De Novo Assembly of Bacterial Genomes from Single Cells

Hamidreza Chitsaz [1,2], Joyclyn L. Yee-Greenbaum [3,2], Glenn Tesler [4], Mary-Jane Lombardo [3], Christopher L. Dupont [3], Jonathan H. Badger [3], Mark Novotny [3], Douglas B. Rusch[5], Louise J. Fraser [6], Niall A. Gormley [6], Ole Schulz-Trieglaff [6], Geoffrey P. Smith [6], Dirk J. Evers [6], Pavel A. Pevzner [1], Roger S. Lasken [3,7]

[1]University of California, San Diego, Department of Computer Science, La Jolla, CA 92093-0404, USA
[2]These authors contributed equally.
[3]J. Craig Venter Institute, San Diego, CA 92121, USA
[4]University of California, San Diego, Department of Mathematics, La Jolla, CA 92093-0112, USA
[5]J. Craig Venter Institute, Rockville, MD 20855, USA
[6]Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterfield, Nr Saffron Walden, Essex CB10 1Xl, UK
[7]Communicating author

Abstract

Whole genome amplification by the multiple displacement amplification (MDA) method allows sequencing of genomes from single cells of bacteria that cannot be cultured. However, genome assembly is challenging because of highly non-uniform read coverage generated by MDA. We describe an assembly approach tailored for single cell Illumina sequences that generates assemblies comparable to those derived from non-MDA templates. Assembly of *E. coli* and *S. aureus* single cell reads captured >91% of genes within contigs, approaching the 95% captured from a multi-cell *E. coli* assembly. We apply this method to assemble a single cell genome of the uncultivated SAR324 clade of Deltaproteobacteria, a cosmopolitan bacterial lineage in the global ocean. Metabolic reconstruction suggests that SAR324 is aerobic, motile and chemotaxic, with novel molybdenum-containing metalloenzymes and a high capacity for lipid degradation. These new methods enable acquisition of genome assemblies for individual uncultivated bacteria, providing cell-specific genetic information absent from metagenomic studies.

Introduction

A myriad of uncultivated bacteria are found in environments ranging from surface ocean [1] to the human body [2]. Advances in DNA amplification technology have enabled genome sequencing directly from individual cells without requiring growth in culture. These genome-centric culture-independent studies are a powerful complement to gene-centric metagenomics studies.

Genome sequencing requires that the femtograms of DNA present in a single cell be 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 multiple displacement amplification (MDA) [4-6] to prepare the template. MDA is now the preferred method for whole genome amplification from single cells [7, 8]. The first attempt to assemble a complete bacterial genome from one cell [9] further explored the challenges of assembly from MDA DNA, including amplification bias and chimeric DNA rearrangements. 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 generates DNA amplification in MDA [10], but increased sequencing coverage helps to alleviate this problem.

Single cell sequencing methods have enabled investigation of novel uncultured microbes [11-13]. However, while recent studies have continued to improve assemblies [14-18], the full potential of single cell sequencing has not yet been realized. As Rodrigue et al., 2009 emphasize [17], the challenges facing single cell genomics are increasingly computational rather than experimental. 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 non-uniform read coverage. Most existing fragment assembly tools implicitly assume nearly uniform coverage, and most produce erroneous contigs (linking non-contiguous genomic fragments) when the rate of chimeric reads (or chimeric read pairs) exceeds a certain threshold. Thus, there is a need to adapt existing fragment assembly tools for single cell sequencing.

We developed a specialized software tool for assembling sequencing reads from single cell MDAs and applied it to assembly of two known genomes and an unknown marine genome. The draft de novo single cell assemblies, with no efforts to close gaps and resolve repeats, are extremely valuable, identifying the vast majority of genes. Single cell sequencing of *E. coli* and *S. aureus* generated contigs that captured over 91% of genes, and single cell assembly of an uncultured SAR324 marine Deltaproteobacterium allowed identification of a majority of gene functions. The ability to generate high-quality draft assemblies that support annotation of the majority of the gene complement from Illumina reads will drive advances in characterizing uncultured organisms from the human microbiome (including pathogens), and from the environment (including bacteria producing antibiotics and bacteria with potential for biofuel production).

Results

Characteristics of single cell sequences

An average of 93% of reads from two single *E. coli* cells and one *S. aureus* cell mapped to the respective reference genomes (Supplementary Table S1), vs. 99% of the reads in the *E. coli* standard dataset. Nonmapping reads in MDA datasets can often be attributed to minor contaminating sequences [17]. 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 S3). These data are consistent with previous data regarding chimeras in MDA sequence datasets [9, 10].

The single cell datasets display highly nonuniform coverage typical of single cell amplification [3] (Supplementary Figs. S2 and S3; Supplementary Table S4). Single cell sequencing at ~600x depth results in 94 and 50 blackout regions for *E. coli* lane 1 and lane 6, respectively, while sequencing of unamplified DNA results in no blackout regions. 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* ~2300x coverage dataset, 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 vs. 6). As evident with the two *E. coli* datasets, blackout regions can potentially be eliminated by combining reads from multiple single cells when available [3] [13, 16].

Velvet-SC: improved assembly of single cell short reads with highly non-uniform coverage

Non-uniform coverage poses a serious problem for existing *de novo* assembly algorithms. Of de Bruijn based assemblers[21], for example, Velvet and ABySS [22] use an average coverage cutoff threshold for contigs to prune out noisy data, while EULER-SR [23] uses a *k*-mer coverage cutoff. This pruning step significantly reduces the complexity of the underlying de Bruijn graph and makes the algorithms practical. For single cell datasets, however, coverage is highly variable, and a single coverage cutoff prevents assembly of a significant portion of the data, as is evident in the assembly statistics (Table 1). Supplementary Figure S3 plots the percentage of positions with given coverage. In the multicell *E. coli* dataset, most positions in the genome have coverage between 450-800x, and pruning by a coverage threshold helps eliminate erroneous reads; only 0.1% of positions have coverage below 450x. In contrast, in the single cell E. coli dataset (lane 1), 5% of positions have coverage below 10x. To assemble genomic DNA reads from a standard multicell sample, current next-generation sequencing assemblers usually require at least 30x coverage of a region for a successful assembly without gaps.

Velvet-SC (http://bix.ucsd.edu/singlecell/) is a modification of the assembly program Velvet that incorporates lower coverage sequences that most existing assemblers discard, as illustrated in Figure 1 and detailed in Methods.

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

De novo assemblies generated by Velvet, Velvet-SC (modified Velvet), and EULER+Velvet-SC (EULER-SR's error correction followed by Velvet-SC), were compared with those generated by several other assemblers (Table 1 and Supplementary Fig. S4). The metrics compared were the percentage of the genome present in the final assembly (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. EULER+Velvet-SC outperforms Velvet-SC, while Velvet-SC outperforms Velvet. Velvet assembled only 73.1% of the single cell *E. coli* genome (lane 6), with an *N*50 of 18,410 bp and an error rate of 4.1 mismatches per 100 kbp. EULER+Velvet-SC assembled 94.2% of the genome with an *N*50 of 36,581 bp and an error rate of 1.7 mismatches per 100 kbp. For single cell *S. aureus* reads, Velvet assembled 93.8% of the *S. aureus* genome with an *N*50 of 15,800 bp (6.2 mismatches per 100 kb) whereas EULER+Velvet-SC assembled 93.1% of the genome with an *N*50 of 32,296 bp (4.7 mismatches per 100 kb). 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 2 of the 3 plasmids in this *S. aureus* strain [24] (pUSA02, 4439 bp in one contig; pUSA03, 37136 bp in 30 contigs) while Velvet only captured sequences from one plasmid. The EULER+Velvet-SC assembly of *E. coli* had no misassembled contigs, while the assembly of *S. aureus* had one misassembled contig. The ability of the Velvet-SC algorithm to extend contigs into regions of low coverage is illustrated in Supplementary Figure S5.

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 EULER+Velvet-SC (Supplementary Table S6). As expected, for single cell *E. coli* datasets (lanes 1 and 6) increased coverage gives better results, whereas for multicell *E. coli*, the quality of the results does not improve significantly above half the original coverage. However, reasonable assembly quality was achieved with half the dataset for the single cells, suggesting sequencing effort could be diminished by half.

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, CA (see Methods). Single cell MDA reactions were screened by 16S PCR and an uncultured SAR324 Deltaproteobacterium was chosen for testing the de novo assembly methods. Like the reference cells, the SAR324_MDA reads were 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.

Assembly statistics

As expected EULER+Velvet-SC outperforms Velvet and Velvet-SC in single-cell assembly of the uncultured Deltaproteobacterium (Table 2), with increased *N*50 and decreased total number of contigs. The ability of the assemblies to support ORF prediction was tested using MetaGene [25], 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 increased bp incorporation and the EULER-SR error correction reduce spurious ORF calls. The EULER+Velvet-SC ORFs were of higher quality as evidenced by increased numbers of ORFs with taxonomic affiliations identified using BLAST and phylogenetic analysis via the Automated Phylogenetic Inference System (APIS, see Methods), by increased numbers of ORFs corresponding to orthologous genes in the COG database [26], and by increased numbers of single copy conserved genes detected. By all these criteria, EULER+Velvet-SC yields 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, as discussed in [17]. Contaminants can potentially be identified from reads or contigs by analysis of GC content, nucleotide frequencies, and BLAST analysis vs. reference bacterial genomes. The purity of the SAR324_MDA assembly was assessed as follows (see Methods for details). 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 is consistent with a single genome being present (Supplementary Fig. S7). 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).

APIS was used to assess the purity of the genome. APIS attempts to generate a phylogenetic tree for each ORF based on BLAST analysis against reference genomes (see Supplementary material). Contaminating contigs might be revealed as having ORFS with phylogeny distinct from the rest of the assembly. Interestingly, the contigs tend to consist of ORFs with a variety of phylogenies, although putative operons with shared phylogeny are present, as expected. While clustering of orfs with phylogenies distinct from the rest of a genome can be indicative of horizontal gene transfer, in this case, we suggest the variety of phylogenies may be due to divergence of the SAR324 genome from available reference genomes. Alternatively, it could be a characteristic of Deltaproteobacterial genomes, as similarly varied ORF phylogenies have been observed for other Deltaproteobacteria [27] (and J.H. Badger, unpublished).

SAR324_MDA Deltaproteobacterium genome: insights from single cell sequencing

Phylogenetic analysis of 16S sequences (Fig. 2) revealed that this organism is a member of the deeply branched and divergent clade of uncultured deltaproteobacteria designated SAR324 [28]. 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 30-40 Kb fosmids for SAR324 are available (representative 16S fosmid sequences from [30] included in Fig. 2), 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 by [31], is an obligate aerobe with a chemoheterotrophic lifestyle [32], and phenotypic characteristics of myxobacteria. However, the significant 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 includes about 4.3 Mb of non-redundant contigs yielding 3811 ORFs (Table 3). We searched the ORFs for two sets of conserved genes typically found in 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 significant insight into SAR324.

The SAR324_MDA assembly has sequences in common (84-99% identity) with two SAR324 fosmids, HF0010_10I05 and HF0070_07E19 [30], and some synteny is evident in alignments (Supplementary Fig. S8). The 16S sequences from these two fosmids and SAR324_MDA cluster tightly (Fig. 2).

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 synthetases (including selenocysteine), and full biosynthetic pathways for all amino-acids and most vitamins are present (see Table 3 for partial data). Complete glycolytic/gluconeogenesis, tricarboxylic acid, and pentose pathways are present, supporting a chemoheterotrophic lifestyle. Many of the components of chemotaxis and flagella synthesis and operation are encoded (Supplementary Fig. S9), as are the components of aerobic metabolism (e.g., cytochrome c oxidase). The presence of putative formate dehydrogenase and carbon monoxide (CO) dehydrogenases within the SAR324 contigs indicates the potential for anaerobic metabolism, thus these enzymes were examined in more detail. A phylogenetic analysis shows that orthologs of the putative formate dehydrogenases are found in other marine aerobes, including *P. pacifica*, and that these proteins form a clade quite divergent from the biochemically-characterized anaerobic formate dehydrogenases (Fig. 3a). This suggests that they act in an unknown aerobic pathway, expanding the metabolic diversity of this ancient protein family. Similarly, a phylogenetic analysis shows that the SAR324 putative CO dehydrogenases are divergent from functionally characterized versions (Fig. 3b), but similar proteins are encoded by two recently sequenced fosmid clones of SAR324 [30] and by the *P. pacifica* genome (Fig. 3b). Protein alignment shows these deltaproteobacterial oxidoreductases contain the Mo-cofactor (MoCo) binding site but lack CO dehydrogenase consensus sequences [36]. 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 support 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 eighteen 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 main challenge for single cell genome assembly is the non-uniformity of coverage, particularly when combined with increased error rates and chimeras. EULER-SR's error correction algorithm [21] was employed to correct read errors prior to Velvet assembly with Velvet-SC, a modified Velvet assembler tailored for single cell data. Validation of de novo assembly by EULER+Velvet-SC with single cells from reference genomes shows that EULER+Velvet-SC successfully copes with the non-uniformity of coverage, incorporating significantly more bases in the assembly than Velvet, 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. The addition of EULER-SR error correction upstream of assembly further enhances contig size and assembly quality.

A test with a novel uncultured organism confirms that a useful genomic draft can be obtained from a single lane of Illumina paired end sequencing reads. The SAR324 single cell genome provides an excellent example of an assembly obtained with minimal effort and reasonable cost (1 or ½ of an Illumina lane, with no closure efforts), that vastly exceeds the information about genomes of uncultivated bacteria that can be extracted via traditional metagenomic approaches. This approach will enable draft assemblies of large numbers of single cell bacterial genomes at affordable cost. Where the organism is of sufficient interest to warrant additional effort, several publications have investigated strategies to approach completion of assemblies [9, 17, 18]. Mate pair sequencing can also assist in assembly; however, the presence of chimeric rearrangements occurring at about one per 10-30 kb [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 level metagenomic data with genome level assemblies and to apply this emerging technology to study uncultured organisms from various environments including marine, soil, and the human microbiome. 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.

Methods

Velvet-SC: Modifications to Velvet assembly algorithm

While 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) at the intermediate stages of assembly that must be removed in the final assembly. In normal multicell assembly, coverage throughout the entire 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 since coverage is highly nonuniform (see Supplementary Figs. S2 and S3; Supplementary Table S4), 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 here (for detailed pseudocode, see Supplementary Fig. S1). Velvet combines reads using a "de Bruijn graph" [21]. Then it removes regions that have low average coverage using a fixed threshold; this is a critical step that attempts to remove errors, but it assumes that coverage is uniform across the genome. 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; see Figure 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. Since 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 [37] combined with Velvet-SC. To test EULER-SR+Velvet-SC, sequencing reads were generated from MDAs performed on single cultured cells of *E. coli* K-12 and *S. aureus* USA300. The *E. coli* (lane 1 and lane 6) and *S. aureus* datasets are 600x-coverage and 2300x-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 dataset was used as a control (EMBL-EBI Sequence Read Archive, ERA000206, average insert length ~215 bp).

Single cell isolation

Single cells of *Escherichia coli* (ATCC 700926) and *Staphylococcus aureus* MRSA USA300 strain FPR3757 ([24] ATCC 25923) were isolated by micromanipulation as described in Supplementary material. Marine cells were sorted by flow cytometry. A marine water sample from the Scripps Research Pier (Scripps Institute for Oceanography, La Jolla, CA, 6 m depth, collected on October 8, 2008 at 9:00 am) was filtered (0.8 µm pore size), flash frozen and stored at -80°C in 30% glycerol. Prior to sorting, the thawed sample was stained with 10x SYBR Green I nucleic acid stain (Invitrogen). Single cells were sorted using a FACS 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.

Multiple Displacement Amplification (MDA) and selection of candidate marine amplified DNAs

MDA of single cell genomes was performed using GenomiPhi HY reagents (GE Healthcare) as detailed in Supplementary material. 16S rRNA gene was amplifed and sequenced (see Supplementary material) 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 (GOS) 16S data[31]. MDAs with 16S rRNA sequences with > 97% identity to operational taxonomic units in the dataset 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[38]. 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). Contigs will be submitted to GenBank prior to publication. Annotation of ORFs, tRNAs, rRNA genes and tRNA synthetases was performed using the JCVI metagenomics annotation pipeline[39] without manual curation. Phylogenetic analysis of select proteins was conducted in Bosque[40], with substantial manual creation. Gene identifiers used in KEGG pathway analysis[41] at http://www.genome.jp/kegg/pathway.html were generated at the KEGG Automatic Annotation Server (KAAS, [42]) using the bidirectional best hit settings. Of 3811 MetaGene ORFs submitted to KAAS, 1415 yielded a gene identifier.

Acknowledgements

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 National Institutes of Health (NIH grant 3P41RR024851-02S1). We thank Maria Kim (JCVI) for bioinformatics support.

Author contributions

All authors analyzed data.

H.C. and G.T. wrote software.

M.N., J.Y.-G, M.-J.L., and L.J.F. performed wetlab experiments.

H.C., J.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.

Figure legends

Figure 1. Assembling single cell reads using Velvet-SC. (**a**) Coverage varies widely along the genome, between 1 and 12 in this cartoon example. Reads (short lines) and potential contigs (thick lines; boxes around the supporting reads) 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 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, 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. The other contigs are combined into a single contig (**b**) of average coverage 9. The purple region is salvaged by Velvet-SC because it was absorbed into a higher coverage region faster than the variable 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**). Each read is traced by colored lines alongside its vertices. The C/T mismatch results in two alternative paths, both with 5 intermediate vertices (since we used 5-mers). 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.

Figure 2. 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 [30]). The sequences were aligned using MOTHUR [43]. The tree was inferred using the nucleotide maximum likelihood feature of PAUP* 4.0b10 [44]. Branches drawn in thick lines are clades with bootstrap support of 75% or greater. Sequences present on fosmids with extensive nucleotide similarity to the SAR324_MDA assembly are indicated (Fosmid-Assembly comparisons), as is a SAR324 fosmid encoding CoxL homologs that were also present in the SAR324_MDA assembly (see Fig. 3).

Figure 3. 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 2. Numbered suffixes indicate different homologs from the source sequence.

Table 1. Comparison of assembly results on known genomes (for contigs >110 bp): number of contigs, genome *N*50, the length of the largest contig, total nucleotides in the assembly, substitution error rate in the correctly assembled contigs (per 100 kbp), the number of genes completely or partially present in the assembly, and the 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 [45] was run with a minimum overlap of 55. SOAPdenovo 1.0.4 [46] 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 were from http://csbl1.bmb.uga.edu/OperonDB/displayNC.php?id=215 [47] (*E. coli*). nd, not done.

Dataset	Assembler	# contigs	N50 (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	1344	26662	**140518**	4369634	16.1	4324	3178	627	884	553	248
	Edena	1592	3919	44031	3996911	**2.6**		2425	1112		317	444
	SOAPdenovo	1240	18468	87533	4237595	98.3		3021	612		520	248
	Velvet	**428**	22648	132865	3533351	3.0		3055	**170**		584	**106**
	Velvet-SC	872	19791	121367	**4589603**	4.4		3617	325		643	184
	E + V-SC	501	**32051**	132865	4570583	2.7		**3753**	185		**713**	109
E. coli lane 6	EULER-SR	1820	29551	170385	4469152	16.6		3339	734		561	283
	Edena	1536	4899	42342	4147566	3.2		2705	1075		368	428
	SOAPdenovo	1397	20319	**204730**	4576388	48.3		3353	646		586	244
	Velvet	522	18410	168533	3753818	4.1		3131	253		566	145
	Velvet-SC	945	27113	144462	**4688759**	3.8		3779	409		694	161
	E + V-SC	**481**	**36581**	173901	4668135	**1.7**		**3943**	**158**		**749**	**101**
E. coli lane normal	EULER-SR	295	**110153**	**221409**	4598020	3.5		4119	115		788	80
	Edena	1673	3814	20470	**4611645**	3.8		3019	1189		317	538
	SOAPdenovo	**192**	62512	172567	4529677	26.8		**4128**	81		802	**53**
	Velvet	408	31503	129378	4569225	1.6		4061	139		760	108
	Velvet-SC	350	52522	166115	4571760	**1.1**		4121	157		804	58
	E + V-SC	339	54856	166115	4571406	1.5		4124	**66**		**808**	57
S. aureus	EULER-SR	4398	7247	66549	**3376776**	53.1	2622	1958	640	-	nd	nd
	Edena	1288	1881	37770	2358911	**3.0**		1222	925			
	SOAPdenovo	2470	5385	37397	3273188	42.9		482	1740			
	Velvet	625	15800	67677	2807042	6.2		2244	268			
	Velvet-SC	1084	20163	76884	3001635	4.2		2100	458			
	E + V-SC	**355**	**32296**	**107657**	2962136	4.7		**2408**	**173**			

Table 2. Comparison of Velvet-based assembler results (*k*=55) on SAR324_MDA assembly: total number of contigs; assembly N50 (for contigs >110 bp); the 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 [25]; number of ORFs with phylogenetic assignments by APIS (see Methods); number of ORFs with COGs identified via BLAST (see Methods); and number of 111 conserved single copy genes present [33]. N50 is defined as the contig length such that using the same length or longer contigs produces half of the total assembly length.

Assembler	**# of contigs**	**N50 (bp)**	**Largest (bp)**	**Total (bp)**	**# ORFs (MetaGene)**	**# ORFs (APIS)**	**# COGs**	**# Conserved single copy genes**
Velvet	1856	11531	100589	3921396	4575	2462	2160	55/111 (46%)
Velvet-SC	933	23230	113282	4284882	4234	2627	2307	75/111 (67%)
E +V-SC	823	30293	113282	4282110	4154	2604	2281	75/111 (67%)

Table 3. Features of the SAR324_MDA single cell assembly (EULER + Velvet-SC). 3811 genes are those >180 bp in length.

Genome size (bp in assembly)	4.3 Mb
Estimated genome size	4.9-6.4 Mb
% GC	43%
# tRNA genes	20 types
# tRNA synthetases	17 of 21 types
# rRNAs	1 each of 5S, 16S, 23S
# genes	3811
# conserved single copy genes	75/111 (67%)
# conserved single copy gene clusters	58/66 (87%)

References

1. Rusch, D.B. et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **13;5(3):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.* **10:1-7** (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.*, 19 (2007).
11. Lasken, R.S. et al. in Whole Genome Amplification: Methods Express. (eds. S. Hughes & R. Lasken) pp. 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(4)**, 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 U S A* **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(10)**, 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 U S A* **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., Brinza, D. & Pevzner, P.A. De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res.* **19**, 336-346 (2009).
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. Noguchi, H., Park, J. & Takagi, T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* **34**, 5623-5630 (2006).
26. Tatusov, R.L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
27. Goldman, B.S. et al. Evolution of sensory complexity recorded in a myxobacterial genome. *Proc. Natl. Acad. Sci. USA* **103**, 15200-15205 (2006).
28. 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).
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.* (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.
35. Consortium, H.M.J.R.S. et al. A catalog of reference genomes from the human microbiome. *Science* **328**, 994-999 (2010).
36. King, G.M. Microbial carbon monoxide consumption in salt marsh sediments. *FEMS Microbiol Ecol* **59**, 2-9 (2007).
37. Chaisson, M.J. & Pevzner, P.A. Short read fragment assembly of bacterial genomes. *Genome Res.* **18**, 324-330 (2008).
38. Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).
39. Tanenbaum, D.M. et al. The JCVI standard operating procedure for annotating prokaryotic metagenomic shotgun sequencing data. *S.I.G.S.* **2** (2010).
40. Ramirez-Flandes, S. & Ulloa, O. Bosque: integrated phylogenetic analysis software. *Bioinformatics* **24**, 2539-2541 (2008).
41. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27-30 (2000).
42. 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).
43. 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).
44. Wilgenbusch, J.C. & Swofford, D. Inferring evolutionary trees with PAUP*. *Curr Protoc Bioinformatics* **Chapter 6**, Unit 6 4 (2003).
45. 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).
46. Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265-272 (2010).

47. Mao, F., Dam, P., Chou, J., Olman, V. & Xu, Y. DOOR: a database for prokaryotic operons. *Nucleic Acids Res* **37**, D459-463 (2009).

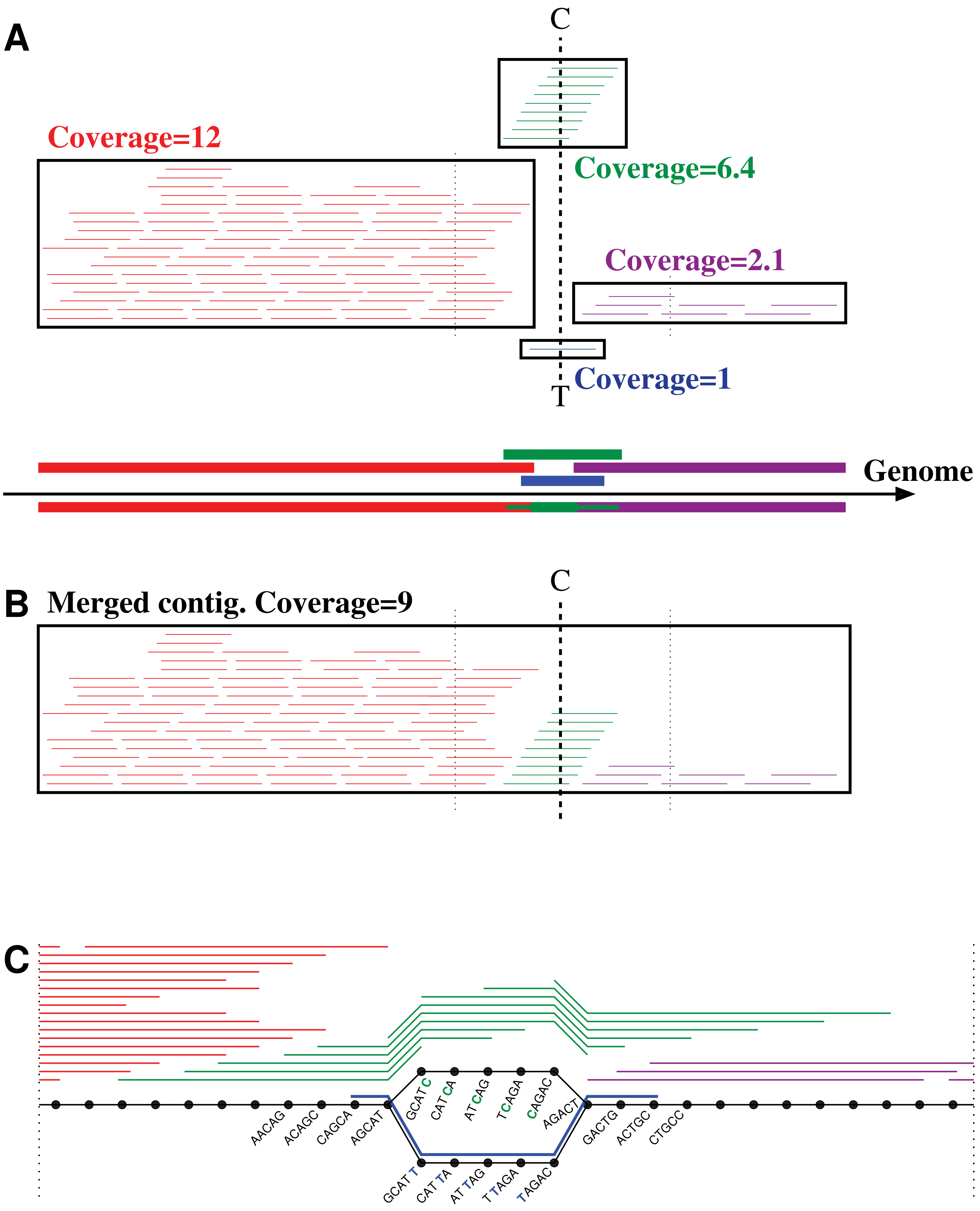

A
C
Coverage=12
Coverage=6.4
Coverage=2.1
Coverage=1
T
Genome
B
Merged contig. Coverage=9
C
C
AACAG
ACAGC
CAGCA
AGCAT
GCATC
CATCA
ATCAG
TCAGA
CAGAC
AGACT
GACTG
ACTGC
CTGCC
GCATT
CATTA
ATTAG
TTAGA
TAGAC

Figure-2 Lasken

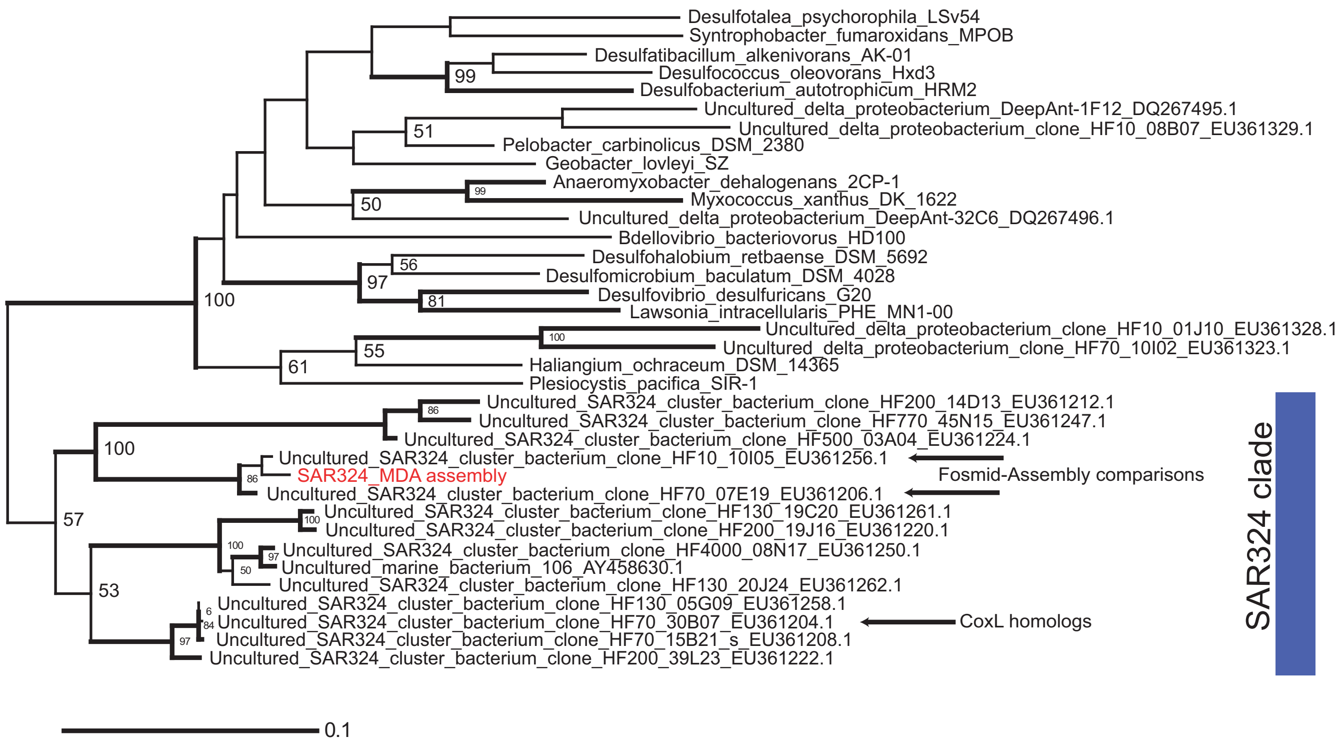
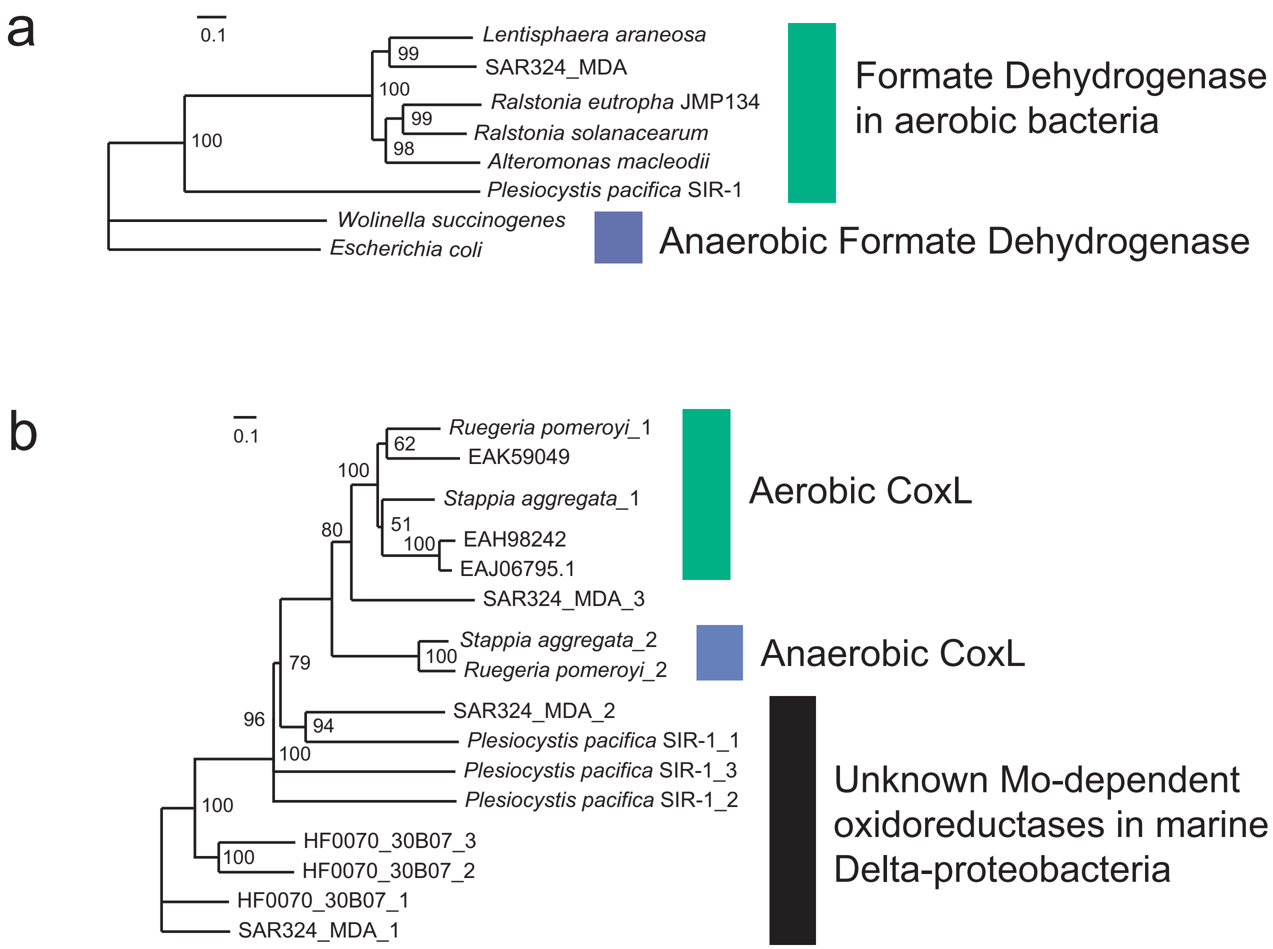

Desulfotalea_psychorophila_LSv54
Syntrophobacter_fumaroxidans_MPOB
Desulfatibacillum_alkenivorans_AK-01
Desulfococcus_oleovorans_Hxd3
99
Desulfobacterium_autotrophicum_HRM2
Uncultured_delta_proteobacterium_DeepAnt-1F12_DQ267495.1
Uncultured_delta_proteobacterium_clone_HF10_08B07_EU361329.1
51
Pelobacter_carbinolicus_DSM_2380
Geobacter_lovleyi_SZ
Anaerомyxobacter_dehalogenans_2CP-1
99
50
Myxococcus_xanthus_DK_1622
Uncultured_delta_proteobacterium_DeepAnt-32C6_DQ267496.1
Bdellovibrio_bacteriovorus_HD100
Desulfohalobium_retbaense_DSM_5692
56
97
Desulfomicrobium_baculatum_DSM_4028
100
Desulfovibrio_desulfuricans_G20
81
Lawsonia_intracellularis_PHE_MN1-00
Uncultured_delta_proteobacterium_clone_HF10_01J10_EU361328.1
100
55
Uncultured_delta_proteobacterium_clone_HF70_10I02_EU361323.1
61
Haliangium_ochraceum_DSM_14365
Plesiocystis_pacifica_SIR-1
Uncultured_SAR324_cluster_bacterium_clone_HF200_14D13_EU361212.1
86
Uncultured_SAR324_cluster_bacterium_clone_HF770_45N15_EU361247.1
Uncultured_SAR324_cluster_bacterium_clone_HF500_03A04_EU361224.1
100
Uncultured_SAR324_cluster_bacterium_clone_HF10_10I05_EU361256.1
SAR324_MDA assembly
86
Fosmid-Assembly comparisons
Uncultured_SAR324_cluster_bacterium_clone_HF70_07E19_EU361206.1
57
Uncultured_SAR324_cluster_bacterium_clone_HF130_19C20_EU361261.1
100
Uncultured_SAR324_cluster_bacterium_clone_HF200_19J16_EU361220.1
100
Uncultured_SAR324_cluster_bacterium_clone_HF4000_08N17_EU361250.1
97
50
Uncultured_marine_bacterium_106_AY458630.1
53
Uncultured_SAR324_cluster_bacterium_clone_HF130_20J24_EU361262.1
6
Uncultured_SAR324_cluster_bacterium_clone_HF130_05G09_EU361258.1
84
Uncultured_SAR324_cluster_bacterium_clone_HF70_30B07_EU361204.1
CoxL homologs
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 Lasken


a
0.1
Lentisphaera araneosa
SAR324_MDA
Ralstonia eutropha JMP134
Ralstonia solanacearum
Alteromonas macleodii
Plesiocystis pacifica SIR-1
Wolinella succinogenes
Escherichia coli
99
100
99
98
100
🟩 Formate Dehydrogenase in aerobic bacteria
🟦 Anaerobic Formate Dehydrogenase
b
0.1
Ruegeria pomeroyi_1
EAK59049
Stappia aggregata_1
EAH98242
EAJ06795.1
SAR324_MDA_3
Stappia aggregata_2
Ruegeria pomeroyi_2
SAR324_MDA_2
Plesiocystis pacifica SIR-1_1
Plesiocystis pacifica SIR-1_3
Plesiocystis pacifica SIR-1_2
HF0070_30B07_3
HF0070_30B07_2
HF0070_30B07_1
SAR324_MDA_1
62
100
51
100
80
100
79
94
96
100
100
100
🟩 Aerobic CoxL
🟦 Anaerobic CoxL
⬛ Unknown Mo-dependent oxidoreductases in marine Delta-proteobacteria