A peer-reviewed version of this preprint was published in PeerJ on 8 October 2015.

View the peer-reviewed version (peerj.com/articles/1319), which is the preferred citable publication unless you specifically need to cite this preprint.

Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319 https://doi.org/10.7717/peerj.1319

Anvi'o: An advanced analysis and visualization platform for 'omics data

A. Murat Eren[1*], Özcan C. Esen[1], Christopher Quince[2], Joseph H. Vineis[1], Mitchell L. Sogin[1] and Tom O. Delmont[1]

[*]Correspondence:
a.murat.eren@gmail.com

[1]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, 02543 MA, USA

Full list of author information is available at the end of the article

Abstract

Comprehensive analysis of shotgun metagenomic assemblies have revolutionized molecular microbial ecology, but few microbiologists command the full suite of bioinformatics skills necessary to process, interact, organize and visualize overlapping DNA sequence contigs. Here we introduce anvi'o, an advanced analysis and visualization platform for 'omics data, and its assembly-based metagenomic workflow. Anvi'o's interactive interface facilitates the management of contigs and associated metadata for automatic or human-guided identification of genome bins, and their curation. Its extensible visualization approach distills multiple dimensions of information about each contig into a single, intuitive display, offering a dynamic and unified work environment for data exploration, manipulation and reporting. Beyond its easy-to-use interface, the advanced modular architecture of anvi'o as a platform allows users with programming skills to implement and test novel ideas with minimal effort. To demonstrate anvi'o's capabilities, we re-analyzed a metagenomic time-series data from an infant gut microbiome. Through the anvi'o interface we identified near-complete draft genomes, and explored temporal genomic changes within the abundant microbial populations through de novo characterization of subtle nucleotide variations. We also used anvi'o to re-analyze a collection of datasets from multiple investigators who studied microbial responses to the Deepwater Horizon oil spill. We linked metagenomic, metatranscriptomic, and single-cell genomic data from the water plume, and used the holistic perspective anvi'o provides to identify the draft genome of a previously uncharacterized, active population of *Oceanospirillales*. We also linked environmental isolates with metagenomes recovered from an oil-contaminated beach, and identified 56 near-complete draft genomes including abundant oil degraders whose functional features suggested an oceanic origin.

Keywords: metagenomics; assembly; genome binning; visualization; SNP profiling

Background

High-throughput sequencing of the environmental DNA is rapidly becoming one of the most effective ways to study naturally occurring microbial communities. By circumventing the need for cultivation, shotgun metagenomics, the direct extraction and sequencing of DNA fragments from a sample, provides access to the enormous

pool of microbial diversity marker gene surveys have unveiled [1, 2]. Early studies using capillary sequencing techniques [3], and more recently massively-parallel techniques [4, 5], led to descriptions of microbial-mediated activities and their functional interactions, which have provided novel insights into medicine [6], biotechnology [7], and evolution [8].

Current high-throughput sequencing technologies generate an astonishing amount of sequence data, although, the length of highly accurate DNA sequence reads falls short of bacterial genome sizes by orders of magnitude. Multiple online resources can annotate metagenomic short reads [9, 10], however, their relatively small information content compared to the length of coding regions, constrains accurate functional inferences [11, 12]. Despite these limitations, researchers have successfully used metagenomic short reads to investigate and compare the functional potential of various environments [13, 14, 15].

The assembly of short reads into contiguous DNA segments (contigs) leads to improved annotations because of the greater information content of longer sequences including the genomic context of multiple coding regions. Multiple factors affect the assembly performance [16, 17, 18], and the feasibility of the assembly-based approaches vary across environments [19, 20]. Nevertheless, increasing read lengths [21], novel experimental approaches [22], advances in computational tools [23], and improvements in assembly algorithms and pipelines [24, 25, 26, 27] continue to make the assembly-based metagenomic workflow more accessible. Additional improvements emerge from genomic binning techniques that employ contextual information to organize unconnected contigs into biologically relevant units, i.e. draft genomes, plasmids, and phages [3, 28]. Draft genomes frequently provide deeper insights into bacterial lifestyles that would otherwise remain unknown [29, 30, 31], and offers an opportunity to identify single-nucleotide polymorphisms that define differences between members or strains in a microbial population [28]. Genome binning processes typically take advantage of sequence composition, and the abundance of contigs across multiple samples. Despite challenges associated with this process [17, 32], researchers have successfully employed these assembly and binning techniques to identify near-complete novel draft genomes from metagenomic datasets generated from various environments [3, 28, 33, 34]. This workflow has become more accessible thanks to the recently introduced human-guided [19, 35] and automatic [36, 37, 38] approaches and software pipelines that lend themselves to the identification of genome bins.

Beyond these advancements, comprehensive analysis of assembled metagenomic data requires the ability to interact, and mine complex datasets within a visualization framework that immediately reports the end result of data manipulations. Available tools for the visualization of metagenomic contigs usually employ self-organizing maps [19], or principal component analysis plots [36, 39]. Although these visualization strategies can describe the organization of contigs, they do not present the distribution of contigs across samples along with the supporting metadata such as the GC-content, inferred taxonomy, or other automatically generated or user-specified information for each contig in one display. Interactive visualization tools

that report the influence of contextual information on the supervised binning of contigs, and that provide the ability to modify the membership of contigs in genome bins would improve the inference of high-quality draft genomes. A platform that consolidates advanced visualization and analysis infrastructure with an open design that allows the addition of novel algorithms could serve as a test bed for sharing new analytical paradigms, and contribute to the dissemination of good practices in the field of metagenomics.

Here we introduce anvi'o, an advanced analysis and visualization platform for 'omics data, and describe its assembly-based metagenomic workflow, which includes supervised and unsupervised metagenomic binning, interactive data exploration, manipulation, visualization, and reporting. To demonstrate anvi'o as a platform we re-analyzed a relatively low-complexity dataset of daily metagenomic sampling of an infant gut microbiome [19], and a collection of datasets that represent the combined efforts of multiple investigators [40, 41, 42, 43, 44], who studied the microbial response to the 2010 Deepwater Horizon (DWH) oil spill [45].

Methods

Anvi'o is an analysis and visualization platform for 'omics data. It provides an interactive and extensible visualization interface that distills multiple dimensions of information into a single, intuitive display. The platform is written predominantly in Python, JavaScript, and C, and relies on scalable vector graphics (SVG) for most visualization tasks. The visualization core, implemented from scratch in JavaScript, uses low-level SVG object manipulation functions with minimal overhead to optimize performance. Anvi'o visualizes tree structures with data or metadata layers that describe properties of each leaf on the tree. The platform stores computed data into self-contained database files that can be interrogated using structured query language (SQL) through SQLite, an open source transactional SQL database engine that does not require any database server or configuration. The user interacts with anvi'o through command line clients or a graphical web browser. The platform generates static HTML web pages to summarize analysis results. Reliance on self-contained database files and static HTML output facilitates transfer of intermediate or final stages of analyses between computers. In this study we emphasize anvi'o's metagenomic workflow, but the platform can also meet the analysis and visualization requirements of other 'omics data types. Anvi'o is a community-driven, open-source project. The source code is licensed under the GNU General Public License, and publicly available at http://merenlab.org/projects/anvio.

Anvi'o metagenomics workflow

Preparing a metagenomic dataset for an analysis with Anvi'o requires a co-assembly step of all short reads from all or a subset of samples to create community contigs, followed by the mapping of short reads from individual samples back to these contigs. A FASTA file for community contigs, and a BAM file for each sample that reports mapping results provides the initial input for anvi'o. The BAM file format is the binary representation of the Sequence Alignment/Map (SAM) format [46],

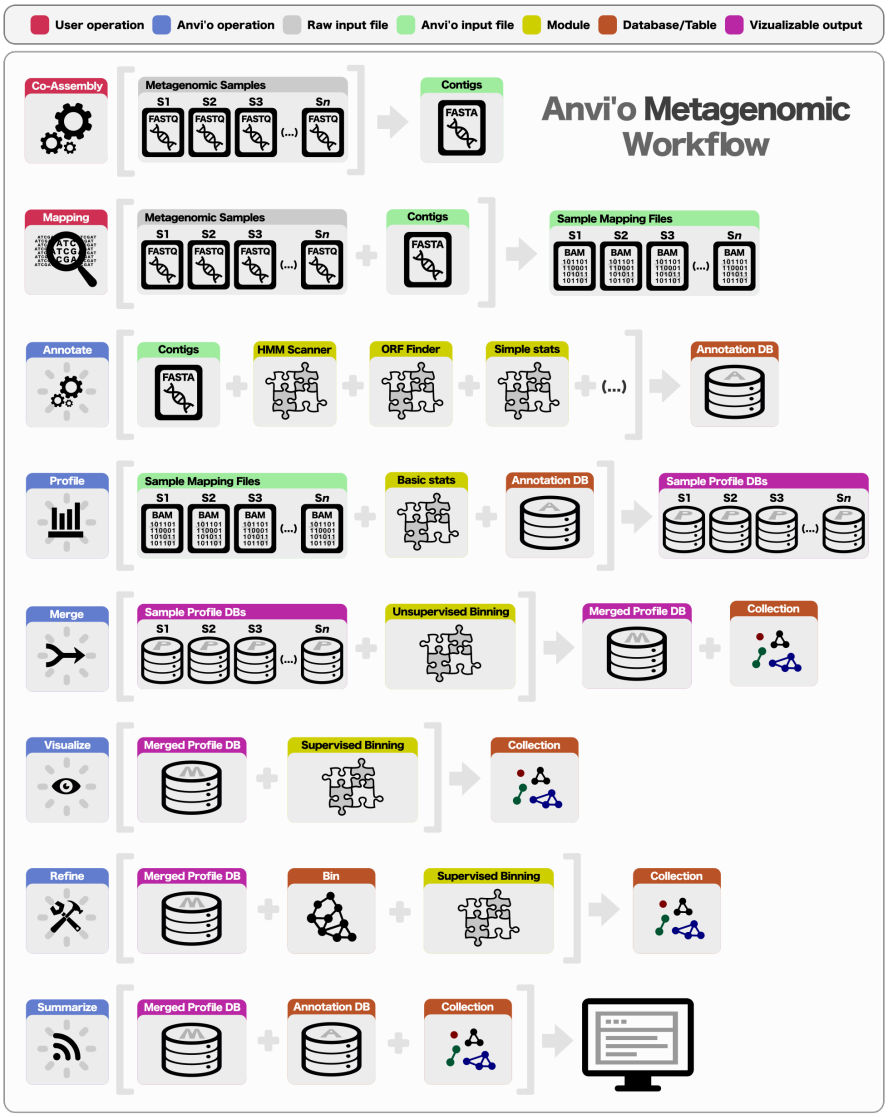

Figure 1 Overview of the anvi'o metagenomic workflow. Anvi'o can perform comprehensive analysis of BAM files following the initial steps of co-assembly and mapping. Annotation of contigs, and profiling each BAM file individually, generate all the essential databases anvi'o uses throughout the downstream processing. Anvi'o can merge single profile databases, during which the unsupervised binning module would exploit the differential distribution patterns of contigs across samples to identify genome bins automatically, and store binning results as a collection. The optional visualization step gives the user the opportunity to interactively work with the data, and perform supervised binning with real-time completion and redundancy estimates based on the presence or absence of bacterial single-copy genes. The user can screen and refine genome bins, and split a single mixed genome bin into multiple bins with low redundancy estimates. Finally, the user can summarize collections that describe genome bins, which would create a static web site that would contain necessary information to review each genome bin, and to analyze their occurrence across samples.

which is the standard output for most widely used mapping software, including BWA [47], Bowtie2 [48], and CLC Genomics Workbench (http://www.clcbio.com). Subsequent to the generation of BAM files, a typical analysis of multiple metagenomic samples with anvi'o entails the following steps (Figure 1): (1) generating an

annotation database, (2) profiling each sample individually, and merging resulting single profiles, (3) visualizing results interactively, performing human-guided binning, or refining automatically identified bins, and (4) summarizing results.

Annotation database. Anvi'o maintains an annotation database that stores the shared information for contigs (or scaffolds) that does not change from sample to sample (i.e., k-mer frequencies of contigs, functional annotation of open reading frames, or GC-content). One or more splits (depending upon split size) break up long contigs, but the workflow soft-links splits to contigs concatenated in the correct order for result summaries. The user can override the default split size of 20,000 bases while creating the annotation database. When the user creates an annotation database for a given FASTA file of contigs, anvi'o identifies splits, computes k-mer frequency tables for each contig and for each split separately. Optionally, anvi'o can identify open reading frames (ORFs), process functional and taxonomical annotations for ORFs, and search contigs for hidden Markov model (HMM) profiles to be stored in the annotation database for later use. Currently anvi'o installs four previously published HMM profiles for bacterial single-copy gene collections [36, 49, 50, 51]. Presence or absence of these genes in contigs provides a metric for estimating the level of completeness of genome bins during the interactive supervised binning (see 'Binning'). The system also generates completion and redundancy (multiple occurrence of one or more single-copy genes in a bin) statistics in real-time to guide the supervised binning processes. Beyond single-copy genes, users can further populate the annotation database with their curated HMM profiles to identify presence of protein families of interest. The annotation database also stores inferred functions and taxonomy for all recognized open reading frames. Users have the option to provide this data as a standard matrix file, or they can use one of the pre-existing parsers. The initial version supports annotation files generated by the RAST annotation server [52], but the design allows inclusion of annotations from other sources.

Profile database. In contrast to the annotation database, an anvi'o profile database stores sample-specific information about contigs (such as the mean coverage values of each contig in a given sample based on the alignment of short reads reported in the corresponding BAM file). Profiling a BAM file creates a single profile that reports properties of each contig in a single sample. Every profile database must link to an annotation database, and anvi'o can merge single profiles that link to the same annotation database into merged profiles. The structure of single and merged profiles differs slightly: when multiple single profiles merge, each attribute reported in the data table for each single profile merges into their own table in the merged-profile database. For instance, the 'mean coverage column' in the data table of single-profile databases generated for sample A and sample B, would become the 'mean coverage table' with sample A and sample B as columns when merged. Anvi'o identifies these merged tables as views, and the user can switch between views in the interactive interface. The profiler computes multiple attributes for each contig, including mean coverage (number of short reads mapping to a contig divided by the length of the contig), portion covered (percentage of nucleotide positions in a given contig covered by at least one short read), and variability (number of

variable nucleotide positions in a contig). The profiler can accommodate new attributes produced by any algorithm that yields a numerical value for a given contig. During the merging step, anvi'o combines each 'attribute' from single profiles, and stores them as a 'view' in the resulting merged profile, which becomes available for the visualization and/or clustering tasks across the platform. This modularity fosters the quick implementation of new binning strategies and evaluation of results without requiring changes in the code. Profile databases also store other essential information such as frequencies of nucleotides at variable positions (see 'Computing variability'), and contig collections (see 'Binning').

De novo characterization of nucleotide variation within samples. The alignment of short reads that map to a particular contig can generate one or more mismatches. The source of a mismatch may be artificial, such as stochastic sequencing or PCR errors, however, some mismatches may represent ecologically informative variation. During the profiling step, anvi'o keeps track of nucleotide variation (base frequencies) among reads from each sample that map to the same community contig and stores that information in the profile database for each sample. To lessen the impact of sequencing and mapping errors in reported frequencies, anvi'o relies on the following conservative heuristic to determine whether to report the variation at a nucleotide position:

$$y = (\frac{1}{b})^{(x^{\frac{1}{b}} - m)} + c \tag{1}$$

where x represents the coverage, and b, m, and c represent the model parameters equal to 3, 1.45, and 0.05, respectively. Assuming n_1 and n_2 represent the frequency of the most frequent and the second most frequent bases in a given nucleotide position, base frequencies are reported only if $n_2/n_1 > y$ criterion is satisfied for a given coverage of x. Briefly, this approach sets a dynamic baseline for minimum variation required for reporting as a function of coverage depth. According to this heuristic, y would be 0.29 for 20X coverage (x), 0.13 for 50X coverage, 0.08 for 100X coverage, and 0.05 for very large values of coverage as y approaches to c. This computation- and storage-efficient strategy reports a short list of sample-specific, variable nucleotide positions that infrequently originate from PCR or sequencing errors. The user also has the option of instructing the profiler to store all observed frequencies for more statistically appropriate but computationally intensive downstream analyses.

Profiling variability. To interpret the ecological significance of sample-specific variable positions across samples, anvi'o installs a helper program, anvi-gen-variability-profile (AGVP). The user can specify filters that employ information from the experimental design to instruct AGVP's generation of a more refined variability profile. The current version of AGVP processes variable positions in a genome bin (see 'Genome binning') based on multiple user-defined, optional filters, including the number of variable positions to sample from each split (-n; default: 0), minimum ratio of the competing nucleotides at a reported variable position (-r; default: 0), minimum number of samples in which a nucleotide position is reported

as a variable position (-x; default: 1), and the minimum scattering power of a variable nucleotide position across samples (-m; default: 0). Samples in a merged profile can be organized into one or more groups (g) based on the nucleotide identity of the competing bases (b) at a given variable position, p. Scattering power then represents the number of samples in the second largest group. For instance, at one extreme b would be identical in all samples at position p, in which case g would be 1, and scattering power of p would be 0. At the other extreme, p would harbor a different b in every sample, in which case g would equal to the number of samples, and the scattering power of p would equal to 1. A value for g between these two extremes would yield a scattering power of > 1. The user can employ scattering power to query only those nucleotide positions that vary consistently across samples, and discard positions that show stochastic behavior that are more likely to result from sequencing or PCR errors, or mapping inconsistencies that are not of interest.

Genome binning. Anvi'o metagenomic workflow offers two modes for binning contigs into draft genomes: automatic binning, and human-guided binning. The result of a binning process corresponds to a collection in a profile database. Each collection consists of one or more bins, with each bin containing one or more splits. When anvi'o merges multiple profiles, it passes coverage values of each split across samples to CONCOCT [36] for the unsupervised identification of genome bins. CONCOCT uses Gaussian mixture models to predict the cluster membership of each contig whilst automatically determining the optimal number of clusters in the data through a variational Bayesian approach [36]. The merged profile database stores the result of unsupervised binning as a collection. Anvi'o provides the user with an easy-to-use interactive interface to visualize unsupervised binning results, and an interface to refine poorly identified bins. CONCOCT automatically installs with anvi'o, but the user can import clustering results from other unsupervised binning techniques into separate collections in the profile database. During the merging step, anvi'o can generate a hierarchical clustering of contigs using multiple clustering configurations. A clustering configuration text file describes one or more data sources for the hierarchical clustering algorithm. A clustering configuration can request the retrieval of data for each contig from a profile database (such as a single attribute or a view), from an annotation database, or from an external user-selected data source. A clustering configuration can also specify normalizations for each data source for anvi'o to employ when mixing multiple sources of information prior to the clustering analysis. The current version of anvi'o uses three default clustering configurations for merged profiles: 'tnf', 'tnf-cov', and 'cov'. Configuration 'tnf' considers k-mer frequencies to represent sequence composition of contigs for clustering. The default 'k' is 4, but the user can set different values for 'k' in new annotation databases. Configuration 'tnf-cov' mixes k-mer frequencies from the annotation database with log-normalized coverage vectors from the merged profile database. This configuration considers both sequence composition and the coverage across samples in a manner similar to CONCOCT. Configuration 'cov' uses only the coverage information from the profile database but ignores sequence composition. Each clustering configuration stores a Newick-formatted tree description of contigs into the profile database, which later becomes the central organizing framework of the interactive

interface. Different clustering configurations can generate alternative organizations of contigs, and the user can switch between visualizations of these organizations while working with the interactive interface to investigate different aspects of the data. The modular design behind the clustering infrastructure allows the user to add new clustering configurations without changing the code base, and improves the supervised binning process. Anvi'o can generate a complete and comprehensive summary of a collection upon completion of the binning process. The summary output is a user-friendly static HTML web site that can be viewed on any computer with or without an anvi'o installation, or network access.

Interactive interface. The interface has the ability to visualize large tree structures, and overlay numerical and categorical data across the tree. This approach allows anvi'o to display splits with a particular organization dictated by a tree structure, and associate each leaf with a single item in each layer mapped across the entire tree. These items can display numerical or categorical information (such as GC-content, or taxonomy). The interface can guide supervised binning and refinement of bins. The user can create a new collection to organize contigs into bins through mouse clicks, or load and modify collections previously stored in the profile database. The advanced search function of the interface can identify contigs that meet specific criteria and highlight their location on the tree, bin them together, or direct their removal from existing bins. The right-click menu provides fast access to NCBI tools to query public databases, and gives access to detailed inspection page for a given contig. The detailed inspection page displays coverage values and frequencies of variable bases for each nucleotide position in each sample for a given contig, and it overlays open reading frames and HMM hits on the contig. The interactive interface uses SVG objects for visualization, and displayed trees can be exported as high-quality, publication-ready figures.

Limitations. Certain steps of the anvi'o metagenomic workflow require intensive computation (such as profiling and merging), while others perform more efficiently on personal computers due to their interactive nature (such as visualization and supervised genome binning). Anvi'o optimally runs on server systems for non-interactive and parallelizable steps, and on personal computers for visualization tasks. However, the design of anvi'o does not impose any limits on different configurations: the entire workflow can be run on a server as an independent web service, or on a personal computer with or without network access. The interactive interface can visualize a very large number of SVG objects, and its performance depends on the user's configuration since all interactive computations are done on the user's web browser. The interface can visualize up to 500,000 SVG objects, and trees that contain up to 25,000 leafs on high-end laptop computers, however large visualization tasks decrease the responsiveness of the interface. One of the biggest limitations of anvi'o is the number of splits that can be clustered for supervised binning. Supervised binning fails for data sets containing 25,000 splits because hierarchical clustering algorithms do not scale well with a time complexity of $O(n^2)$ or more. To work around this limitation, the user can mix unsupervised and supervised approaches by starting with unsupervised clustering, and refining coarse genome bins through the anvi-refine program. In this workflow, the user employs

CONCOCT results, and then refines bins with high redundancy estimations into high-quality draft genomes. The URL http://merenlab.org/projects/anvio provides a detailed guide for best practices.

Preparation of publicly available sequencing datasets

Noise filtering, assembly, mapping, and functional characterization of contigs. For each dataset, we analyzed the raw metagenomic data with illumina-utils library [53] version 1.4.1 (available from https://github.com/meren/illumina-utils) to remove noisy sequences using 'iu-filter-quality-minoche' program with default parameters, which implements the noise filtering described by Minoche *et al.* [54]. CLC Genomics Workbench (version 6) (http://www.clcbio.com) performed all assembly and mapping tasks. For mapping, we required 97% sequence identity over 100% of the read length, and exported results as BAM files. We used RAST [52] and myRAST (available from http://blog.theseed.org/downloads/) for functional characterization of contigs.

Infant gut metagenomes. Sharon *et al.* [19] collected daily infant gut samples 15-19 and 22-24 days after birth including biological replicate samples on days 15, 17 and 22. Shotgun metagenomic analyses for the 11 samples share the SRA accession ID SRA052203. We co-assembled all samples after quality filtering, and discarded contigs shorter than 1,000 base pairs. After mapping short reads from each sample back to these contigs (Additional file 1), we used anvi'o to perform profiling and merging of samples, and supervised binning. After splitting draft genomes from our supervised binning into 1,000 bp pieces, we used blastn version 2.2.28+ [55] to determine their level of concordance with the draft genomes published by Sharon *et al.* (available at http://ggkbase.berkeley.edu/carrol). To simplify computational complexity, the analyses of variability between closely related draft genomes only included a single shotgun metagenome for each sampling day (inclusion of biological replicate metagenomes with the largest number of reads for days 15, 17 and 22). We used 'anvi-gen-variability-profile' (AGVP) program to access the variable positions reported in the merged profile database by specifying a maximum of 5 nucleotide positions from each split (-n 5), and only retaining positions with a scattering power of three (-m 3) (see 'Profiling variabilty' for the definition). For supervised genome binning we used the interactive interface.

Deep Horizon samples. We used anvi'o to interrogate several previously published cultivar and single cell genomic, metagenomic, and metatranscriptomic datasets for environmental nucleic acid preparations from Pensacola Beach (Florida, USA) sand samples and Gulf of Mexico (GOM) water samples before and after the 2010 Deep Horizon oil spill.

Overholt isolates. Data for ten culture genomes from Overholt *et al.* [41] has the NCBI BioProject ID 217943. We concatenated all 10 cultivar genomes into a single FASTA file for downstream analyses.

Rodriguez-R metagenomes. Raw metagenomic sequencing data for 16 samples from Rodriguez-R *et al.* [40] has the SRA ID PRJNA260285. After noise filtering,

we mapped short reads from each sample back to Overholt isolates (Additional file 1). Anvi'o profiled and merged resulting BAM files. In parallel, we co-assembled the metagenomic dataset, and discarded contigs smaller than 1,000 base pairs. After mapping short reads back to the co-assembled contigs (Additional file 1), anvi'o profiled individual BAM files and CONCOCT version 0.4.0 [36] performed an unsupervised binning. We summarized the CONCOCT results using anvi-summarize, and anvi-refine to interactively split CONCOCT bins and used anvi-refine to interactively partition CONCOCT bins into high-quality draft genomes with high-completion and low-redundancy estimates

Mason single-cell genomes, metagenomes, and metatranscriptomes, and Yergeau metagenomes. The web site http://mason.eoas.fsu.edu/ posts quality-filtered data for three single-cell genomes (SAGs), three metagenomes, and two metatranscriptomes [42, 43]. We obtained quality-filtered data for metagenomes previously reported by Yergeau *et al.* [44] from http://goo.gl/pfHfl8. From Yergeau metagenomes we only used the three samples collected from BM57 station, which is 3.87 km from the wellhead. Figure S1 summarizes our co-assembly, mapping, and analysis steps for these data sets. We first co-assembled short reads from the three Mason SAGs, and independently co-assembled short reads from the three Mason metagenomes. Next, we mapped short reads from each of the Mason metagenomic, metatranscriptomic, and SAG samples, as well as the three Yergeau metagenomes to the co-assembled metagenomic dataset, and separately to the co-assembled SAG genome dataset generating two BAM files for each sample (Figure S1)(Additional file 1). We independently profiled each of the resultant BAM files (16 from Mason, 6 from Yergeau samples), and merged the 11 profiles from BAM file mappings to the metagenomic co-assemblies and separately merged the 11 profiles from BAM file mappings to the SAG co-assemblies. We instructed anvi'o through an additional clustering configuration to employ only three Mason metagenomes for hierarchical clustering of contigs. We subsequently processed the merged profiles (1) to quantify the presence of short reads from metagenomic and metatranscriptomic reads matching to SAGs, (2) to quantify the presence of short reads from SAGs in the metagenomic contigs, and (3) to identify draft genomes through supervised binning. To compare variability across samples, we generated variability profiles with AGVP program for each genome bin we identified in the metagenomic assembly. We instructed AGVP to sample up to 5 co-occurring variable nucleotide positions from each split in proximal and distal samples to generate variability profiles for each genome bin separately.

We used R version 3.1.2 [56] for statistical analyses, the R library ggplot version 1.0.0 [57] for all visualizations that were not done by anvi'o, and Inkscape version 0.48 (https://inkscape.org/) to finalize figures for publication. https://github.com/meren/anvio-methods-paper-analyses gives access to the shell and R scripts we implemented to generate variability profiles and to visualize results.

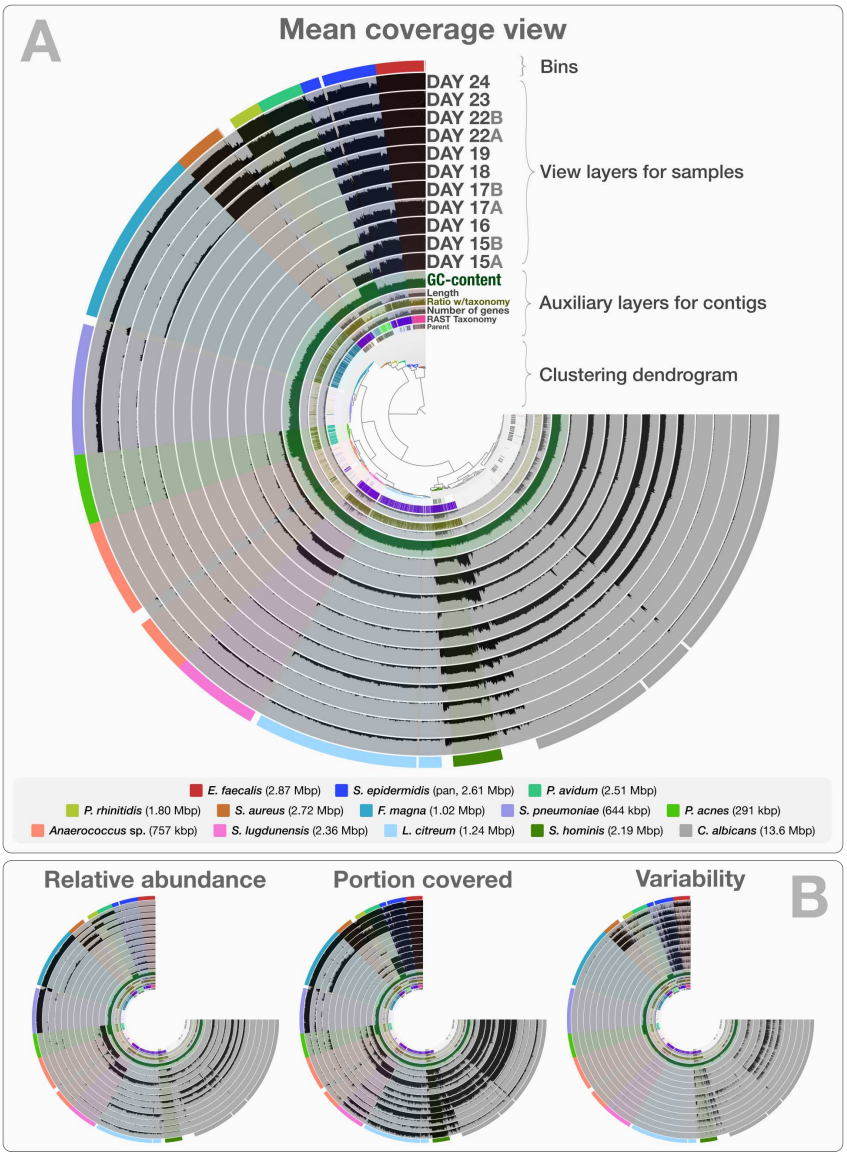

Figure 2 Static images from the anvi'o interactive display for the infant gut dataset with genome bins. The clustering dendrogram in the center of Panel A displays the hierarchical clustering of contigs based on their sequence composition, and their distribution across samples. Each tip on this dendrogram represents a split (anvi'o divides a contig into multiple splits if it is longer than a certain amount of nucleotides, which is 20,000 bps in this example). Each auxiliary layer represents essential information for each split that is independent of their distribution among samples. In this example auxiliary layers from the inside out include (1) the parent layer that marks splits originate from the same contigs with gray bars, (2) the RAST taxonomy layer that shows the consensus taxonomy for each open reading frame found in a given split, (3) the number of genes layer that shows the number of open reading frames identified in a given split, (4) the ratio with taxonomy layer that shows the proportion of the number of open reading frames with a taxonomical hit in a given split, (5) the length layer that shows the actual length of a given split, and finally (6) the GC-content layer. The view layers layers for samples follow the auxiliary layers section. In the view layers section each layer represents a sample, and each bar represents a datum computed for a given split in a given sample. Panel A demonstrates the "mean coverage", where the datum for each bar is the average coverage of a given split in a given sample. Panel B exemplifies three other views for the same display: "relative abundance", "portion covered", and "variability" of splits among samples.

Results and discussion

Characterization of variable nucleotide positions in genome bins

The co-assembly of 11 samples in the infant gut dataset yielded 4,189 contigs with a minimal length of 1,000 bps, a total assembly size of 35.8 Mbp and an N50 of 36.4 kbp. On average 92.4% (std: 4.43%) of all reads mapped back to contigs from each sample. The supervised binning of the infant gut data with anvi'o converged upon 12 bacterial and one fungal genome bin, that largely agree with the draft genomes Sharon *et al.* reported [19]. Additional file 1 reports the quality filtering and mapping statistics, as well as the attributes of recovered genome bins. Figure 2 demonstrates the interactive interface of anvi'o, as it visualizes (1) the clustering dendrogram for contigs based upon their composition and differential coverage, (2) auxiliary layers that report information about contigs stored in the annotation database (GC-content, RAST taxonomy, number of genes, etc.), (3) view layers that report information about contigs across samples stored in the profile database (while Panel A shows the mean coverage view, panel B exemplifies three other views), and (4) our draft genome bins. Having access to sample-independent auxiliary layers as well as sample-specific view layers that provide information for each contig in one interactive display improves the user's ability to work interactively with a given co-assembly. The URL http://merenlab.org/data/ gives read-only access to the interactive interface shown in Figure 2, and the automatically generated anvi'o summary for this analysis.

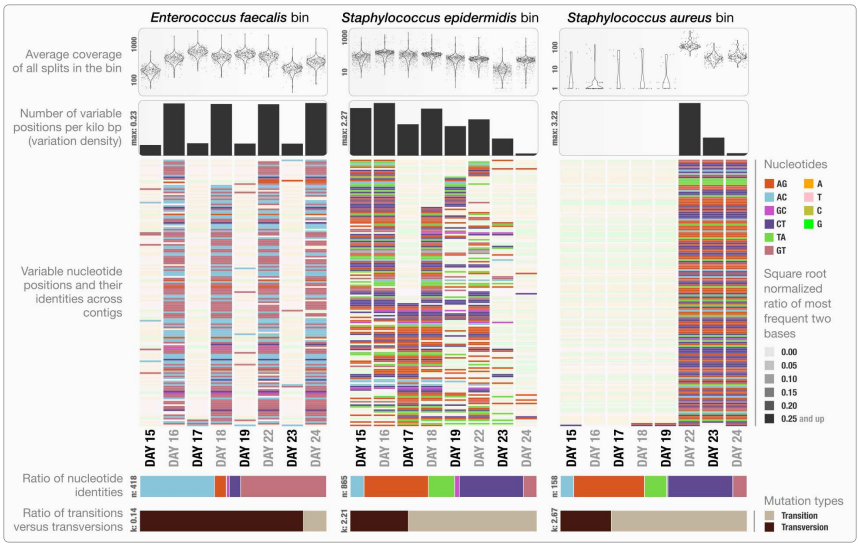

Figure 3 Variable nucleotide positions in contigs for three draft genome bins. The figure displays for each genome bin in each sample (from top to bottom), (1) average coverage values for all splits, (2) variation density (number of variable positions reported during the profiling step per kilo base pairs), (3) heatmap of variable nucleotide positions, (4) ratio of variable nucleotide identities, and finally (5) the ratio of transitions (mutations that occur from A to G, or T to C, and vice versa) versus transversions. In the heatmap, each row represents a unique variable nucleotide position, where the color of each tile represents the nucleotide identity, and the shade of each tile represents the square root-normalized ratio of the most frequent two bases at that position (i.e., the more variation in a nucleotide position, the less pale the tile is).

Anvi'o can characterize positional nucleotide variation during the profiling step without requiring reference genomes. This information provides the basis for inferring subtle population dynamics within genome bins. We applied our analysis of nucleotide variation to three genome bins in the infant gut dataset: the two most abundant bins, *Enterococcus faecalis* and *Staphylococcus epidermidis* with average coverage of 480X, and 60X respectively, as well as the *Staphylococcus aureus* bin that becomes abundant during the final three days of sampling with an average coverage of 50X. Anvi'o's profiling reported across all samples 3,241, 29,682 and 12,194 variable positions for the *E. faecalis*, *S. epidermidis*, and *S. aureus* bins respectively. Using the raw numbers for each sample in the three bins (Additional file 1), we first analyzed the variation density, which we define as the number of variable positions per kbp of contigs in a genome bin. *S. epidermidis* exhibited the highest variation density with a value of 2.27 on day 16 (second day of sampling). We then used 'anvi-gen-variability-profile' to focus only on those nucleotide positions that showed consistent variation across samples by randomly sampling up to five nucleotide positions from each split. This analysis reported 418 positions for *E. faecalis*, 865 positions for *S. epidermidis*, and 158 positions for *S. aureus*. The *Staphylococcus* bins exhibited transition/transversion ratios of 2.21-2.67 consistent with expectations that transitions (mutations that occur from A to G, or T to C, and vice versa) usually occur more commonly than transversions [58]. In contrast, the *E. faecalis* bin displayed a transition/transversion ratio of 0.14. Our analysis also revealed very different nucleotide substitution patterns among the three groups. Increased variation density within contigs from the *E. faecalis* bins on even days alternates with lower variation density on odd numbered days (Figure 3). The variation pattern, which includes conservation of nucleotide substitution patterns on alternate days at the same sites for *E. faecalis* bins suggests an underlying mechanism that does not affect other metrics such as coverage, and variation density. Initial inspection of this pattern suggests the possibility of 24-hour clonal sweeps that succumb to re-establishing a mixed population of a few different strains 24 hours later. More likely, differences in methodology account for these patterns: Sharon *et al.* used two different size selections during the library preparation for these data: while they prepared samples from odd days with an insert size of 900 bp, they used 400 bp insert size for samples from even days in their otherwise identical sample preparation. Variation in error frequencies between different Illumina sequencing runs or possibly differences in insert length that will affect cluster density might explain these patterns. Yet their non-random occurrence including clear patterns for each of the different major bins remains unexplained. In contrast, *S. epidermidis* and *S. aureus* bins did not show a bi-daily trend, and changes in their variability patterns did not follow the variability patterns anvi'o reported for *E. faecalis*. In their detailed analysis, Sharon *et al.* detected multiple strains in the *S. epidermidis* bin, members of which shifted throughout the sampling period. In our analysis we detected a high variation density for the *S. epidermidis* bin, resonating with the highly mixed nature of this population. Variation density decreased in the *S. epidermidis* bin in time, and while the coverage of this bin did not change dramatically, the nucleotide variation nearly disappeared in samples from the last day (Figure 3), suggesting a shift in the population with the dominance of a relatively small number of *S.*

epidermidis genomes. The absence of variability for *S. aureus* during the initial five-day sampling period reflects the mapping of very few metagenomic reads to these genomes but by the 22nd day, *S. aureus* flourished with a very high variation density, which steadily decreased independent of the stable coverage.

Other investigators have utilized single bp changes to compare different variants of the same species based on reference genomes [59, 60]. While less frequent, identification of single bp changes have also been used to characterize heterogeneity in naturally occurring microbial populations through metagenomics [28, 61, 62]. However, recovering detailed reports of single bp change patterns has not been straightforward due to the lack of adequate algorithms that can automatically identify and report nucleotide positions of high-variability inferred from multiple samples using contigs constructed de novo as reference for metagenomic short reads. The default metagenomic workflow of anvi'o now makes the under-exploited variability patterns accessible for every level of analysis. Application of our approach to draft genomes may lead to novel observations as well as more targeted investigations to describe underlying mechanisms that drive ecological processes. For instance, why does the *E. faecalis* population show bi-daily patterns in Sharon *et al.*'s dataset when S. epidermidis and *S. aureus* populations do not? Although exploring this question further falls outside the scope of our study, the observation of the single bp substitution patterns demonstrates the utility of anvi'o at providing deeper insights into metagenomic data.

Holistic analysis of the microbial response to the Deep Water Horizon

In contrast to the infant gut dataset, the datasets related to the Deep Water Horizon (DWH) oil spill represent a more challenging case with their size and complex nature. Following the DWH oil spill on April 20, 2010, investigators launched numerous molecular surveys to uncover bio-indicators of oil pollution and to investigate the bioremediation capacity of indigenous bacteria. Multiple studies described the strong influence of oil on the bacterial community composition in the water plume, ocean sediments, and the shoreline, as well as enrichment of oil degradation genes in affected environments [40, 41, 42, 43, 63, 64, 65]. Our DWH collection included a metagenomic dataset generated by Rodriguez-R *et al.* [40] from 16 sand samples collected from Pensacola Beach (Florida) during the three periods of beach oiling following the April 2010 DWH explosion: 'before' the oil has reached to the shore, 'during' the oil contamination, and 'after' the oil was removed (Additional file 1). The dataset includes 1) four May 2010 samples collected before oil began to wash ashore the first week of June, 2010, 2) four July 2010 and four October 2010 samples collected during the oiling event (the July and October samples each included one weathered sample with lower oil concentrations), and 3) four June 2011 samples collected after removal of oil from the beach. The original investigation of this dataset relied on taxonomic assignments of contigs from individually assembled samples without binning, and the authors observed a functional transition from generalist taxa during the oil pollution to specialists after the event. Our DWH collection also included genomes of 10 proteobacterial strains isolated from Pensacola Beach and Elmer's Island Beach (Louisiana) by Overholt *et al.* [41] using samples collected in

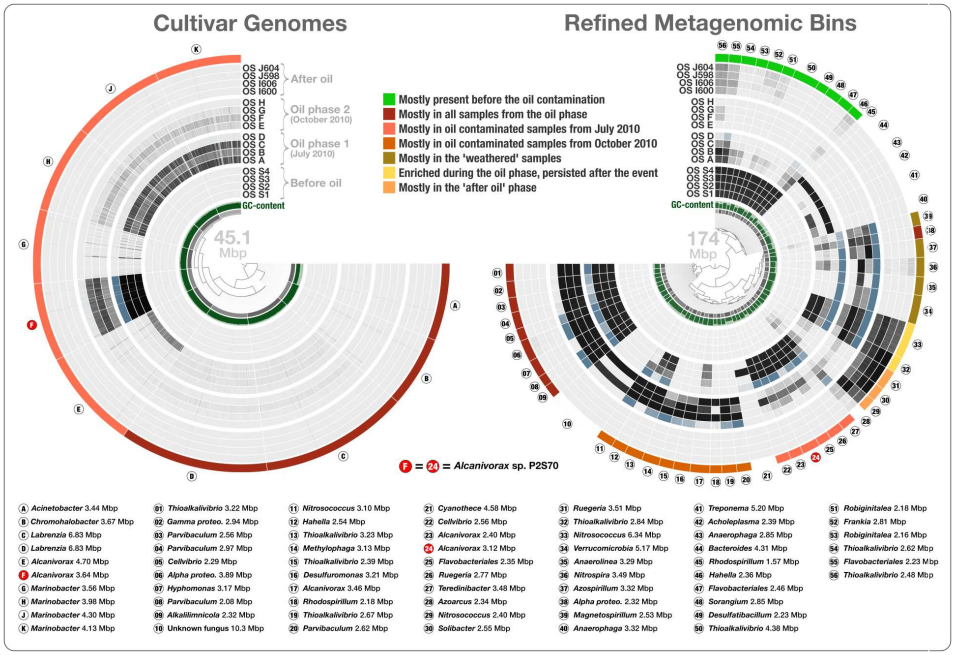

Figure 4 Overholt culture isolates linked to the Rodriguez-R metagenomes of the beach sand microbial community. The tree on the left displays the hierarchical clustering of 10 culture genomes based on sequence composition. Each view layer represents the "percent coverage" of each split in the Pensacola beach metagenomic dataset. The tree on the right displays the coverage-based hierarchical clustering of 56 environmental draft genomes we determined from the co-assembly of Pensacola Beach metagenomic dataset. The view layers display the "mean coverage" of each split in samples from the Pensacola beach metagenomic dataset. The most outer layer in both trees show the ecological pattern of a given genome bin during the period of sampling. Letters A to J identify culture genomes, and numbers 1 to 56 identify each metagenomic bin. The letter F, and the number 24, identifies two bins that represent the only genome that was present in both collections (*Alcanivorax* sp. P2S70). All genus- and higher-level taxonomy assignments are based on the best-hit function in RAST.

June and July 2010. In the original study the authors suggested that these isolates represented the dominant oil degrading microbial populations by comparing their taxonomy to an independent 16S rRNA gene-based survey of the same environment [65]. The final dataset in our DWH collection included metagenome, metatranscriptome, and single-cell genome (SAG) data generated by Mason *et al.* [42, 43] and Yergeau *et al.* [44] from the oil spill water plume samples (Additional file 1). Mason *et al.* reported a rapid response of members of the *Oceanospirillales* to aliphatic hydrocarbons [42]. Yergeau *et al.* [44] investigated the same location one year after the event and detected *Oceanospirillales* in relatively low abundance. Our re-analysis of these data using anvi'o tests some of the previous assertions by providing contextual information and determining key genomic structures that were previously overlooked.

Linking culture genomics to metagenomics

To estimate the abundance of Overholt isolates in the Pensacola Beach before, during, and after the oil contamination, we mapped the short reads from Rodriguez-R metagenomes to these 10 cultivar genomes. Overholt isolates recruited on average

0.00097% of the May 2010, 1.16% of the July 2010, 0.088% of the October 2010, and 0.0024% of the June 2011 metagenomic reads (Figure 4, and Additional file 1). Anvi'o indicates high completion with little redundancy for these genomes (Additional file 1). Among the ten cultivars, *Alcanivorax* sp. P2S70 corresponded to the most frequently detected genome (Additional file 2). On average, the July 2010 metagenomes covered 96% of the *Alcanivorax* sp. P2S70 genome to 8X depth while the October 2010 metagenomes covered only 35% of the *Alcanivorax* sp. P2S70 genome with an average depth of 0.6X. Reads from the metagenome data set of 452 million sequences mapped at very low levels to five of the isolates. Nonetheless, we observed a clear increase in the abundance of the ten genomes from 'before' to 'during' phases of the oil contamination, with a striking four thousand-fold increase of *Alcanivorax* sp. P2S70 between May and July 2010. The recovery of these genomes diminished in the two 'weathered' samples. Finally, the absence of short reads matching any of these ten genomes in samples from the 'after' phase, suggests these isolates might have depended upon oil for their primary carbon source, or their growth might requires syntrophic partnerships with other oil degrading microbes. The metagenomic data in our combined analysis supports the hypothesis that increased oil concentration creates a niche for the cultivars from Pensacola Beach. However, as these cultivars recruited only 0.0098% to 1.84% of the metagenomic reads from the same environment, our results also show that they were not the most abundant oil degraders (Figure 4, panel A) and contradict with Overholt *et al.*'s 16S rRNA gene-based estimations [41].

To access genomes of dominant oil degraders in the Gulf of Mexico shoreline without relying on cultivation, we co-assembled the Rodriguez-R dataset of 452 million reads. The de novo assembly yielded 56,804 contigs with a minimal length of 2.5 kbp and a total assembly size of 325.2 Mbp. The assembled bins recruited in average 20.4% of each sand metagenome during mapping (Additional file 1). This score represents only of 0.31% for the cultivar genomes. The large size and fragmentation of the metagenomic assembly prevented us from a direct hierarchical clustering and visualization of all contigs for supervised binning. For large datasets anvi'o offers a workflow that combines the unsupervised and supervised binning steps. CONCOCT's unsupervised binning during anvi'o merge generated 81 bins with an average redundancy of 31.7%. We then processed, visualized and manually partitioned these bins using anvi'o, leading to the determination of 162 refined bins with an average redundancy of 1.96% (Additional file 2). For a more focused analysis, we used genome bins larger than 2 Mbp and/or more than 80% complete. The 56 draft genomes that fit to these criteria had an average length 3.11 Mbp (std: 1.31 Mbp), and their GC-content varied from 32.2% to 71.0%. We compared these drafts genomes, along with the Overholt cultivars, to the best matching reference genomes using the best-hit function implemented in RAST (Additional file 1). The RAST taxonomical inference supported Overholt's *et al.*'s assignments for 9 out of the 10 genomes derived from cultivation (one misclassified draft genome was identified as *Chromohalobacter* instead of *Halomonas*), and detected a total of 33 genera within the 56 draft genomes, which included a fungus (10.3 Mbp in length), and a *Cyanobacterium* affiliated with Cyanothece that harbors 60 genes encoding for

the photosynthesis aparatus. These taxonomical inferences largely agree with analysis of sample-centric contigs by Rodriguez-R *et al.* [40]. The only organism that co-occurred in both Overholt cultivars and draft genomes from the metagenomic binning was *Alcanivorax* sp. P2S70. The metagenomic binning process recovered 86% of its genome ('bin 24' in Figure 4), with 1,858 identical proteins. An overall protein identity of 99.2% for 95.9% of proteins from this bin matched to *Alcanivorax* sp. P2S70 (Additional file 2).

Seven among the 66 culture and draft genomes occurred primarily in only a single sample. In addition, one draft genome was not characteristic to any phase (bin 28), and one draft genome represented a fungal organism (bin 10). The remaining 57 bacterial genomes exhibited one of seven distinct ecological patterns (Figure 4 and Additional file 2): 1) mostly present before the oil contamination (n=11), 2) characteristic of all samples from the oil phase (n=14, includes 4 cultivars), 3) characteristic of oil contaminated samples from July 2010 (n=12, includes the 6 remaining cultivars), 4) characteristic of oil contaminated samples from October 2010 (n=10), 5) characteristic of the weathered samples (n=5), 6) enriched during the oil phase and persisted after the event (n=2), and finally 7) characteristic of the "recovered" phase (n=3) (Figure 4). Interestingly, the most frequently represented genus (*Thioalkalivibrio*, n=8) occurred in four of the seven ecological patterns, emphasizing the importance of sensitive microbial population partitioning, and limitation of taxonomy-based binning. We grouped functions that occurred in our collection of bacterial draft genomes based on these seven ecological patterns. 2,621 of 12,982 functions differentially occurred across different ecological responses (ANOVA, Tukey-Kramer post-hoc test, p ¡ 0.05; Additional file 2).

Genes involved in oil degradation and described by Rodriguez-R *et al.* [40] likely drive shifts in beach microbial community during oil spills. The occurrence of oil-degrading microbes detected in beach sand might reflect the presence of members of the rare biosphere, and/or the ocean. Here we screened for genes in our bins to investigate whether they might offer information about the origin of oil-degrading bacteria. Among the functions characteristic of genomes enriched during the oil phase, several traits represented acquisition and metabolism of urea (Additional file 2). Urea is a dissolved organic nitrogen compound that can occur at highly abundant levels in coastal oceanic systems and serves as a main source of nitrogen for marine bacteria [66]. The apparent lack of urea metabolism in genomes characteristic of the uncontaminated beach samples in this dataset suggest this compound does not serve as a primary source of nitrogen in the innate microbial populations. On the other hand, the acquisition of carbon sources through oil degradation processes likely triggers an increased need for micronutrients such as nitrogen, and urea might represent an important source of nitrogen to support the bioremediation process. Urea functional traits suggest a life style more adapted to the marine environment, giving support to the hypothesis of an oceanic origin for microbes involved in the bioremediation process at the oil-contaminated Pensacola Beach.

Co-assembly of the metagenomic data, and the identification of draft genomes through anvi'o, revealed a more comprehensive perspective on community changes in response to the oil spill relative to the cultivars alone, which depicted only two

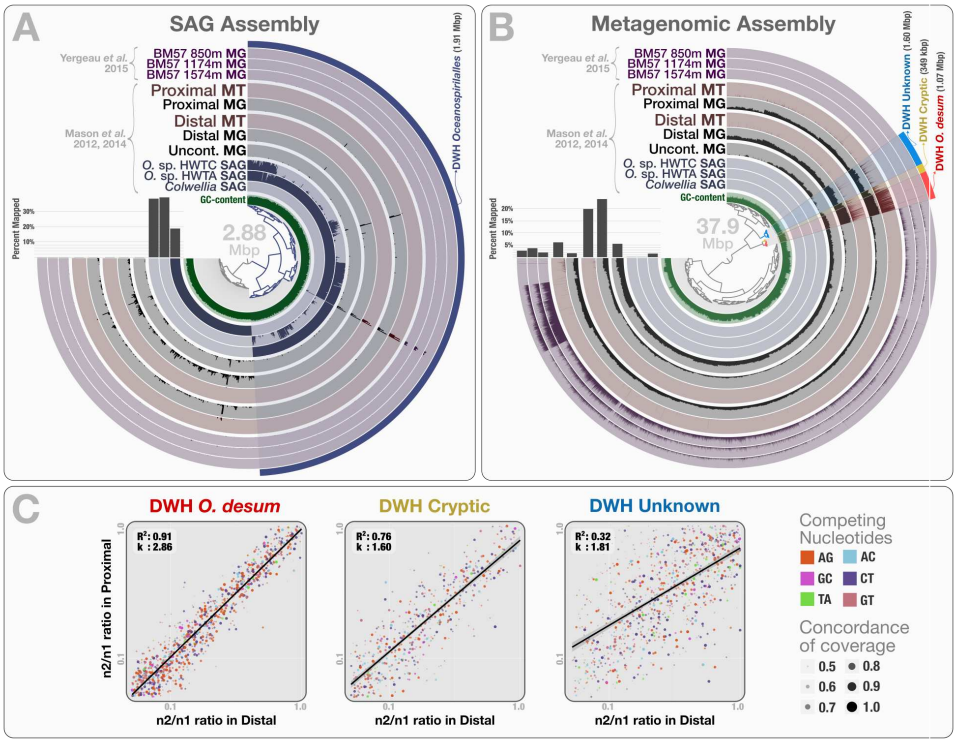

Figure 5 Mapping of samples to SAGs and metagenomic assembly, and nucleotide frequencies and identities of variable positions in three bins. Panel A shows the mapping of Mason *et al.* (2012, 2014) samples, as well as the three Yergeau *et al.* (2015) depth profiles collected from a location close to Mason *et al.*'s proximal station, to the co-assembly of the three SAGs. The dendrogram shows the sequence composition-based hierarchical clustering of the community contigs with the "portion covered" view, where each bar in the sample layers represents the percentage of coverage of a given contig by at least one short read in a given sample (i.e., if each nucleotide position in a contig is covered by at least one read, the bar is full). Panel B shows the mapping of the same samples to the co-assembly of the three Mason *et al.* metagenomes. The dendrogram shows the sequence composition- and coverage-based hierarchical clustering of the community contigs with the "mean coverage" view, where each bar in the sample layers represents the average coverage of a given contig in a given sample. Bar charts on the left-side of dendrograms both in Panel A and Panel B show the percent mapped reads from each sample to the assembly. Panel C compares the identity and frequency of the competing nucleotides at the co-occurring variable positions in three bins identified in the Panel B: DWH *O. desum*, DWH Cryptic, and DWH Unknown. X- and Y-axes in each of the three plots represent the ratio of the second most frequent base (n2) in a variable position to the most frequent base (n1) in distal, and proximal samples, respectively. Each dot on a plot represents a variable nucleotide position. The color of a given dot represents the identity of competing nucleotides. The size of a given dot increases if the coverage of it is similar in both samples, where size equals to '1 - std(coverage in proximal, coverage in distal)'. Linear regression lines show the correlation between the base frequencies at variable nucleotide positions. Each plot also displays the R2 values for linear regressions, and the ratio of transition versus transversion rates (k).

ecological patterns and represented relatively low abundance populations. The most significant functional difference between the 10 cultivars and the 59 draft genomes involved the arsenic resistance protein ArsH (p: 4.01e-21), which occurred in all culture genomes, but only in one bacterial draft genome. While multiple factors likely affect the cultivability of microbes when using oil as a sole source of carbon, arsenic, a toxic consequence of most oil spills [67], might differentially impact the fitness of oil degraders and prevent the isolation of some of the most promising populations for bioremediation processes.

Linking single-cell genomes, metatranscriptomes, and metagenomes

The Mason data [42, 43] contained metagenomes of ocean water samples collected five weeks after the oil spill at three locations: 1.5 km from the wellhead ('proximal' sample), 11 km from the wellhead ('distal' sample), and 40 km from the wellhead ('uncontaminated' sample). In addition to the metagenomes, the authors generated metatranscriptomic data from the proximal and distal samples, and isolated three single-cell genomes (SAGs) from the proximal sample (Additional file 1). The Yergeau data [44] contained metagenomes of ocean water samples collected one year after the oil spill at multiple depths of two locations: 3.87 km (BM57) and 37.8 km (A6, control station outside the plume) from the wellhead (Table S1). Consistent with previous studies [63, 68], Mason *et al.*'s analysis suggested that the taxonomical group DWH *Oceanospirillales* dominated the bacterial community composition and activity within the oil plume. Furthermore, Mason *et al.* suggested through their standalone analysis of SAGs, metagenomic, and metatranscriptomic datasets that the dominant and active *Oceanospirillales* possessed genes that encode the nearly complete pathway for cyclohexane degradation. The dataset from Mason *et al.*'s multifaceted sampling effort soon after the event and delayed sampling of Yergeau *et al.* provide an opportunity to investigate the microbial response to the DWH oil spill in a comprehensive manner. Anvi'o facilitated a holistic analysis of this composite dataset by linking separate sources of data into one unified perspective that led to a high-resolution genomic analysis of the dominant DWH *Oceanospirillales* population in time and space.

The co-assembly of 46.8 million reads representing 3 SAGs yielded 941 contigs with a minimal length of 1 kbp, a total assembly size of 2.88 Mbp and an N50 score of 3.88 kbp. Clustering of contigs based on their sequence composition (k=4) formed two distinct groups that represent for genetic structures originating from *Colwellia*, and *Oceanospirillales*, in agreement with Mason *et al.*'s findings (Figure 5 panel A). When combined, the two *Oceanospirillales* SAGs provided a draft genome of 1.91 Mbp that included 1.3 Mbp of shared contigs with a sequence identity over 99%. However, only 0.16-0.64% of the metagenomic and metatranscriptomic reads mapped to the *Oceanospirillales* SAGs which indicates low levels of relative abundance (Additional file 1). Moreover, a majority of mapped reads represented non-specific regions of ribosomal RNA operons (Figure 5 panel A). These results disagree with previous findings, and suggest that the recovered SAGs do not represent the dominant or active members of the microbial community at the time of sampling. Why did all the three single-cell captured organisms fail to represent an abundant member of the microbial community? This may reflect a methodological bias, and the population structure of single-cell captured members in a sample might not follow the rank abundance curve of the organisms that occur in the environment.

To recover the draft genome of DWH *Oceanospirillales* population, we co-assembled the metagenomic dataset of Mason *et al.* (397.9 million reads), which yielded 19,954 contigs longer than 1 kbp (N50: 1.88 kbp), with a total length of 37.9 Mbp. These contigs recruited 5.83% to 23.6% of Mason metagenomes, 1.52% to 3.58% of Yergeau metagenomes and 1.58% to 6.12% of Mason metatranscriptomes during the mapping

(Additional file 1). Clustering of contigs with respect to their sequence composition and coverage patterns across the three Mason metagenomes revealed a distinct bin that contained 1.07 Mbp with a completion score of 62.8%. Here we temporarily name this bin as "DWH *Oceanospirillales* desum" to avoid confusion with the previously identified DWH *Oceanospirillales* through SAGs. DWH *O. desum* recruited 77.8% and 79.5% of all mapped metagenomic reads in the proximal and distal samples, respectively. In contrast, DWH *O. desum* recruited only 3.55% of mapped reads in the uncontaminated sample, emphasizing the dramatic shift in its abundance between uncontaminated and contaminated samples five weeks after the oil spill (Figure 5, panel B). Furthermore, only 0.08-0.98% of mapped reads from Yergeau metagenomes were recruited by DWH *O. desum*, indicating that the abundance of this microbial population was not only restrained in space but also in time. It also suggest that the so called "uncontaminated station" from Mason *et al.*, might have been contaminated by oil already at the time of sampling, as the relative abundance of DWH *O. desum* was ¿20 fold higher in the corresponding metagenome compared to its average in the six Yergeau metagenomes.

In the distal and proximal samples from Mason *et al.*, DWH *O. desum* also recruited 97% and 99% of the mapped metatranscriptomic reads, respectively. Since we have not used the metatranscriptomic data for clustering, the increased mapping of the transcriptome reads to DWH *O. desum* bin confirms the link between its abundance in this dataset and its activity in the environment. The 1,375 nt long 16S rRNA gene from DWH *O. desum* matched the uncultured *Oceanospirillales* bacterium clones from proximal and distal stations published by Hazen *et al.* [63] with over 99% sequence identity. The first cultured organism hit in the NCBI's refseq_genomic database was *Oleispira antarctica* strain RB-8 with 92% identity, and the 23S rRNA gene we acquired from *O. desum* also matched to *Oleispira antarctica* with 93% identity. These results indicate that DWH *O. desum* represents the abundant and active *Oceanospirillales* population in the environment at the time of sampling. We also analyzed the variable positions that occurred in DWH *O. desum* population in proximal, distal, and uncontaminated samples. Despite the high variation density across samples, frequencies of the competing bases at positions of high nucleotide variation for DWH *O. desum* bin were nearly identical in proximal, and distal samples, indicating a similar population structure for DWH *O. desum* at both sampling stations (Figure 5 panel C). Our analysis of the metatranscriptomic data that mapped to the DWH *O. desum* bin revealed the expression of genes encoding the synthesis and export of lipids (lipid-A-disaccharide synthase, lipid A export), lipoproteins (protein LolC) and capsular polysaccharides (proteins LptB, KpsD, KpsE, KpsM and KpsT), known to act as bio-surfactants in oil degrading bacterial models by increasing the solubility of hydrocarbons [69]. Aside from the ribosomal machinery, one of the most expressed genes encoded for a cold-shock protein, suggesting that it played an important role sustaining the metabolism of this psychrophilic population in a suboptimal temperature for their growth. Overall, the functional activity of DWH *O. desum* suggests an important extracellular activity consistent with known oil degradation mechanisms coupled with a state of cellular stress.

We also identified two other bins adjacent to DWH *O. desum* that were strongly enriched in proximal and distal samples compared to the uncontaminated station and samples collected one year after the event. These clusters showed remarkable activity and coverage that were distinct from DWH *O. desum*, and from each other (Additional file 3). One of these two clusters has the size of a bacterial genome (1.6 Mbp), yet we found no single-copy gene markers, and hence a puzzling completion level of 0%. We refer to this cluster as "DWH Unknown". The second bin had a total length of only 0.35 Mbp, and we refer to it as "DWH Cryptic". We also performed an analysis of polymorphism on these bins to compare populations they represent in distal and proximal samples. Our analysis indicated that the frequencies of bases at variable positions showed much less agreement compared to DWH *O. desum* between proximal and distal samples. This observation suggests a subtle change in the population structure between the two stations. However, since the coverage of both bins in distal stations were much lower compared to DWH *O. desum*, these results may reflect a technical shortcoming rather than a biologically-relevant inference. Figure S2 demonstrates the change in coverage, the reported variable nucleotide positions in three contigs that represent each genome bin. The overall functional profiles of these two clusters did not resemble a typical bacterial genome: while the genes encoding for the ribosomal machinery were largely missing, pathways for phage machinery and protection against phages (CRISPRs and the type I restriction-modification system) were dramatically enriched (Additional file 3). In the case of DWH Unknown, most expressed genes encoded for proteins involved in the synthesis, transport, and export of capsular polysaccharides. The most expressed gene for DWH Cryptic encoded for a cytochrome P450 hydroxylase, enzyme involved in the metabolism of hydrocarbon [70]. Other highly expressed genes encoded for the transport and export of capsular polysaccharides, as well as CRISPR-associated proteins. These bins are likely represent phages or plasmids. However, their coverage values do not fully support this hypothesis, as there is no other bacterial genome bin in our assembly with comparable coverage. We did not detect any ribosomal proteins in these bins despite their rather large size, therefore it is likely that their presence in the environment were transparent to 16S rRNA gene-based surveys, as well as metagenomic analyses that do not perform genome binning. Their enrichment in the polluted stations and metabolic activity centered on polysaccharide synthesis and export suggests a role in hydrocarbon degradation, yet the origin of these two genetic structures remains unclear. The anvi'o summary of the three bins is available at address http://merenlab.org/data/

Anvi'o as a community platform

The ability to interact with metagenomic and metatranscriptomic data, identify and refine draft genome bins with real-time feedback, and report final results in a comprehensive and reproducible manner are essential needs for the rapidly growing field of metagenomics. Anvi'o introduces a high-level, dynamic visualization framework to better guide 'omics analyses and to communicate results, while it empowers its users with easy-to-use interfaces that require minimal bioinformatics skills to operate. Because of its modular structure, anvi'o can mix information the profiling

step generates from the raw input files with additional user-provided information in a seamless manner (i.e., external supervised or unsupervised binning results, experimental organization of contigs, views, or simply additional data or metadata layers). Through this flexibility, anvi'o does not impose specific analysis practices, and encourages question-driven exploration of data.

Anvi'o is an open source project, and it welcomes developers. By abstracting the monotonous steps of characterizing and profiling metagenomic data, the platform gives its users with programming skills the ability to access internal data structures and implement novel ideas quickly. We developed anvi'o using modern programming languages and paradigms, relied on easy-to-query and self-contained database files for data storage, and used open technologies for visualization tasks. These properties leverage anvi'o as a community platform that can support the development, testing, and dissemination of new approaches.

Conclusions

Anvi'o is an open-source, extensible software platform built upon open technologies and standard file formats to study 'omics data. In this study we used anvi'o to combine environmentally linked datasets of different nature from multiple investigators, identify draft genomes in supervised and unsupervised manner, infer population dynamics within draft genome bins through de novo characterization of nucleotide variation, and visualize our findings. Through anvi'o we identified systematic emergence of nucleotide variation in an abundant draft genome bin in an infant's gut, and extend our understanding of the microbial response to the 2010 Deepwater Horizon Oil Spill. Anvi'o's ability to integrate, analyze, and visualize data of diverse origin empowers its users to explore their sequencing datasets to address wide variety of questions.

Availability of supporting data

http://merenlab.org/data/ and additional files provide access to supporting data. In addition, the infant gut metagenome data from Sharon *et al.* is accessible via http://ggkbase.berkeley.edu/carrol. Overholt isolates are stored under NCBI BioProject ID 217943, Rodriguez-R metagenomes are stored under SRA ID PRJNA260285. http://mason.eoas.fsu.edu/ gives access to Mason single-cell genomes, metagenomes, and metatranscriptomes. Finally, Yergeau metagenomes are stored at http://goo.gl/pfHfl8.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

AME and TOD conceived the study, and performed the data analyses. AME and OCE implemented the software platform. CQ, JHV, and MLS participated the design and coordination of data analyses. CQ implemented the CONCOCT module for the software platform. MLS edited the manuscript. All authors contributed to writing of the manuscript, and all authors read and approved the final study.

Acknowledgements

We thank Faruk Uzun, Doğan Can Kilment, Gökmen Göksel, and Gökmen Görgen for their contributions to the code base. We thank Inés Martínez for testing anvi'o, and for her valuable suggestions throughout the development of the platform. We thank Sheri Simmons for suggesting the application of oligotyping to the metagenomic data to characterize single-nucleotide variation. We also thank Itai Sharon, Luis M. Rodridugez-R, Will A. Overholt, Olivia U. Mason, Etienne Yergeau, and their colleagues for making valuable datasets available to the science community, and for answering our questions. AME was supported by the G. Unger Vetlesen Foundation. This project was

supported by the Frank R. Lillie Research Innovation Award given by the University of Chicago and the Marine Biological Laboratory.

Author details

[1]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, 02543 MA, USA. [2] Warwick Medical School, University of Warwick, CV4 7AL Coventry, United Kingdom.

References

1. Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., Goodman, R.M.: Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology **5**(10), 245–249 (1998). doi:10.1016/S1074-5521(98)90108-9

2. Sogin, M.L., Morrison, H.G., Huber, J.A., Mark Welch, D., Huse, S.M., Neal, P.R., Arrieta, J.M., Herndl, G.J.: Microbial diversity in the deep sea and the underexplored "rare biosphere". Proceedings of the National Academy of Sciences of the United States of America **103**(32), 12115–12120 (2006)

3. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H., Smith, H.O.: Environmental genome shotgun sequencing of the Sargasso Sea. Science (New York, NY) **304**(5667), 66–74 (2004)

4. Angly, F.E., Felts, B., Breitbart, M., Salamon, P., Edwards, R.A., Carlson, C., Chan, A.M., Haynes, M., Kelley, S., Liu, H., Mahaffy, J.M., Mueller, J.E., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C.A., Rohwer, F.: The marine viromes of four oceanic regions. PLoS biology **4**(11), 368 (2006). doi:10.1371/journal.pbio.0040368

5. Edwards, R.A., Rodriguez-Brito, B., Wegley, L., Haynes, M., Breitbart, M., Peterson, D.M., Saar, M.O., Alexander, S., Alexander, E.C., Rohwer, F.: Using pyrosequencing to shed light on deep mine microbial ecology. BMC genomics **7**, 57 (2006). doi:10.1186/1471-2164-7-57

6. Turnbaugh, P.J., Ley, R.E., Mahowald, M.A., Magrini, V., Mardis, E.R., Gordon, J.I.: An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444**(7122), 1027–1031 (2006)

7. Lorenz, P., Eck, J.: Metagenomics and industrial applications. Nature reviews. Microbiology **3**(6), 510–516 (2005). doi:10.1038/nrmicro1161

8. Woyke, T., Teeling, H., Ivanova, N.N., Huntemann, M., Richter, M., Gloeckner, F.O., Boffelli, D., Anderson, I.J., Barry, K.W., Shapiro, H.J., Szeto, E., Kyrpides, N.C., Mussmann, M., Amann, R., Bergin, C., Ruehland, C., Rubin, E.M., Dubilier, N.: Symbiosis insights through metagenomic analysis of a microbial consortium. Nature **443**(7114), 950–955 (2006). doi:10.1038/nature05192

9. Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., Edwards, R.A.: The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC bioinformatics **9**(1), 386 (2008). doi:10.1186/1471-2105-9-386

10. Zakrzewski, M., Bekel, T., Ander, C., Pühler, A., Rupp, O., Stoye, J., Schlüter, A., Goesmann, A.: MetaSAMS–a novel software platform for taxonomic classification, functional annotation and comparative analysis of metagenome datasets. Journal of biotechnology **167**(2), 156–65 (2013). doi:10.1016/j.jbiotec.2012.09.013

11. Wommack, K.E., Bhavsar, J., Ravel, J.: Metagenomics: read length matters. Applied and environmental microbiology **74**(5), 1453–63 (2008). doi:10.1128/AEM.02181-07

12. Carr, R., Borenstein, E.: Comparative analysis of functional metagenomic annotation and the mappability of short reads. PloS one **9**(8), 105776 (2014). doi:10.1371/journal.pone.0105776

13. Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W., Podar, M., Short, J.M., Mathur, E.J., Detter, J.C., Bork, P., Hugenholtz, P., Rubin, E.M.: Comparative metagenomics of microbial communities. Science (New York, NY) **308**(5721), 554–557 (2005)

14. Dinsdale, E.A., Edwards, R.A., Hall, D., Angly, F., Breitbart, M., Brulc, J.M., Furlan, M., Desnues, C., Haynes, M., Li, L., McDaniel, L., Moran, M.A., Nelson, K.E., Nilsson, C., Olson, R., Paul, J., Brito, B.R., Ruan, Y., Swan, B.K., Stevens, R., Valentine, D.L., Thurber, R.V., Wegley, L., White, B.A., Rohwer, F.: Functional metagenomic profiling of nine biomes. Nature **452**(7187), 629–32 (2008). doi:10.1038/nature06810

15. Delmont, T.O., Prestat, E., Keegan, K.P., Faubladier, M., Robe, P., Clark, I.M., Pelletier, E., Hirsch, P.R., Meyer, F., Gilbert, J.A., Le Paslier, D., Simonet, P., Vogel, T.M.: Structure, fluctuation and magnitude of a natural grassland soil metagenome. The ISME journal **6**(9), 1677–87 (2012). doi:10.1038/ismej.2011.197

16. Pop, M.: Genome assembly reborn: recent computational challenges. Briefings in bioinformatics **10**(4), 354–66 (2009). doi:10.1093/bib/bbp026

17. Luo, C., Tsementzi, D., Kyrpides, N.C., Konstantinidis, K.T.: Individual genome assembly from complex community short-read metagenomic datasets. The ISME journal **6**(4), 898–901 (2012). doi:10.1038/ismej.2011.147

18. Mende, D.R., Waller, A.S., Sunagawa, S., Järvelin, A.I., Chan, M.M., Arumugam, M., Raes, J., Bork, P.: Assessment of metagenomic assembly using simulated next generation sequencing data. PloS one **7**(2), 31386 (2012). doi:10.1371/journal.pone.0031386

19. Sharon, I., Morowitz, M.J., Thomas, B.C., Costello, E.K., Relman, D.A., Banfield, J.F.: Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome research **23**(1), 111–20 (2013). doi:10.1101/gr.142315.112

20. Iverson, V., Morris, R.M., Frazar, C.D., Berthiaume, C.T., Morales, R.L., Armbrust, E.V.: Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science (New York, N.Y.) **335**(6068), 587–90 (2012). doi:10.1126/science.1212665

21. Sharon, I., Kertesz, M., Hug, L.A., Pushkarev, D., Blauwkamp, T.A., Castelle, C.J., Amirebrahimi, M., Thomas, B.C., Burstein, D., Tringe, S.G., Williams, K.H., Banfield, J.: Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Research, 183012–114 (2015). doi:10.1101/gr.183012.114

22. Delmont, T.O., Eren, A.M., Maccario, L., Prestat, E., Esen, O.C., Pelletier, E., Le Paslier, D., Simonet, P., Vogel, T.M.: Reconstructing rare soil microbial genomes using in situ enrichments and metagenomics. Frontiers in microbiology **6**, 358 (2015). doi:10.3389/fmicb.2015.00358

23. Brown, C.T., Howe, A., Zhang, Q., Pyrkosz, A.B., Brom, T.H.: A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data (2012). 1203.4802

24. Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., Corbeil, J.: Ray Meta: scalable de novo metagenome assembly and profiling. Genome biology **13**(12), 122 (2012). doi:10.1186/gb-2012-13-12-r122

25. Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L.: IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics (Oxford, England) **28**(11), 1420–8 (2012). doi:10.1093/bioinformatics/bts174

26. Zerbino, D.R., Birney, E.: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research **18**(5), 821–9 (2008). doi:10.1101/gr.074492.107

27. Treangen, T.J., Koren, S., Sommer, D.D., Liu, B., Astrovskaya, I., Ondov, B., Darling, A.E., Phillippy, A.M., Pop, M.: MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome biology **14**(1), 2 (2013). doi:10.1186/gb-2013-14-1-r2

28. Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S., Banfield, J.F.: Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428**(6978), 37–43 (2004). doi:10.1038/nature02340

29. Stein, J.L., Marsh, T.L., Wu, K.Y., Shizuya, H., DeLong, E.F.: Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. Journal of bacteriology **178**(3), 591–9 (1996)

30. Alonso-Sáez, L., Waller, A.S., Mende, D.R., Bakker, K., Farnelid, H., Yager, P.L., Lovejoy, C., Tremblay, J.-E., Potvin, M., Heinrich, F., Estrada, M., Riemann, L., Bork, P., Pedrós-Alió, C., Bertilsson, S.: Role for urea in nitrification by polar marine Archaea. Proceedings of the National Academy of Sciences of the United States of America **109**(44), 17989–94 (2012). doi:10.1073/pnas.1201914109

31. Kantor, R.S., van Zyl, A.W., van Hille, R.P., Thomas, B.C., Harrison, S.T.L., Banfield, J.F.: Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unravelled with genome-resolved metagenomics. Environmental microbiology (2015). doi:10.1111/1462-2920.12936

32. Wooley, J.C., Godzik, A., Friedberg, I.: A primer on metagenomics. PLoS computational biology **6**(2), 1000667 (2010). doi:10.1371/journal.pcbi.1000667

33. Hess, M., Sczyrba, A., Egan, R., Kim, T.-W., Chokhawala, H., Schroth, G., Luo, S., Clark, D.S., Chen, F., Zhang, T., Mackie, R.I., Pennacchio, L.A., Tringe, S.G., Visel, A., Woyke, T., Wang, Z., Rubin, E.M.: Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. Science (New York, N.Y.) **331**(6016), 463–7 (2011). doi:10.1126/science.1200387

34. Raveh-Sadka, T., Thomas, B.C., Singh, A., Firek, B., Brooks, B., Castelle, C.J., Sharon, I., Baker, R., Good, M., Morowitz, M.J., Banfield, J.F.: Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. eLife **4** (2015). doi:10.7554/eLife.05477

35. Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.r.L., Tyson, G.W., Nielsen, P.H.: Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature biotechnology **31**(6), 533–8 (2013). doi:10.1038/nbt.2579

36. Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., Quince, C.: Binning metagenomic contigs by coverage and composition. Nature Methods **11**(11), 1144–1146 (2014). doi:10.1038/nmeth.3103

37. Wu, Y.-W., Tang, Y.-H., Tringe, S.G., Simmons, B.A., Singer, S.W.: MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome **2**(1), 26 (2014). doi:10.1186/2049-2618-2-26

38. Imelfort, M., Parks, D., Woodcroft, B.J., Dennis, P., Hugenholtz, P., Tyson, G.W.: GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ **2**, 603 (2014). doi:10.7717/peerj.603

39. Cantor, M., Nordberg, H., Smirnova, T., Hess, M., Tringe, S., Dubchak, I.: Elviz – exploration of metagenome assemblies with an interactive visualization tool. BMC Bioinformatics **16**(1), 130 (2015). doi:10.1186/s12859-015-0566-4

40. Rodriguez-R, L.M., Overholt, W.A., Hagan, C., Huettel, M., Kostka, J.E., Konstantinidis, K.T.: Microbial community successional patterns in beach sands impacted by the Deepwater Horizon oil spill. The ISME journal (2015). doi:10.1038/ismej.2015.5

41. Overholt, W.A., Green, S.J., Marks, K.P., Venkatraman, R., Prakash, O., Kostka, J.E.: Draft genome sequences for oil-degrading bacterial strains from beach sands impacted by the deepwater horizon oil spill. Genome announcements **1**(6) (2013). doi:10.1128/genomeA.01015-13

42. Mason, O.U., Hazen, T.C., Borglin, S., Chain, P.S.G., Dubinsky, E.A., Fortney, J.L., Han, J., Holman, H.-Y.N., Hultman, J., Lamendella, R., Mackelprang, R., Malfatti, S., Tom, L.M., Tringe, S.G., Woyke, T., Zhou, J., Rubin, E.M., Jansson, J.K.: Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. The ISME journal **6**(9), 1715–27 (2012). doi:10.1038/ismej.2012.59

43. Mason, O.U., Han, J., Woyke, T., Jansson, J.K.: Single-cell genomics reveals features of a Colwellia species that was dominant during the Deepwater Horizon oil spill. Frontiers in microbiology **5**, 332 (2014). doi:10.3389/fmicb.2014.00332

44. Yergeau, E., Maynard, C., Sanschagrin, S., Champagne, J., Juck, D., Lee, K., Greer, C.W.: Microbial community composition, functions and activities in the Gulf of Mexico, one year after the Deepwater Horizon accident. Applied and environmental microbiology, 01470–15 (2015). doi:10.1128/AEM.01470-15

45. Atlas, R.M., Hazen, T.C.: Oil biodegradation and bioremediation: a tale of the two worst spills in U.S. history. Environmental science & technology **45**(16), 6709–15 (2011). doi:10.1021/es2013227

46. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.: The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) **25**(16), 2078–9 (2009). doi:10.1093/bioinformatics/btp352

47. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England) **25**(14), 1754–60 (2009). doi:10.1093/bioinformatics/btp324

48. Langmead, B., Salzberg, S.L.: Fast gapped-read alignment with Bowtie 2. Nature methods **9**(4), 357–9 (2012). doi:10.1038/nmeth.1923

49. Campbell, J.H., O'Donoghue, P., Campbell, A.G., Schwientek, P., Sczyrba, A., Woyke, T., Söll, D., Podar, M.: UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. Proceedings of the National Academy of Sciences of the United States of America **110**(14), 5540–5 (2013). doi:10.1073/pnas.1303090110

50. Dupont, C.L., Rusch, D.B., Yooseph, S., Lombardo, M.-J., Richter, R.A., Valas, R., Novotny, M., Yee-Greenbaum, J., Selengut, J.D., Haft, D.H., Halpern, A.L., Lasken, R.S., Nealson, K., Friedman, R., Venter, J.C.: Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. The ISME journal **6**(6), 1186–99 (2012). doi:10.1038/ismej.2011.189

51. Creevey, C.J., Doerks, T., Fitzpatrick, D.A., Raes, J., Bork, P.: Universally distributed single-copy genes indicate a constant rate of horizontal transfer. PloS one **6**(8), 22099 (2011). doi:10.1371/journal.pone.0022099

52. Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O.: The RAST Server: rapid annotations using subsystems technology. BMC genomics **9**, 75 (2008). doi:10.1186/1471-2164-9-75

53. Eren, A.M., Vineis, J.H., Morrison, H.G., Sogin, M.L.: A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. PLoS ONE **8**(6), 66643 (2013). doi:10.1371/journal.pone.0066643

54. Minoche, A.E., Dohm, J.C., Himmelbauer, H.: Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome biology **12**(11), 112 (2011). doi:10.1186/gb-2011-12-11-r112

55. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of molecular biology **215**(3), 403–410 (1990)

56. R Development Core Team, R.: R: A Language and Environment for Statistical Computing (2011). doi:10.1007/978-3-540-74686-7. http://www.r-project.org

57. Ginestet, C.: ggplot2: Elegant Graphics for Data Analysis. Journal of the Royal Statistical Society: Series A (Statistics in Society) **174**(1), 245–246 (2011). doi:10.1111/j.1467-985X.2010.00676_9.x

58. Lawrence, J.G., Ochman, H.: Amelioration of Bacterial Genomes: Rates of Change and Exchange. Journal of Molecular Evolution **44**(4), 383–397 (1997). doi:10.1007/PL00006158

59. Zhang, W., Qi, W., Albert, T.J., Motiwala, A.S., Alland, D., Hyytia-Trees, E.K., Ribot, E.M., Fields, P.I., Whittam, T.S., Swaminathan, B.: Probing genomic diversity and evolution of Escherichia coli O157 by single nucleotide polymorphisms. Genome research **16**(6), 757–67 (2006). doi:10.1101/gr.4759706

60. Morelli, G., Song, Y., Mazzoni, C.J., Eppinger, M., Roumagnac, P., Wagner, D.M., Feldkamp, M., Kusecek, B., Vogler, A.J., Li, Y., Cui, Y., Thomson, N.R., Jombart, T., Leblois, R., Lichtner, P., Rahalison, L., Petersen, J.M., Balloux, F., Keim, P., Wirth, T., Ravel, J., Yang, R., Carniel, E., Achtman, M.: Yersinia pestis genome sequencing identifies patterns of global phylogenetic diversity. Nature genetics **42**(12), 1140–3 (2010). doi:10.1038/ng.705

61. Simmons, S.L., Dibartolo, G., Denef, V.J., Goltsman, D.S.A., Thelen, M.P., Banfield, J.F.: Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. PLoS biology **6**(7), 177 (2008). doi:10.1371/journal.pbio.0060177

62. Morowitz, M.J., Denef, V.J., Costello, E.K., Thomas, B.C., Poroyko, V., Relman, D.A., Banfield, J.F.: Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. Proceedings of the National Academy of Sciences of the United States of America **108**(3), 1128–1133 (2011)

63. Hazen, T.C., Dubinsky, E.A., DeSantis, T.Z., Andersen, G.L., Piceno, Y.M., Singh, N., Jansson, J.K., Probst, A., Borglin, S.E., Fortney, J.L., Stringfellow, W.T., Bill, M., Conrad, M.E., Tom, L.M., Chavarria, K.L., Alusi, T.R., Lamendella, R., Joyner, D.C., Spier, C., Baelum, J., Auer, M., Zemla, M.L., Chakraborty, R., Sonnenthal, E.L., D'haeseleer, P., Holman, H.-Y.N., Osman, S., Lu, Z., Van Nostrand, J.D., Deng, Y., Zhou, J., Mason, O.U.: Deep-sea oil plume enriches indigenous oil-degrading bacteria. Science (New York, N.Y.) **330**(6001), 204–8 (2010). doi:10.1126/science.1195979

64. Kimes, N.E., Callaghan, A.V., Aktas, D.F., Smith, W.L., Sunner, J., Golding, B., Drozdowska, M., Hazen, T.C., Suflita, J.M., Morris, P.J.: Metagenomic analysis and metabolite profiling of deep–sea sediments from the Gulf of Mexico following the Deepwater Horizon oil spill. Frontiers in Microbiology **4**, 50 (2013). doi:10.3389/fmicb.2013.00050

65. Kostka, J.E., Prakash, O., Overholt, W.A., Green, S.J., Freyer, G., Canion, A., Delgardio, J., Norton, N., Hazen, T.C., Huettel, M.: Hydrocarbon-degrading bacteria and the bacterial community response in gulf of Mexico beach sands impacted by the deepwater horizon oil spill. Applied and environmental microbiology **77**(22), 7962–74 (2011). doi:10.1128/AEM.05402-11

66. Solomon, C.M., Collier, J.L., Berg, G.M., Glibert, P.M.: Role of urea in microbial metabolism in aquatic systems: A biochemical and molecular review (2010). doi:10.3354/ame01390

67. Cozzarelli, I.M., Schreiber, M.E., Erickson, M.L., Ziegler, B.A.: Arsenic Cycling in Hydrocarbon Plumes: Secondary Effects of Natural Attenuation. Ground water (2015). doi:10.1111/gwat.12316

68. Redmond, M.C., Valentine, D.L.: Natural gas and temperature structured a microbial community response to the Deepwater Horizon oil spill. Proceedings of the National Academy of Sciences of the United States of America **109**(50), 20292–7 (2012). doi:10.1073/pnas.1108756108

69. Ron, E.Z., Rosenberg, E.: Biosurfactants and oil bioremediation. Current Opinion in Biotechnology **13**(3), 249–252 (2002). doi:10.1016/S0958-1669(02)00316-6

70. Ortiz de Montellano, P.R.: Hydrocarbon hydroxylation by cytochrome P450 enzymes. Chemical reviews **110**(2), 932–48 (2010). doi:10.1021/cr9002193

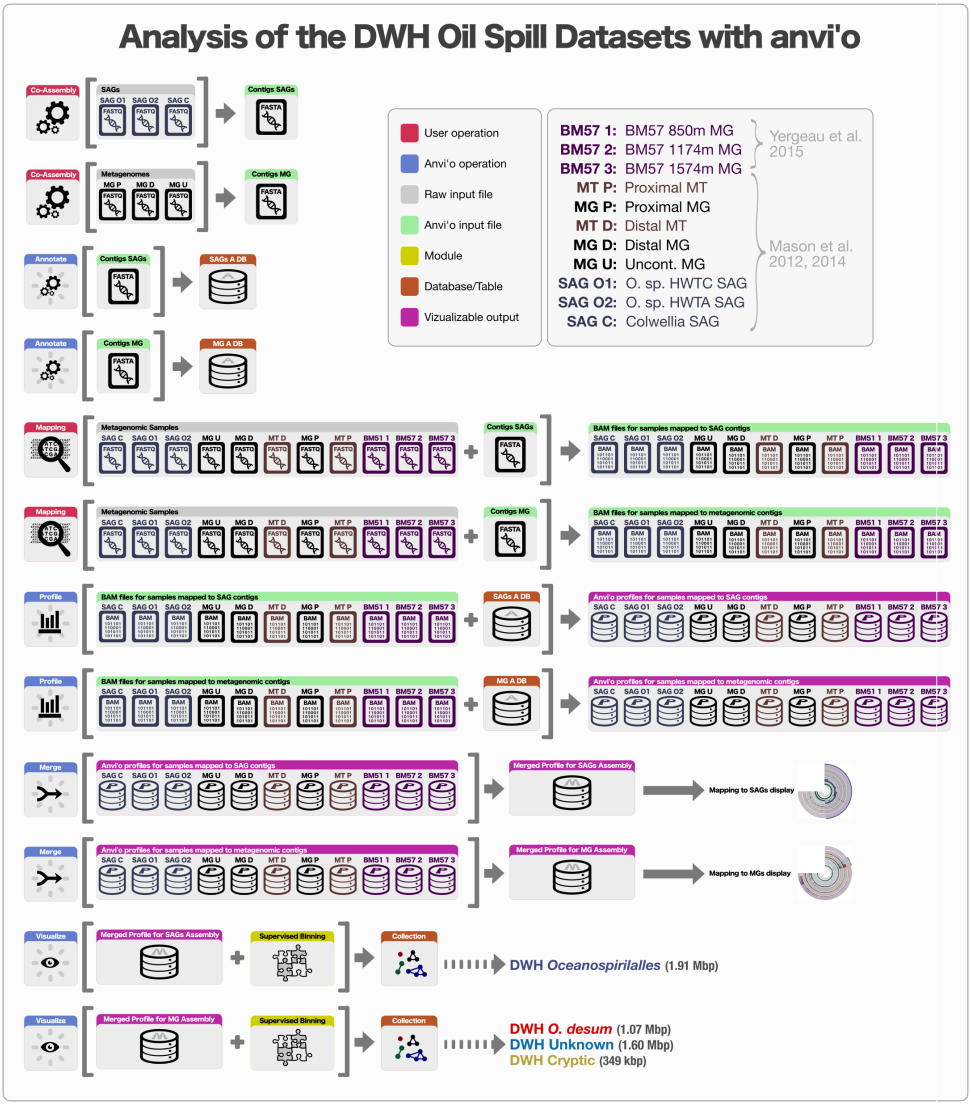

Figure 6 (**Figure S1**) Co-assembly, mapping, and anvi'o profiling steps for the combined analysis of single-cell, metagenomic, and metatranscriptomic data from Mason *et al.* and metagenomic data from Yergeau *et al.*

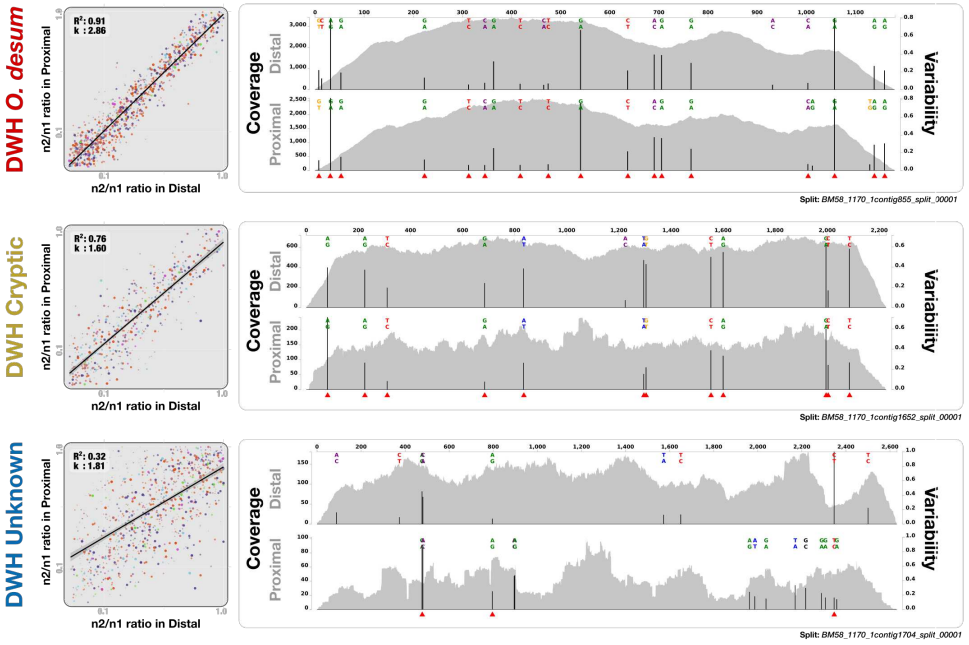

Figure 7 (**Figure S2**) Three contigs from the Mason data (shown in Figure 5) to demonstrate anvi'o's representation of coverage, variable nucleotide positions, and base frequencies. In each panel, plots on the left show the summary of all variable positions (see Figure 5 and its caption for details) in a given genome bin, while each coverage/variability plot on the right demonstrates an example contig from a given genome bin. Red triangles underneath the variable nucleotide positions identify the positions that contribute to the generation of the plots on the left side.