kbp in lane 6. There are only two small blackout regions in the *S. aureus* ~2,300× coverage data set, comprising just 143 bases. These observations illustrate the substantial variability in coverage even for MDAs generated from single cells processed in parallel from the same culture (lane 1 versus 6). As evident with the two *E. coli* data sets, blackout regions can potentially be eliminated by combining reads from multiple single cells when available[3,13,16].

De novo single-cell assembly of *E. coli* and *S. aureus*

De novo assemblies generated by Velvet, Velvet-SC and EULER+Velvet-SC were compared with those generated by several other assemblers (**Table 1**, **Fig. 2** and **Supplementary Fig. 6**). The metrics compared were the percentage of the genome present in the final assembly

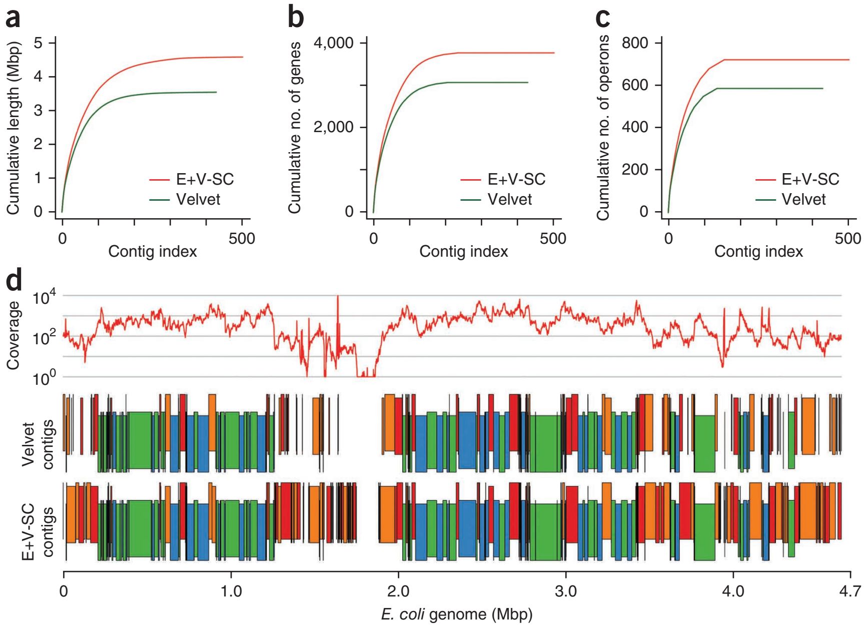

Figure 2 Comparison of contigs generated by Velvet versus EULER+Velvet-SC for single-cell *E. coli* lane 1. (**a**–**c**) Contigs are those presented in **Table 1** and are ordered from largest to smallest number of bases. The *y* axis shows the cumulative length (**a**), the cumulative number of genes (**b**) and the cumulative number of operons in the contigs (**c**). EULER+Velvet-SC improves upon Velvet in all three plots. (**d**) Average read coverage over a 1,000-bp window (top, log scale), Velvet contigs (middle) and EULER+Velvet-SC contigs (bottom) mapped along the *E. coli* reference genome, with vertical staggering to help visualize small contigs. Contigs in blue or green match between the assemblies. Contigs in red or orange differ between the assemblies; they either have substantially different lengths, are broken into a different number of contigs, or are present in one assembly but missing in the other.

(in terms of bases, genes and operons), N_{50} (the contig length at which all longer contigs represent half of the total genome length) and substitution error rate per 100 kbp. We note that the single-cell assemblies include some contigs that do not map to the *E. coli* or *S. aureus* genomes. As both nonmapping and mapping reads are informative with regard to assembler functionality, and for simplicity in presenting the data, we have not removed them from the analysis.

We find that EULER+Velvet-SC outperforms Velvet-SC, and Velvet-SC outperforms Velvet. For example, from the *E. coli* single-cell lane 1 data set EULER+Velvet-SC assembled 4.57 Mb of contigs whereas Velvet assembled only 3.53 Mb. EULER+Velvet-SC assembled with an N_{50} of 32.1 kbp compared to an N_{50} of 22.6 kbp with Velvet. EULER+Velvet-SC achieved an error rate of 2.7 mismatches per 100 kbp compared to an error rate of 3.0 mismatches per 100 kbp with Velvet. For single-cell *S. aureus* reads, Velvet assembled 2.81 Mb of contigs with an N_{50} of 15.8 kbp and 6.2 mismatches per 100 kbp whereas EULER+Velvet-SC assembled 2.96 Mb of contigs with an N_{50} of 32.3 kbp and 4.7 mismatches per 100 kbp.

Single-cell assembly of *E. coli* (lane 6) with EULER+Velvet-SC captured 91.2% of *E. coli* genes and 84.7% of *E. coli* operons in single contigs, slightly less than 95.4% and 91.4%, respectively, captured in a multicell *E. coli* assembly. Single-cell assembly of *S. aureus* captured 91.8% of *S. aureus* genes in single contigs. EULER+Velvet-SC captured sequences from two of the three plasmids in this *S. aureus* strain[24] (pUSA02: 4,439 bp in one contig; pUSA03: 37,136 bp in 30 contigs) whereas Velvet captured sequences from only one plasmid. The EULER+Velvet-SC assembly of *E. coli* had no misassembled contigs as determined by BLAST analysis (**Supplementary Fig. 7**), whereas the assembly of *S. aureus* had one misassembled contig. The Velvet-SC algorithm extended contigs into regions of low coverage (**Fig. 2** and **Supplementary Fig. 8**). EULER+Velvet-SC captured more bases, genes and operons, and assembled more regions with low-read coverage than Velvet on single-cell *E. coli* lane 1 (**Fig. 2**). By these tests, EULER+Velvet-SC outperformed the other assemblers, generating higher quality single-cell assemblies.

To test the effect of sequencing with lower coverage, we randomly selected a fraction of the input reads ranging from 0.1 to 0.9 of the total and assembled them with both EULER+Velvet-SC and Velvet (**Supplementary Fig. 9**). As expected, for single-cell *E. coli* data sets (lanes 1 and 6), increased coverage gave better results, and EULER+Velvet-SC outperformed Velvet for total bp assembled at all coverage depths. Assembly of half the reads allowed capture of 3,579 of the 3,753 complete genes captured in lane 1 using all the reads, with a similar result for lane 6 (**Supplementary Fig. 9**) and data not shown), suggesting sequencing effort could be diminished by half with minimal effect on the resulting assembly.

Single-cell assembly of an uncultured Deltaproteobacterium

To demonstrate the performance of EULER+Velvet-SC with an uncultivated organism, a genome of a marine bacterium was sequenced from a single cell isolated from a marine sample collected at La Jolla, California (Online Methods). Single-cell MDA reactions were screened by 16S PCR and an uncultured Deltaproteobacterium was chosen for testing the *de novo* assembly methods. Phylogenetic analysis of 16S sequences (**Fig. 3**) revealed that this organism is a member of the deeply branched and divergent clade of uncultured deltaproteobacteria designated SAR324 (ref. 25). As for the *E. coli* and one *S. aureus* cells, we generated SAR324_MDA reads (reads available at http://bix.ucsd.edu/singlecell/; Short Read Archive (SRA) accession SRA043956) from a 100-bp paired-end run of the Illumina GA pipeline, and 57,816,790 of 67,995,232 reads passed the Illumina purity filter.

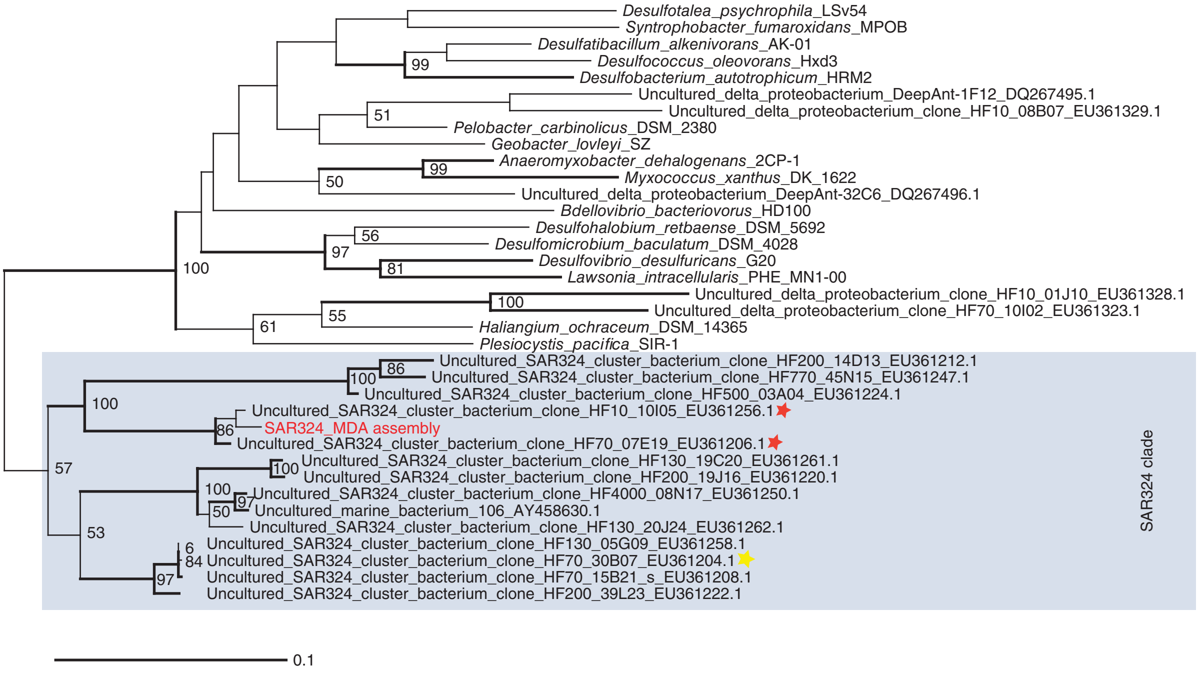
Desulfotalea_psychrophila_LSv54
Syntrophobacter_fumaroxidans_MPOB
99
Desulfatibacillum_alkenivorans_AK-01
Desulfococcus_oleovorans_Hxd3
Desulfobacterium_autotrophicum_HRM2
Uncultured_delta_proteobacterium_DeepAnt-1F12_DQ267495.1
51
Uncultured_delta_proteobacterium_clone_HF10_08B07_EU361329.1
Pelobacter_carbinolicus_DSM_2380
Geobacter_lovleyi_SZ
99
Anaeromyxobacter_dehalogenans_2CP-1
50
Myxococcus_xanthus_DK_1622
Uncultured_delta_proteobacterium_DeepAnt-32C6_DQ267496.1
Bdellovibrio_bacteriovorus_HD100
56
Desulfohalobium_retbaense_DSM_5692
97
Desulfomicrobium_baculatum_DSM_4028
100
81
Desulfovibrio_desulfuricans_G20
Lawsonia_intracellularis_PHE_MN1-00
100
Uncultured_delta_proteobacterium_clone_HF10_01J10_EU361328.1
55
Uncultured_delta_proteobacterium_clone_HF70_10I02_EU361323.1
61
Haliangium_ochraceum_DSM_14365
Plesiocystis_pacifica_SIR-1
86
Uncultured_SAR324_cluster_bacterium_clone_HF200_14D13_EU361212.1
100
Uncultured_SAR324_cluster_bacterium_clone_HF770_45N15_EU361247.1
100
Uncultured_SAR324_cluster_bacterium_clone_HF500_03A04_EU361224.1
Uncultured_SAR324_cluster_bacterium_clone_HF10_10I05_EU361256.1 ★
86
SAR324_MDA assembly
Uncultured_SAR324_cluster_bacterium_clone_HF70_07E19_EU361206.1 ★
100
Uncultured_SAR324_cluster_bacterium_clone_HF130_19C20_EU361261.1
57
Uncultured_SAR324_cluster_bacterium_clone_HF200_19J16_EU361220.1
100
97
Uncultured_SAR324_cluster_bacterium_clone_HF4000_08N17_EU361250.1
50
Uncultured_marine_bacterium_106_AY458630.1
Uncultured_SAR324_cluster_bacterium_clone_HF130_20J24_EU361262.1
53
6
Uncultured_SAR324_cluster_bacterium_clone_HF130_05G09_EU361258.1
84
Uncultured_SAR324_cluster_bacterium_clone_HF70_30B07_EU361204.1 ★
97
Uncultured_SAR324_cluster_bacterium_clone_HF70_15B21_s_EU361208.1
Uncultured_SAR324_cluster_bacterium_clone_HF200_39L23_EU361222.1
SAR324 clade
0.1

Figure 3 A 16S maximum likelihood tree of Deltaproteobacterial 16S sequences including SAR324_MDA (red). Sequences with species identification are from representative Deltaproteobacterial reference genomes in GenBank. The environmental 16S sequences (designated uncultured SAR324 or uncultured Deltaproteobacteria) were retrieved from GenBank based on their accession numbers (see Fig. S3 of ref. 30). The sequences were aligned using MOTHUR[38]. The tree was inferred using the nucleotide maximum likelihood feature of PAUP* 4.0b10 (ref. 39). Branches drawn in thick lines are clades with bootstrap support of 75% or greater. Scale at bottom of figure indicates the branch length associated with 0.1 substitutions per position. Sequences present on fosmids with extensive nucleotide similarity to the SAR324_MDA assembly are indicated (red star), as is a SAR324 fosmid (yellow star) encoding CoxL homologs also present in the SAR324_MDA assembly (**Supplementary Fig. 13**).

Table 2 Comparison of Velvet-based assembler results ($k = 55$) on SAR324_MDA assembly

Assembler	No. of contigs	N_{50} (bp)	Largest (bp)	Total (bp)	No. ORFs (MetaGene)	No. ORFs (APIS)	No. COGs	No. conserved single-copy genes
Velvet	1,856	11,531	100,589	3,921,396	4,575	2,462	2,160	55/111 (46%)
Velvet-SC	933	23,230	113,282	4,284,882	4,234	2,627	2,307	75/111 (67%)
E +V-SC	823	30,293	113,282	4,282,110	4,154	2,604	2,281	75/111 (67%)

Total number of contigs; assembly N_{50} (for contigs >110 bp); length of the largest contig (for contigs >110 bp); total nucleotides in the assembly (for contigs >110 bp); number of ORFs > 20 bp predicted by MetaGene; number of ORFs with phylogenetic assignments by APIS; number of ORFs with COGs identified by BLAST; and number of 111 conserved single-copy genes present.

Assembly statistics

As expected, EULER+Velvet-SC outperformed Velvet and Velvet-SC in single-cell assembly of the uncultured Deltaproteobacterium (**Table 2**), with a higher N_{50} and lower total number of contigs. The ability of the assemblies to support open reading frame (ORF) prediction was tested using MetaGene[26], a program designed for annotation of metagenomic sequences that uses less stringent criteria than traditional annotation tools. The decreased number of ORFs from Velvet to Velvet-SC to EULER+Velvet-SC suggests that both the greater bp incorporation and the EULER-SR error correction reduced spurious ORF calls. The EULER+Velvet-SC ORFs were of higher quality as evidenced by the greater numbers of ORFs with taxonomic affiliations identified using BLAST and phylogenetic analysis using the Automated Phylogenetic Inference System (APIS; **Supplementary Methods**), by the greater numbers of ORFs corresponding to orthologous genes in the Clusters of Orthologous Groups (COG) database[27] and by greater numbers of single-copy conserved genes detected. By all these criteria, EULER+Velvet-SC yielded the most robust assembly for annotation.

Assembly purity

Single-cell MDAs may sometimes contain DNA from other organisms originating from the MDA reagents or the biological source material[17]. We assessed contamination in the SAR324_MDA assembly by analyzing the GC content and nucleotide frequencies of reads and contigs and by comparing them against reference bacterial genomes and the NCBI nr database using BLAST. Briefly, BLAST analysis of contigs revealed that all top BLAST hits for contigs >500 bp were to uncultured marine organisms (data not shown), supporting the novelty of the single-cell genome, its marine origin and an absence of known DNA contaminants. Principal component analysis of nucleotide frequencies of the contigs examined in three-dimensional space was consistent with a single genome being present (**Supplementary Fig. 10**). A plot of the GC content of the reads forms a unimodal distribution, also consistent with the presence of a single genome (data not shown).

We also assessed the purity of the SAR324 genome by attempting to generate a phylogenetic tree for each ORF using the APIS software program, which performs BLAST analysis of each ORF against reference genomes and, when possible, generates a phylogenetic tree for each ORF (**Supplementary Methods**). APIS was originally designed to detect genes potentially acquired by horizontal gene transfer, for example, clusters of ORFs with phylogenies different from the rest of a genome. We reasoned that APIS might also help identify contaminant contigs as those containing ORFs that were phylogenetically similar to each other and distinct from the rest of the assembly. However, our analysis did not result in most contigs having a majority classification as expected. Instead, most contigs consisted of ORFs displaying a variety of predicted phylogenetic origins. Further investigation revealed that this might not be unexpected for SAR324 given that inconsistent phylogenetic signal has been observed in other Deltaproteobacteria[28] (and J.H.B., unpublished data), and that SAR324 is quite distant from available reference genomes.

Insights from the SAR324_MDA Deltaproteobacterium genome

Spatial and temporal studies of oceanic bacterial diversity show that SAR324 is cosmopolitan, appearing in both surface and deep ocean[29,30]. Although sequences of several diverse 30- to 40-kbp fosmids for SAR324 are available (representative 16S fosmid sequences from ref. 30 included in **Fig. 3**), the lack of even a draft reference genome for SAR324 prevents elucidation of its ecophysiological role. *Pleocystis pacifica*, the closest cultured relative for which a complete genome is available, as noted previously[31], is an obligate aerobe with a chemoheterotrophic lifestyle[32], and phenotypic characteristics of myxobacteria. However, the substantial phylogenetic distance between the SAR324 clade and *P. pacifica* suggests that the latter may have limited relevance to SAR324. Genome assembly attained using a culture-independent approach represents the best possibility for elucidating the ecological role of SAR324.

The SAR324_MDA assembly (assembly available at http://bix.ucsd.edu/singlecell/; Whole Genome Shotgun (WGS) accession AGAU00000000) includes about 4.3 Mb of nonredundant contigs yielding 3,811 ORFs (**Table 3**). We searched the ORFs for two sets of conserved genes typically found in a single copy within bacterial genomes, which can be used to estimate genome size[33–35]. Seventy-five of a set of 111 (67%) conserved single-copy genes[33] were represented, and 58 out of 66 (87%) single-copy gene clusters[34,35] were represented. A criterion of 90% of the 66 gene clusters is a suggested passing metric for draft genomes of cultured strains[35]. Extrapolating from these results suggests a complete genome size of 4.95–6.42 Mb and that the assembly contains a majority of the gene complement and should provide insight into SAR324. That the genome is not complete is typical of the majority of single-cell genomes published to date[17,36], and most likely due to the absence of sequences in the amplified DNA (as in the *E. coli* and *S. aureus* data sets (**Supplementary Fig. 1**), as described previously[3,17]).

The SAR324_MDA assembly has sequences in common (84–99% identity) with two SAR324 fosmids, HF0010_10I05 and HF0070_07E19 (ref. 30), and some synteny is evident in alignments (**Supplementary Fig. 11**). The 16S sequences from these two fosmids and SAR324_MDA clustered tightly (**Fig. 3**).

The assembly appears to contain a majority of the genome by other criteria as well: all 20 tRNA types, 17 of the 21 types of tRNA

Table 3 Features of the SAR324_MDA single-cell assembly (EULER+Velvet-SC)

Genome size (bp in assembly)	4.3 Mb
Estimated genome size	4.9–6.4 Mb
Percent GC	43%
No. tRNA genes	20 types
No. tRNA synthetases	17 of 21 types
No. rRNAs	1 each of 5S, 16S, 23S
No. genes	3,811
No. conserved single-copy genes	75/111 (67%)
No. conserved single-copy gene clusters	58/66 (87%)

3,811 genes are those >180 bp in length.

synthetases (including selenocysteine) and full biosynthetic pathways for all amino acids and most vitamins were present (**Table 3** for partial data). Complete glycolytic/gluconeogenesis, tricarboxylic acid and pentose pathways were present, supporting a chemoheterotrophic lifestyle. Many of the components of chemotaxis and flagella synthesis and operation were encoded (**Supplementary Fig. 12**), as were the components of aerobic metabolism (for example, cytochrome *c* oxidase). Putative formate dehydrogenase and carbon monoxide (CO) dehydrogenases within the SAR324 contigs that indicate the potential for anaerobic metabolism were examined in more detail. Phylogenetic analysis shows that the putative formate dehydrogenases and orthologs found in other marine aerobes, including *P. pacifica*, form a clade quite divergent from the biochemically characterized anaerobic formate dehydrogenases (**Supplementary Fig. 13a**). This suggests that they act in an unknown aerobic pathway, expanding the metabolic diversity of this ancient protein family. Similarly, phylogenetic analysis (**Supplementary Fig. 13b**) of SAR324_MDA putative CO dehydrogenases, and similar proteins encoded by two recently sequenced fosmid clones of SAR324 (ref. 30) and the *P. pacifica* genome shows they are divergent from functionally characterized versions. Protein alignment shows these deltaproteobacterial oxidoreductases contain the molybdenum-cofactor (MoCo) binding site but lack CO dehydrogenase consensus sequences[37]. All the deltaproteobacterial putative CO dehydrogenase genomic clusters encode both the MoCo-binding large subunit and the Fe-S–binding small subunit, but not the flavoprotein-binding medium subunit, providing more evidence that they are not CO dehydrogenases. In summary, the more detailed sequence analysis of these proteins is inconsistent with roles in anaerobic metabolism. One of the most striking features of the SAR324 assembly is the presence of 18 putative phytanoyl dioxygenases, which catalyze the degradation of the lipid chain on chlorophyll a. The metabolic features of SAR324, and its dominance in the upper mesopelagic, suggest they track and degrade sinking photosynthetic biomass as it leaves the sunlit surface ocean.

DISCUSSION

A major challenge in single-cell genome assembly is the nonuniformity of coverage, particularly when combined with increased error rates and chimeras. To address this problem, we developed Velvet-SC, a modified version of the Velvet assembler tailored for single-cell data, and applied it with the read error correction algorithm from EULER-SR[23]. By validating the performance of EULER+Velvet-SC on data from single cells of organisms with reference genomes, we showed that our approach successfully copes with the nonuniformity of coverage, incorporating many more bases in the assembly than Velvet (4.57 Mb of contigs in EULER+Velvet-SC vs. 3.53 Mb in Velvet, in the *E. coli* single-cell lane 1 data set) and increasing the quality of the assembly. Although using a lower cutoff initially creates a noisier graph with more contig fragmentation, the iterative cutoff approach used by Velvet-SC overcomes the fragmentation and results in longer contigs. Applying EULER-SR error correction prior to assembly further improves contig size and assembly quality. A test with a novel uncultured organism confirmed that a draft-quality genome sequence can be obtained with minimal effort and reasonable cost (one or one-half of an Illumina lane, with no closure efforts) providing much more information about uncultivated bacteria than can be acquired using traditional metagenomic approaches.

This approach should facilitate the creation of draft assemblies of large numbers of single-cell bacterial genomes. In cases where the organism is of sufficient interest to warrant additional effort, there are several strategies for completing the assemblies[9,17,18]. Mate-pair sequencing can also assist in assembly; however, the presence of chimeric rearrangements (**Supplementary Tables 3** and **5**) occurring at about 1 per 10–30 kbp[9,10] of amplified DNA may limit the useful length of inserts. The optimal use of mate pairs for single-cell sequencing remains to be investigated. The rapid improvement of sequencing technologies and reduction of cost also promises to accelerate progress.

A major goal of single-cell genomics is to complement the large volume of gene-centric metagenomic data with whole-genome assemblies of uncultivated organisms that support the annotation of the majority of the gene complement. This emerging technology will drive studies of uncultured organisms from the human microbiome (including pathogens) and from marine and soil environments (including bacteria producing antibiotics and bacteria with potential for biofuel production). The cost-effective approach demonstrated here should contribute to exploration of microbial taxonomy and evolution and facilitate the mining of environmental organisms for genes and pathways of interest to biotechnology and biomedicine. We also envision further development of EULER+Velvet-SC and applications in metagenomics and transcriptome sequencing projects, which are also characterized by highly nonuniform coverage.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession codes. SAR324_MDA reads were submitted to Short Read Archive under the accession SRA043956. The SAR324_MDA assembly was submitted to Whole Genome Shotgun project and deposited at DDBJ/EMBL/GenBank under the accession AGAU00000000. The version described in this paper is the first version, AGAU01000000. The NCBI BioProject accession ID is PRJNA71321 (Taxonomy ID 1073573) under the organism name "SAR324 cluster bacterium JCVI-SC AAA005". Reads are also available at http://bix.ucsd.edu/singlecell/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This work was partially supported by grants to R.S.L. from the National Human Genome Research Institute (NIH-2 R01 HG003647) and the Alfred P. Sloan Foundation (Sloan Foundation-2007-10-19), and by a grant to P.A.P. and G.T. from the US National Institutes of Health (NIH grant 3P41RR024851-02S1). We thank M. Kim (J. Craig Venter Institute) for bioinformatics support.

AUTHOR CONTRIBUTIONS

All authors analyzed data. H.C. and G.T. wrote software. M.N., J.L.Y.-G., M.-J.L. and L.J.F. performed wet lab experiments. Illumina sequencing was performed at Illumina Cambridge Ltd. O.S.-T. analyzed sequencing data at Illumina. H.C., J.L.Y.-G., G.T., C.L.D., M.-J.L., L.J.F., N.A.G., P.A.P. and R.S.L. wrote the manuscript. H.C., G.T., M.-J.L., C.L.D., J.H.B., D.B.R. and N.A.G. created figures and tables. R.S.L. and M.-J.L. supervised the JCVI group. P.A.P. and G.T. supervised the UCSD group. N.A.G. and D.J.E. supervised the Illumina group. G.P.S. initiated the Illumina-JCVI collaboration.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/nbt/index.html.

Published online at http://www.nature.com/nbt/index.html.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Rusch, D.B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5**, e77 (2007).
2. Gill, S.R. *et al.* Metagenomic analysis of the human distal gut microbiome. *Science* **312**, 1355–1359 (2006).
3. Raghunathan, A. *et al.* Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* **71**, 3342–3347 (2005).

4. Dean, F.B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. USA* **99**, 5261–5266 (2002).
5. Dean, F.B., Nelson, J.R., Giesler, T.L. & Lasken, R.S. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11**, 1095–1099 (2001).
6. Hosono, S. *et al.* Unbiased whole-genome amplification directly from clinical samples. *Genome Res.* **13**, 954–964 (2003).
7. Lasken, R.S. Single cell genomic sequencing using Multiple Displacement Amplification. *Curr. Opin. Microbiol.* **10**, 510–516 (2007).
8. 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).
9. Zhang, K. *et al.* Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* **24**, 680–686 (2006).
10. Lasken, R.S. & Stockwell, T.B. Mechanism of chimera formation during the Multiple Displacement Amplification reaction. *BMC Biotechnol.* **7**, 19 (2007).
11. Lasken, R.S. *et al.* Multiple displacement amplification from single bacterial cells in *Whole Genome Amplification: Methods Express* (eds. Hughes, S. & Lasken, R.) 119–147 (Scion Publishing Ltd., UK, 2005).
12. Kvist, T., Ahring, B.K., Lasken, R.S. & Westermann, P. Specific single-cell isolation and genomic amplification of uncultured microorganisms. *Appl. Microbiol. Biotechnol.* **74**, 926–935 (2007).
13. Mussmann, M. *et al.* Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. *PLoS Biol.* **5**, e230 (2007).
14. 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. USA* **104**, 11889–11894 (2007).
15. Podar, M. *et al.* Targeted access to the genomes of low abundance organisms in complex microbial communities. *Appl. Environ. Microbiol.* **73**, 3205–3214 (2007).
16. Hongoh, Y. *et al.* Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *Proc. Natl. Acad. Sci. USA* **105**, 5555–5560 (2008).
17. Rodrigue, S. *et al.* Whole genome amplification and de novo assembly of single bacterial cells. *PLoS ONE* **4**, e6864 (2009).
18. Woyke, T. *et al.* Assembling the marine metagenome, one cell at a time. *PLoS ONE* **4**, e5299 (2009).
19. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
20. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
21. Pevzner, P.A., Tang, H. & Waterman, M.S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA* **98**, 9748–9753 (2001).
22. Simpson, J.T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
23. Chaisson, M.J. & Pevzner, P.A. Short read fragment assembly of bacterial genomes. *Genome Res.* **18**, 324–330 (2008).
24. 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).
25. Wright, T.D., Vergin, K.L., Boyd, P.W. & Giovannoni, S.J. A novel delta-subdivision proteobacterial lineage from the lower ocean surface layer. *Appl. Environ. Microbiol.* **63**, 1441–1448 (1997).
26. Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* **34**, 5623–5630 (2006).
27. Tatusov, R.L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
28. Goldman, B.S. *et al.* Evolution of sensory complexity recorded in a myxobacterial genome. *Proc. Natl. Acad. Sci. USA* **103**, 15200–15205 (2006).
29. DeLong, E.F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496–503 (2006).
30. 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.* **13**, 116–134 (2010).
31. Yooseph, S. *et al.* Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468**, 60–66 (2010).
32. Iizuka, T. *et al.* Plesiocystis pacifica gen. nov., sp. nov., a marine myxobacterium that contains dihydrogenated menaquinone, isolated from the Pacific coasts of Japan. *Int. J. Syst. Evol. Microbiol.* **53**, 189–195 (2003).
33. Callister, S.J. *et al.* Comparative bacterial proteomics: analysis of the core genome concept. *PLoS ONE* **3**, e1542 (2008).
34. Mitreva, M. Bacterial core gene set. <http://www.hmpdacc.org/doc/sops/reference_genomes/metrics/Bacterial_CoreGenes_SOP.pdf> (2008).
35. Nelson, K.E. *et al.* A catalog of reference genomes from the human microbiome. *Science* **328**, 994–999 (2010).
36. Woyke, T. *et al.* One bacterial cell, one complete genome. *PLoS ONE* **5**, e10314 (2010).
37. King, G.M. Microbial carbon monoxide consumption in salt marsh sediments. *FEMS Microbiol. Ecol.* **59**, 2–9 (2007).
38. Schloss, P.D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
39. Wilgenbusch, J.C. & Swofford, D. Inferring evolutionary trees with PAUP*. *Curr. Prot. Bioinformatics*, Unit 6.4 6.4.1–6.4.28 (2003).
40. Hernandez, D., Francois, P., Farinelli, L., Ostera, M. & Schrenzel, J. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Res.* **18**, 802–809 (2008).
41. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
42. Mao, F., Dam, P., Chou, J., Olman, V. & Xu, Y. DOOR: a database for prokaryotic operons. *Nucleic Acids Res.* **37**, D459–D463 (2009).

ONLINE METHODS

Velvet-SC: modifications to Velvet assembly algorithm. Although Velvet[19], ABySS[22] and EULER-SR[23] generate many correct contigs, they also generate many erroneous regions (caused by errors in reads as well as assembly errors) during intermediate stages of assembly that must be removed in the final assembly. In normal multicell assembly, coverage throughout the genome is fairly uniform, so all these tools use a fixed coverage cutoff to eliminate erroneous contigs. This strategy, however, fails in single-cell assembly as coverage is highly nonuniform (**Supplementary Figs. 1** and **2** and **Supplementary Table 4**), and low-coverage regions can represent correct contigs.

The Velvet-SC (http://bix.ucsd.edu/singlecell/) algorithm is designed to salvage low-coverage regions. We give an informal explanation of Velvet-SC here (for detailed pseudocode, see **Supplementary Fig. 5**). Velvet combines reads using a de Bruijn graph[21]. Vertices of the de Bruijn graph correspond to all k-mers present in reads and edges correspond to all $(k+1)$-mers present in reads. An edge corresponding to a $(k+1)$-mer $a_1a_2 \ldots a_ka_{k+1}$ connects a vertex $a_1a_2 \ldots a_k$ (prefix k-mer) with a vertex $a_2 \ldots a_{k-1}a_{k+1}$ (suffix k-mer). The coverage of an edge ($(k+1)$-mer) in the de Bruijn graph is the number of times this $(k+1)$-mer appears in all reads.

The resulting de Bruijn graph is very complex (even for small bacterial genomes), necessitating a step of removing low-coverage edges. Velvet removes such edges (and entire low-coverage regions) using a fixed threshold; this is a critical assembly step that attempts to remove errors, but it assumes that coverage is uniform across the genome. However, this approach leads to removal of many correct edges with low coverage as such edges are prevalent in many low-coverage genomic regions in single-cell assembly projects. The Velvet-SC algorithm instead uses a variable threshold that starts at 1 and gradually increases. Some contigs may potentially be linked by two possible intermediate linker sequences, one with high coverage and one with low coverage (**Fig. 1**). Velvet-SC removes the low-coverage linker sequence, allowing the neighboring sequences to be merged into a longer contig. This procedure is iterated with a gradually increasing low-coverage cutoff. Because single-cell sequencing results in a mosaic of short low-coverage regions and (typically longer) higher coverage regions, Velvet-SC typically merges low-coverage regions with high-coverage regions (resulting in a region with high coverage), thus rescuing low-coverage regions from elimination.

EULER+Velvet-SC is EULER-SR's error correction[23] combined with Velvet-SC. To test EULER-SR+Velvet-SC, we generated sequencing reads from MDAs done on single cultured cells of *E. coli* K-12 and *S. aureus* USA300. The *E. coli* (lanes 1 and 6) and *S. aureus* data sets are 600×-coverage and 2,300×-coverage, 100-bp, paired-end runs of the Illumina Genome Analyzer IIx pipeline, respectively (~270 bp average insert length for *E. coli* and ~220-bp average insert length for *S. aureus*). A standard unamplified genomic DNA-derived *E. coli* K-12 data set was used as a control (EMBL-EBI Sequence Read Archive, ERA000206, average insert length ~215 bp).

Single-cell isolation. Single cells of *E. coli* (American Type Culture Collection (ATCC) 700926) and *S. aureus* MRSA USA300 strain FPR3757 (ref. 24 and ATCC 25923) were isolated by micromanipulation as described in **Supplementary Methods**. Marine cells were sorted by flow cytometry. A marine water sample from the Scripps Research Pier (Scripps Institute for Oceanography, La Jolla, California, 6 m depth, collected on 8 October 2008, at 9:00 am) was filtered (0.8 μm pore size), flash frozen and stored at –80 °C in 30% glycerol. Before it was sorted, the thawed sample was stained with 10× SYBR Green I nucleic acid stain (Invitrogen). Single cells were sorted using a FACS (fluorescence-activated cell sorting) Aria II flow cytometer (BD Biosciences) equipped with a custom forward scatter (FSC)-PMT using detection by the FSC-PMT and green fluorescence, and at the highest purity setting and a low flow rate to avoid sorting of coincident events. Cells were sorted into 384-well plates containing 4 μl of TE buffer per well, and stored at –80 °C.

MDA and selection of candidate marine amplified DNA. MDA of single-cell genomes was performed using GenomiPhi HY reagents (GE Healthcare) as detailed in **Supplementary Methods**. The 16S rRNA gene was amplified and sequenced (**Supplementary Methods**) and marine MDAs of interest were selected by BLAST analysis of their 16S sequences against a curated marine 16S rRNA database derived from the Global Ocean Sampling 16S data[31]. MDAs with 16S rRNA sequences with >97% identity to operational taxonomic units in the data set were selected for sequencing, including the SAR_324 MDA described here.

Library generation and sequencing. Short insert paired-end libraries were generated from amplified single *E. coli* cell DNA following the standard Illumina protocol[43]. PCR-free paired-end libraries were generated for *S. aureus* and Deltaproteobacteria (to avoid possible uneven representation of AT-rich sequences) from 15 μg of amplified DNA using the adapters:

5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′ and 5′-GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG-3′, and selecting an average insert size of ~250 bp. Sequencing was carried out on a Genome Analyzer IIx using standard reagents. PCR-free libraries were sequenced using the sequencing primer 5′-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3′ in read 2.

Analysis and annotation of the single-cell assembly. Contigs were analyzed by BLAST against a nucleotide sequence database with entries from GenBank and RefSeq (excluding whole genome shotgun assemblies). Annotation of ORFs, tRNAs, rRNA genes and tRNA synthetases was performed using the JCVI metagenomics annotation pipeline[44] without manual curation. Phylogenetic analysis of select proteins was conducted in Bosque[45], with substantial manual creation. Gene identifiers used in KEGG pathway analysis[46] at http://www.genome.jp/kegg/pathway.html were generated at the KEGG Automatic Annotation Server (KAAS[47]) using the bidirectional best hit settings. Of 3,811 MetaGene ORFs submitted to KAAS, 1,415 yielded a gene identifier.

43. Bentley, D.R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
44. Tanenbaum, D.M. *et al.* The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. *Stand. Genomic Sci.* **2**, 229–237 (2010).
45. Ramirez-Flandes, S. & Ulloa, O. Bosque: integrated phylogenetic analysis software. *Bioinformatics* **24**, 2539–2541 (2008).
46. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
47. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.* **35**, W182–185 (2007).

doi:10.1038/nbt.1966