AMERICAN SOCIETY FOR MICROBIOLOGY | mBio®

RESEARCH ARTICLE

The Source and Evolutionary History of a Microbial Contaminant Identified Through Soil Metagenomic Analysis

Matthew R. Olm,[a] Cristina N. Butterfield,[a] Alex Copeland,[b] T. Christian Boles,[c] Brian C. Thomas,[a] Jillian F. Banfield[a]

University of California, Berkeley, California, USA[a]; Joint Genome Institute, Walnut Creek, California, USA[b]; Sage Science, Inc., Beverly, Massachusetts, USA[c]

ABSTRACT In this study, strain-resolved metagenomics was used to solve a mystery. A 6.4-Mbp complete closed genome was recovered from a soil metagenome and found to be astonishingly similar to that of *Delftia acidovorans* SPH-1, which was isolated in Germany a decade ago. It was suspected that this organism was not native to the soil sample because it lacked the diversity that is characteristic of other soil organisms; this suspicion was confirmed when PCR testing failed to detect the bacterium in the original soil samples. *D. acidovorans* was also identified in 16 previously published metagenomes from multiple environments, but detailed-scale single nucleotide polymorphism analysis grouped these into five distinct clades. All of the strains indicated as contaminants fell into one clade. Fragment length anomalies were identified in paired reads mapping to the contaminant clade genotypes only. This finding was used to establish that the DNA was present in specific size selection reagents used during sequencing. Ultimately, the source of the contaminant was identified as bacterial biofilms growing in tubing. On the basis of direct measurement of the rate of fixation of mutations across the period of time in which contamination was occurring, we estimated the time of separation of the contaminant strain from the genomically sequenced ancestral population within a factor of 2. This research serves as a case study of high-resolution microbial forensics and strain tracking accomplished through metagenomics-based comparative genomics. The specific case reported here is unusual in that the study was conducted in the background of a soil metagenome and the conclusions were confirmed by independent methods.

IMPORTANCE It is often important to determine the source of a microbial strain. Examples include tracking a bacterium linked to a disease epidemic, contaminating the food supply, or used in bioterrorism. Strain identification and tracking are generally approached by using cultivation-based or relatively nonspecific gene fingerprinting methods. Genomic methods have the ability to distinguish strains, but this approach typically has been restricted to isolates or relatively low-complexity communities. We demonstrate that strain-resolved metagenomics can be applied to extremely complex soil samples. We genotypically defined a soil-associated bacterium and identified it as a contaminant. By linking together snapshots of the bacterial genome over time, it was possible to estimate how long the contaminant had been diverging from a likely source population. The results are congruent with the derivation of the bacterium from a strain isolated in Germany and sequenced a decade ago and highlight the utility of metagenomics in strain tracking.

Received 9 November 2016 **Accepted** 25 January 2017 **Published** 21 February 2017

Citation Olm MR, Butterfield CN, Copeland A, Boles TC, Thomas BC, Banfield JF. 2017. The source and evolutionary history of a microbial contaminant identified through soil metagenomic analysis. mBio 8:e01969-16. https://doi.org/10.1128/mBio.01969-16.

Invited Editor C. Titus Brown, Michigan State University

Editor Dianne K. Newman, California Institute of Technology/HHMI

Address correspondence to Jillian F. Banfield, jbanfield@berkeley.edu.

Microbial strains of the same species can have very different traits, including virulence and drug resistance (1–3); thus, tracking of specific strain populations is important in a number of different contexts. Microbial source tracking (MST) via

quantitative PCR is routinely used to determine the source of fecal bacteria in environmental waters (4) and, in some cases, can discriminate between the fecal profiles of different types of animals (5). Tracing pathogenic strains within hospitals via sequencing of isolated strains can uncover vectors of nosocomial infections (6), and larger-scale studies have improved our understanding of the intercontinental spread of pathogens (7). The forensic investigation launched following the 2001 *B. anthracis* bioterrorism attack has been called "one of the largest and most complex in the history of law enforcement" (8). Significant effort since has been invested to develop new methods, including clustered regularly interspaced short palindromic repeat(s) (CRISPR)-Cas analysis (9), to deploy in the case of future bioterrorism events (10).

A key component of microbial forensics is strain typing. Strain resolution is essential to trace the spread of a population, and more sensitive strain typing methods can provide higher-resolution transmission maps. The most common methods identify unique (but small) markers of the microbial population's genomic DNA sequence and take advantage of the fact that random mutations develop in all growing populations. Examples of methods to identify these mutations include restriction endonuclease analysis, pulsed-field gel electrophoresis, ribotyping, and multilocus sequencing typing (11, 12). Decreasing sequencing costs have also allowed an increasing number of studies to take advantage of genome sequencing, the "gold standard" of microbial typing (6). This approach can discriminate between microbial populations that differ by even a single nucleotide, but typically this requires culturing of the organism before DNA extraction, and this is not feasible in all cases. Metagenomics, on the other hand, has the potential to trace and characterize virulent strains without cultivation and could be used to detect strains of interest in environmental samples (13).

The National Academy of Sciences stated that metagenomics "will bring about a transformation in biology, medicine, ecology, and biotechnology that may be as profound as that initiated by the invention of the microscope" (14). In metagenomics, shotgun sequencing of DNA extracted directly from environmental samples allows characterization of microbes without the need for cultivation. Assembly and binning of short metagenomic reads can yield hundreds of genomes from metagenomic samples (15–18). However, sequencing projects can be contaminated with exogenous DNA (19). Significant efforts have been made to determine where these contaminants originate (20), but precise sources of contaminant sequences are seldom identified. The mystery of determining the source of contaminant DNA in the background of a complex metagenomic sample represents a useful test case for genomics-based microbial forensics.

RESULTS

Recovery of a complete *Delftia acidovorans* genome from soil. Soil samples were collected from two locations in the Angelo Coast Range Reserve in northern California. A meadow within the reserve was sampled at two soil depths (20 and 40 cm) over a 6-week period, and a ridge nearby was sampled at six depths (15 to 115 cm, including both soil and underlying weathered shale) during one sampling event. The sites are located within the Eel Critical Zone Observatory. Metagenomic DNA was extracted from all of the samples, and between 15.7 and 44.0 Gbp of Illumina shotgun paired-end sequencing was generated for each sample. In total, 0.4 Tbp of sequence data was generated (see Table S1 in the supplemental material). Shannon diversity was calculated for all of the samples by using ribosomal protein S3 (rpS3) genes. The mean alpha diversity of the soil collected in this study was 4.65 (standard deviation, 0.28), similar to that of many previously studied soils (21–23) (see Table S1).

An initial binning analysis revealed genome fragments that were profiled as deriving from *D. acidovorans* populations in six samples (other genomes were reported by Butterfield et al. [24]). We reconstructed a 6.41-Mbp high-quality draft genome of *D. acidovorans* consisting of 53 scaffolds ranging in length from 6 to 500 kbp. The genome was further curated by read mapping to fill scaffolding gaps and extend and join the contigs. The resulting 16 contigs could be ordered and oriented to the

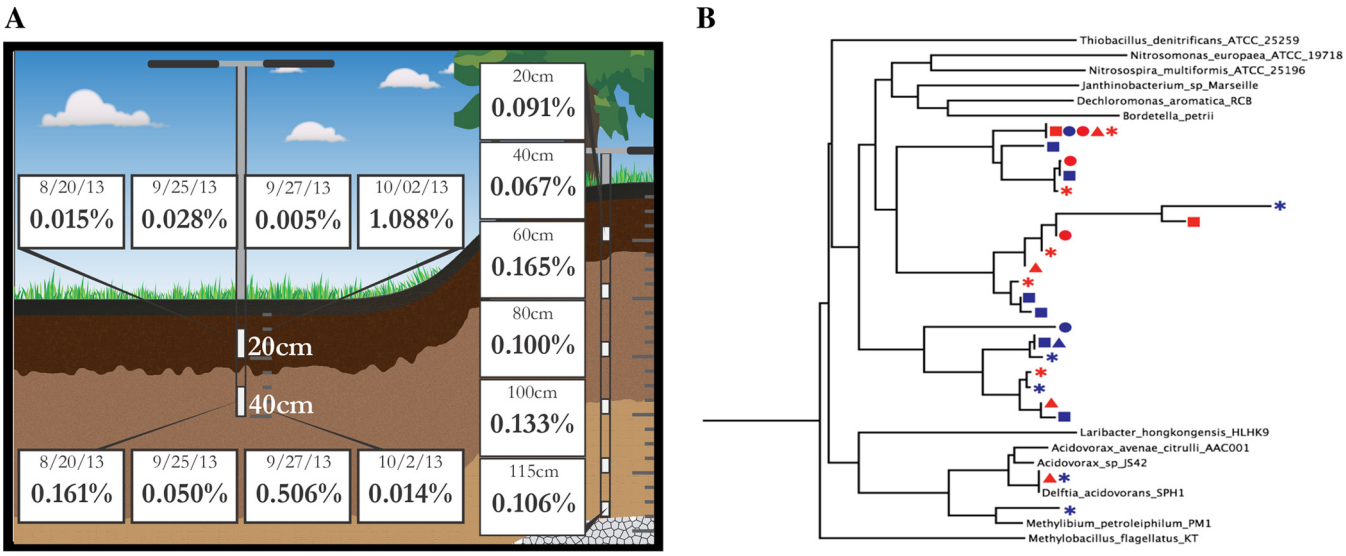

FIG 1 *D. acidovorans* is present in all of the soil and sediment samples sequenced, but the population structure is distinct from that of other bacteria. (A) The percentage of the reads from each sample that map to the *D. acidovorans* ANG1 genome. (B) An rps3 phylogenetic tree for a typical bacterial group (*Betaproteobacteria*) showing a dandelion-like pattern of strain diversity that is dissimilar from that of *D. acidovorans* (only the full-length assembled sequences are shown). Branches ending with a taxonomic identification are reference sequences, and the soil sequences are indicated by colored shapes representing their soil sample depth and time of origin around the first rainfall from August to October 2013 (10 to 20 cm, blue; 30 to 40 cm, red) before (squares) and after the rain events (4 days after the second rain, circles; 6 days after the second rain, triangles; 2 days after the second rain, asterisks).

previously sequenced *D. acidovorans* SPH-1 genome (GenBank accession no. NC_010002.1). The contigs spanned 97.95% of the SPH-1 genome, with an average nucleotide identity (ANI) of 99.99% (25). Because of the remarkably high nucleotide similarity between our recovered genome and that of strain SPH-1, we further curated the 16 contigs into one circular genome by using the SPH-1 genome to identify reads in our metagenomic data set that filled the gaps. When visualized with the software package Geneious (26), single nucleotide polymorphisms (SNPs) and insertions and deletions relative to the isolate genome sequence were easily identified (see Fig. S1). The circular closed genome represents the bacterial strain *D. acidovorans* ANG1.

Identification of *D. acidovorans* as a contaminant. On the basis of metagenomic read mapping, *D. acidovorans* ANG1 was detected in all 14 samples from two sites. No pattern between the sampling location and the abundance of *D. acidovorans* could be detected (Fig. 1A). On the basis of mapping of reads to the ANG1 genome, it was determined that the *D. acidovorans* reads in all of the samples derived from a single genotype (see Table S1). This result contrasts with our general findings from soil metagenomics, which typically indicate the presence of multiple closely related strains. For example, a phylogenetic tree constructed with rpS3 gene sequences shows dandelion-like strain diversity patterns in betaproteobacteria from the same samples (Fig. 1B). The lack of strain diversity in *D. acidovorans*, in combination with the high similarity of the *D. acidovorans* genome to that of the SPH-1 strain isolated years earlier, raised the possibility that *D. acidovorans* ANG1 was not a native member of the soil community. A PCR test with primers designed to target random regions of the *D. acidovorans* genome failed to detect DNA from this bacterium in the original DNA extracted from the samples (Fig. S2). This indicated that the DNA was probably introduced into our samples during sequencing. Moreover, the detection of the same genotype in samples sequenced at different times suggested that *D. acidovorans* was a persistent contaminant at this facility.

Source tracking of contaminant *D. acidovorans*. To further investigate the possibility that DNA from *D. acidovorans* ANG1 was introduced in the sequencing facility, we screened 43 publicly available metagenomes sequenced at this facility between June 2012 and January 2015. Seventeen of these projects had ≥20% of the *D. acidovorans* genome present, with coverage of >5× (see Table S2). Multiple strains were

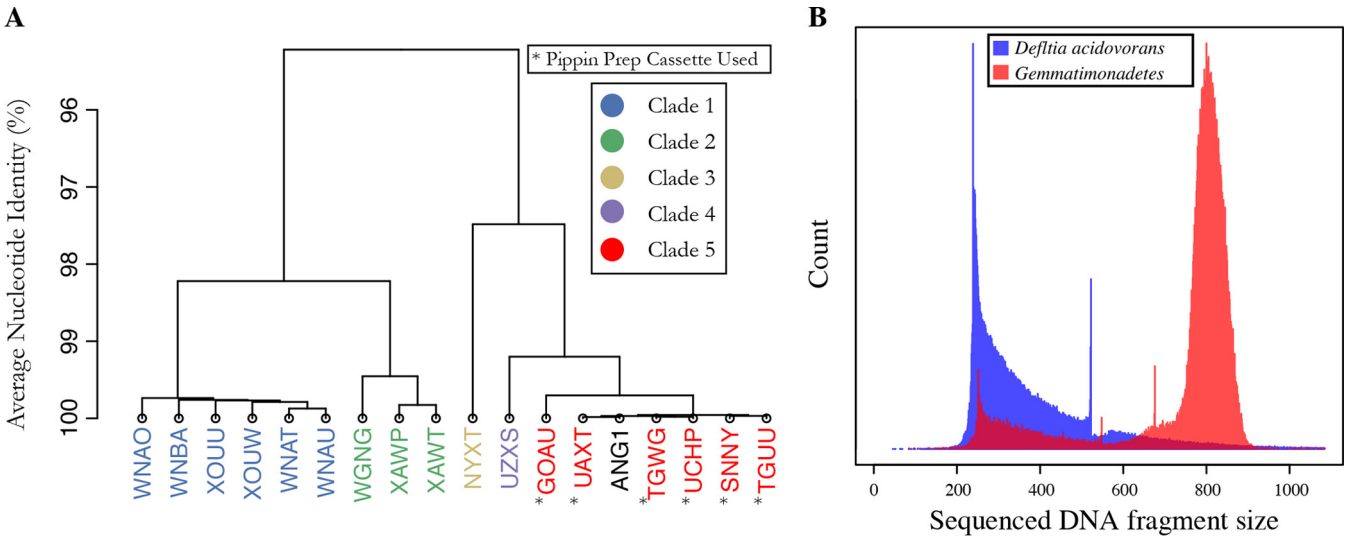

FIG 2 *D. acidovorans* contamination originates from library size selection cassettes. (A) Hierarchical clustering of *D. acidovorans* strains present in shotgun sequencing projects sequenced at the same facility as soil metagenomes in this study, on the basis of shared SNP frequencies. Samples prepared with Pippin Size Selection Cassettes (Sage Science) are marked with asterisks. Sequences of *D. acidovorans* in those samples form a monophyletic group that includes the ANG1 genome. (B) Histogram of sequenced DNA fragment sizes of metagenomic reads mapping to the recovered *D. acidovorans* genome and the second most abundant recovered genome in the sample, from a *Gemmatimonadetes* bacterium. The placements of reads mapping to the *Gemmatimonadetes* genome indicate the expected 800-bp fragment length, yet reads mapping to *D. acidovorans* show a different and skewed profile, consistent with the introduction of DNA after size selection.

present in these projects, on the basis of analysis of the patterns of SNPs relative to the *D. acidovorans* ANG1 genome (Fig. 2A; see Table S3). Six projects, including three plant genomes and our soil samples, contained sequences that clustered with the *D. acidovorans* ANG1 genome. We refer to this as the contaminant clade and hypothesized that some reagent in the sequencing pipeline may have been the source of the *D. acidovorans* DNA in these six projects. Records provided by the sequencing facility revealed that each project containing the *D. acidovorans* ANG1 contaminant clade used a Pippin Prep size selection cassette. All of the other projects analyzed did not use these cassettes (Fig. 2A).

To test the hypothesis that the Pippin Prep size selection cassettes were the source of *D. acidovorans* ANG1 contamination, we analyzed the insert sizes of reads from all of the projects mapped to the ANG1 genome. If the contaminant DNA was introduced from the library preparation cassette, it should have a more random fragment size profile, and thus insert size profile, than the tight size profile generated during library preparation. A histogram of reads mapping to the *D. acidovorans* ANG1 genome from a sample exhibiting substantial contamination is shown in Fig. 2B, along with a histogram of reads mapping to the genome of a *Gemmatimonadetes* strain known to be native to the same sample. The fragment size of paired reads mapped to the *Gemmatimonadetes* genome was ~800 bp. In contrast, read mapping to *D. acidovorans* indicates that many of the sequencing reads were generated from fragments of around 250 bp (in these cases, the 250-bp *D. acidovorans* reads overlap completely). However, the *D. acidovorans* peak is strongly skewed and some of the fragments were >1,000 bp in length. These observations indicate that the contaminant DNA was present in the gel and/or buffer used for size selection. All of the projects that contained *D. acidovorans* from the contaminant clade had similar insert size histograms, whereas projects with other *D. acidovorans* clades had normal insert sizes.

Sage Science, the producer of the Pippin size selection cassettes, was contacted regarding our observations. They explained that bacterial biofilms were present in tubing that delivered buffer to the cassettes and stated that the problem was corrected in 2013. All six projects that contain the contaminant clade used cassettes made prior to the correction, and libraries made with cassettes produced after Sage Science revised its manufacturing process to keep buffer tubing bacterium free did not reveal *D. aci-*

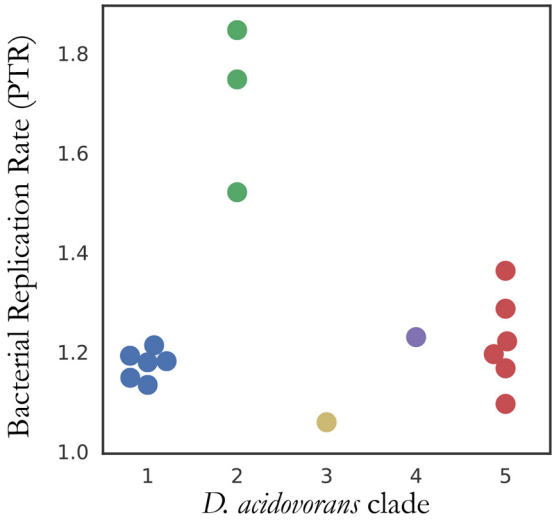

FIG 3 *D. acidovorans* strains native to a bioreactor have distinct replication rates. Clades of similar *D. acidovorans* genomes were clustered on the basis of shared SNP frequencies (Fig. 2A). The PTRs are significantly higher ($P = 0.005$) for the strains growing in a thiocyanate bioreactor (clade 2) than for strains growing in the other environments. As the other clades are likely made up of contaminants (see text), this observation supports the conclusion that the bioreactor strains are active community members and not contaminants.

dovorans ANG1 contamination (data not shown). The sequencing facility continued to detect *D. acidovorans* in samples sequenced after Sage Science corrected the problem and so concluded that sequences were not from the Pippin cassettes. We show here that the newly detected sequences were from the other clades, either strains actually present in the samples or contaminants from a different source. For example, clade 1 is associated with six metagenomes of the upper troposphere and clade 2 is associated with three metagenomes of thiocyanate bioreactor communities (see Table S2). In the case of clade 2, we conclude that *D. acidovorans* was native to the sample because the population was growing rapidly, on the basis of differential coverage at the origin and terminus of replication (peak-to-trough ratios [PTRs]) (full range, 1.5 to 1.9) (27, 28) (see Table S4). In contrast, bacteria of the contaminant clade (clade 5) had consistently low growth rates, based on PTRs between 1.4 and 1.1 (Fig. 3).

***In situ* evolution of a bacterial contaminant.** Fifteen large (>100-bp) insertions/deletions distinguished the *D. acidovorans* SPH-1 and ANG1 genomes (see Table S5). Four of these were insertions, and 11 were deletions (Fig. 4A). Specifically, two prophages were inserted into the ANG1 genome relative to the SPH-1 genome, but six prophages and five transposons were lost. Another difference involved three adjacent CRISPR repeat-spacer sequences. Finally, we identified a 195-bp insertion that added a pair of transmembrane helices that converted a predicted major facilitator superfamily protein into a predicted transporter that confers drug resistance.

We documented SNP-based genomic variation over the 1.5-year period to determine whether an evolutionary rate could be measured. A total of 203 SNPs distinguished the ANG1 genome from the *D. acidovorans* SPH-1 isolate genome (see Table S5). Thirty-seven SNPs were very closely spaced on the genome (Fig. 4A) and are statistically unlikely to have formed by individual random mutation events (see Materials and Methods). Specifically, 11 SNPs and four single-base-pair indels occurred in a 64-bp intergenic region upstream of an integrase and 24 SNPs occurred in a 447-bp region within a YD repeat-containing protein possibly involved in carbohydrate binding. We infer that these regions may have been acquired via homologous recombination with a very distinct genotype (<95% nucleotide identity); thus, we did not include them in our analysis of *in situ* genomic change via SNP formation.

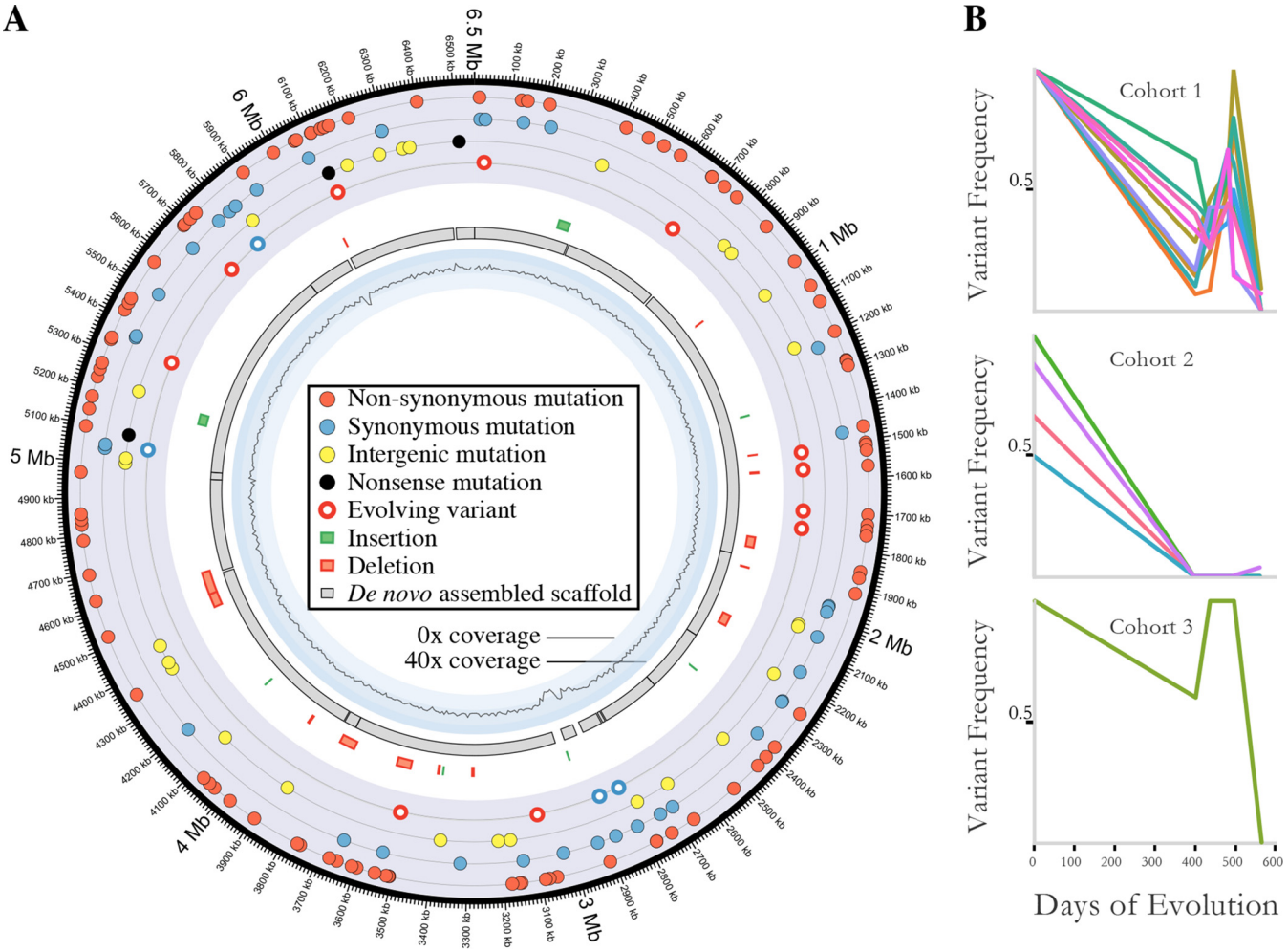

FIG 4 Detailed analysis of *D. acidovorans* sequence variation indicates the presence of subpopulations that are evolving *in situ*. (A) Shown are the locations of all of the differences between the contaminant *D. acidovorans* ANG1 genome recovered in this study and the *D. acidovorans* SPH-1 genome, as well as the alignment of the assembled contigs and coverage of the genome. (B) The frequencies of specific SNPs in samples sequenced at different time points exhibit three distinct patterns suggestive of three subpopulations (cohorts). Abundances fluctuate, but there is an overall progression toward fixation.

We identified 15 SNPs that distinguished the genotypes present in the first and last metagenomes (a separation time of ~1.5 years) and tracked their frequencies over time. Notably, the SNP frequencies cluster into three patterns, consistent with linkage and thus the existence of three subpopulations. Frequencies do not show a simple trend toward fixation, but all SNPs are fixed by the last time point (Fig. 4B) (see Table S6). Cohort three is defined by a single nonsynonymous mutation (Gly to Ser) that becomes fixed in a gene encoding a rod shape-determining protein.

By correcting for missing information due to lack of coverage at SNP sites in the ANG1 genome, we estimate that ~16 SNPs were fixed over the 1.5-year period. Thus, the rate of fixation is estimated as 10.1 SNPs/year. The remaining 150 SNPs that distinguish the SPH-1 genome sequenced in 2006 (S. Kjelleberg, personal communication) and ANG1 genomes likely arose between the time the SPH-1 genome was sequenced and the first metagenome time point (~6.5 years).

Given information about the SNP accumulation rate, we estimated how long ago the strain was separated from the original source culture. Assuming that the measured value approximates the rate of SNP formation over longer time periods, we calculated that the ANG1 genotype present in the soil was separated from the SPH-1 population 16.5 years ago.

We classified all 166 fixed SNPs that distinguish the SPH-1 and ANG1 genomes as synonymous, nonsynonymous, or intergenic (Fig. 4A). Overall, the *pN*/*pS* ratio is 0.98.

DISCUSSION

We reconstructed a complete *D. acidovorans* genome from a soil metagenome. Recovery of a complete genome from soil is an extremely unusual achievement, given that assembly becomes more difficult as sample complexity increases. We infer that this occurred in the present study because the genome was relatively abundant (~1% of the total DNA) and the population was nearly clonal, avoiding assembly problems that arise because of strain variation. The lack of microheterogeneity is atypical of soil populations. This and the uncanny similarity to an isolate genome raised the possibility that the genome derived from a contaminant, a hypothesis that was confirmed by PCR-based testing. The extremely high similarity between the contaminant *D. acidovorans* populations and the *D. acidovorans* SPH-1 genome gives us high confidence that one derived from the other. In this specific case, we could determine the source of the contaminant DNA. Consequently, this research serves as a case study of microbial metagenomics-based forensics. Importantly, we showed that a contaminant clade could be identified through base-pair-by-base-pair analysis of variable sites in population genomic data sets collected over time. This made it possible to distinguish strains that were native to the samples in which they occurred from those that were the contaminant. This approach should be generally applicable in strain-tracking investigations so long as high-quality genomes can be recovered.

Metagenomics-derived genomes have only very rarely been used to identify and track strains. The closest example to our work was a study by Loman et al. (29) that targeted pathogenic *Escherichia coli* in fecal samples and associated it with disease. However, their methods were specifically designed for the clinical setting and generated a draft genome that was too incomplete and fragmented for use in accurate strain tracking and *in situ* evolutionary analysis. Long stretches of contiguous sequence are needed to identify indels and regions associated with horizontal gene transfer, and higher genome completeness leads to more accurate estimates of evolutionary distance.

During analysis of the anthrax attack, the crux of the analysis related to determining the similarity between the weaponized strain and the Ames Laboratory strain (30, 31). As we show here, high-resolution determination of strain relatedness is possible given comprehensive comparative genomic information about the bacterium of interest. Given a reasonable estimate of the mutation rate and information about the genomic similarity between the weaponized strain and the Ames Laboratory strain, it would be possible to estimate how long ago the cultures were separated.

Most short-term evolution experiments performed to determine mutation rates are carried out under laboratory conditions. These rates may differ significantly from rates that are applicable under conditions experienced by populations growing in natural systems. Because we had access to data sets archived over a 1.5-year period, we could estimate the rate of evolution of the contaminant population in the relevant environment (tubing) and use this calibration to approximate the length of time separating it from the previously sequenced isolate. Our calculated value (10.1 substitutions/year) is very close to previously reported values of evolving pathogen populations (9.6 substitutions/year) (7) and orders of magnitude greater than estimates of natural *E. coli* populations (32). Using our calculated value, we estimate that separation of ANG1 from the source population occurred 16 years ago. SPH-1 was sequenced in 2006, and ANG1 was sequenced in 2014; hence, the longest time that the contaminant ANG1 population could have been separated from the SPH-1 isolate is 9 years. This result is within a factor of 2 of the expected value, despite the possibility of large errors in mutation accumulation rates due to variations in growth rates and stress (33). Thus, we conclude that the comparative genomics approach used here can constrain the time of separation of two populations. From a technical perspective, it is worth noting that this detailed analysis was successful even in a background of soil.

We defined three distinct SNP cohorts at time points intermediate between the first and last metagenomic sequencing events, suggesting the existence of three genotypic

variants. The fact that all 15 SNPs are fixed in the final ANG1 population is unexpected and inconsistent with separate coevolving populations. Further, the frequencies of SNPs in cohorts 1 and 3 undergo dramatic fluctuations in frequency (in some cases dropping below our detection level) rather than exhibit a simple trajectory toward fixation. We attribute both observations to extensive redistribution of SNPs through homologous recombination, a common process among closely related bacteria (34–36). The homogenization of population variation is consistent with recombination acting as a cohesive force, countering the diversifying SNP formation processes that otherwise could lead to speciation.

The original SPH-1 strain was isolated from municipal sewage sludge, an environment very different from the biofilm tubing environment in which the ANG1 strain was growing (37). The *pN*/*pS* ratio (the population-based equivalent to the *dN*/*dS* ratio [the ratio of nonsynonymous to synonymous evolutionary changes or substitutions]) of ~1, determined on the basis of a comparison of the isolate and the ANG1 population, is inconsistent with stabilizing selection. It could reflect minimal selective pressure or the combination of positive and negative selection. Most of the populations studied previously have values consistent with stabilizing selection (*dN*/*dS* ratios of ~0.1) because they are shaped by overall negative selection with some genes under positive selection (2, 38). Given this, we consider it more likely that the *D. acidovorans* population was experiencing a mixture of positive and negative selection rather than no selection. Positive selection is not surprising, given a population that is evolving and adapting to an environment different from that where it was isolated. An alternative explanation is that the nonsynonymous mutations have not yet had enough time to be selected against (39). The larger number of deletions (particularly of phage and transposon sequences) compared to insertions in the ANG1 relative to the SPH-1 genome is indicative of genome streamlining, consistent with its adaptation to a defined laboratory environment. Again, the observation of streamlining rather than genomic expansion is informative regarding the recent population history.

In conclusion, we show that detailed strain-resolved metagenomic studies can detect a specific organism of interest in very complex samples, provide evidence that the strain is not native to the environment from which the sample was collected, and constrain its recent history. We demonstrate this by using the example of a contaminant that was introduced during the laboratory handling of metagenomic samples, but the approach is far more broadly applicable. We used statistical analysis to link the contaminant genotypic group to its source and comparative genomics to uncover aspects of its recent evolutionary history. Because there was confirmation of many conclusions by independent methods, this work serves as a case study of strain-resolved forensic metagenomics.

MATERIALS AND METHODS

Sample collection. Soil collection and DNA extraction were performed as reported previously (24) and briefly described here. Soil samples were collected (with permission under application no. 27790) from the Angelo Coast Range Reserve meadow (39°44′21.4″N, 123°37′51.0″W) and from a nearby ridge within the Eel Critical Zone Observatory. Approximately 1 kg of soil was removed with sterilized stainless steel hand trowels at each depth. Samples were immediately flash frozen in a mixture of dry ice and ethanol and placed on dry ice for transport to the lab. DNA was extracted with Mo Bio Laboratories PowerMax Soil DNA Isolation kits from 10 g of soil from each depth. We optimized the protocol for our samples to maximize the DNA yield while minimizing shearing; each sample was vortexed for only 1 min, followed by a 30-min heating step at 65°C with inversion every 10 min. We performed two elution steps of 5 ml each and precipitated the DNA with sodium acetate and glycogen, resuspending it in 100 μl of 10 mM Tris buffer.

Metagenomic sequencing, assembly, and binning. Metagenomic DNA was sequenced at the Joint Genome Institute (JGI). DNA was sheared to 800 bp with the Covaris LE220 (Covaris) and size selected with the Pippin Prep (Sage Science). The fragments were treated by end repair, A tailing, and ligation of Illumina compatible adapters (IDT, Inc.) with the KAPA-Illumina library creation kit (KAPA Biosystems). Paired-end reads of 250 bp were generated on an Illumina Hiseq2500 with sequencing depth enumerated in Table S1. Reads were trimmed with Sickle (40) and assembled with IDBA-UD (41). Resulting scaffolds >1 kb in length were annotated with Prodigal (42) to predict open reading frames (ORFs) by using default metagenomic settings. Annotated protein sequences were searched against the KEGG (43), UniRef100 (44), and UniProt databases with USEARCH (45). All matches with bit scores of >60 were

saved, and reciprocal best hits with bit scores of >300 were also cataloged. We identified rRNA sequences with Infernal (46) by searching against databases from the SSU-Align package (47) and tRNAs with tRNAscan_SE (48). Genome binning was carried out with the online interface within ggKbase as described previously (49) (http://ggkbase.berkeley.edu/). This method takes into account the phylogenetic profile, GC content, and coverage information. The completeness of bacterial bins was evaluated on the basis of the presence or absence of single-copy genes. Shannon diversity calculations were performed as described previously (50) with assembled rps3 genes.

Genome curation. Once *D. acidovorans* scaffolds were binned, they were manually curated in order to close gaps between scaffolds. This was done within the Geneious software package version 8.1.5 (26). Project reads were first mapped back to the contigs with Bowtie 2 (51), and Geneious was used to extend the contigs. Contigs were next ordered and oriented by being mapped to the *D. acidovorans* SPH-1 reference genome with ABACAS (52). Areas of overlap between adjacent contigs were used to combine two contigs into one. To circularize the genome, reads were mapped directly to the reference genome. The reference genome was edited to be in agreement with the reads, and regions between contigs were filled in with corresponding pieces of the edited reference genome. Reads were finally mapped back to the circularized genome, and every base was manually inspected to ensure sequence read support. To catalog the differences between genomes, Mauve (53) was used to align the genomes and the alignment was manually inspected for mutations.

Phylogenetic tree. An alignment was generated with all of the rpS3 genes in the Angelo metagenomes, as well as previously published rpS3 sequences identified as similar to Angelo rpS3 sequences by BLAST. All rpS3 amino acid sequences longer than 180 amino acids were aligned with MUSCLE (54). The full alignments were stripped of columns containing 95% or more gaps. A maximum-likelihood phylogeny was inferred with RAxML (55) run with the PROTGAMMLG model of evolution. The RAxML interface included calculation of 100 bootstrap iterations (MRE [majority rules extended]-based bootstrapping criterion).

PCR testing. Primers amplifying randomly selected portions of the *D. acidovorans* genome were synthesized by IDT, Inc., to generate a 500-bp insert (forward, 5′GGGTTGCACCATTGGTATT; reverse, 5′GTCAGCGCCTTCTTTTCAA). Primers that amplify a 150-bp product of the 16S rRNA gene were also synthesized (forward, 5′GTGSTGCAYGGYTGTCGTCA; reverse, 5′ACGTCRTCCMCACCTTCCTC) (56). Pure *D. acidovorans* DNA was purchased from the DSMZ culture collection (DSMZ reference no. 14801). Reactions were performed with 5 PRIME MasterMix 50-μl reaction mixtures with 1 μl of each primer at 10 mM run in a ThermoCycler for 35 cycles with a melting temperature of 50°C and an extension time of 1 min 30 s. Both sets of primers and a no-primer control were run on (i) 0.05 ng of *D. acidovorans* DNA, (ii) 4.95 ng of original extracted soil DNA (sample 13_1_20cm_4), and (iii) both of the above combined into a single reaction mixture.

Paired read insert length profiling. Reads were mapped from each analyzed project to the *D. acidovorans* SPH-1 genome with Bowtie 2 (51). For comparison, reads from project 13_1_20cm_4 were also mapped to the second most abundant bacterial genome, that of a bacterium belonging to the phylum *Gemmatimonadetes*. The resulting .sam file was converted to a sorted .Bam file with SAMtools (57), and the insert size was profiled with Picard (58) (command, Java –jar Picard_tools/CollectInsertSizeMetrics.jar MINIMUM_PCT=0.4).

Comparison of *D. acidovorans* strains. To determine if soil and sediment metagenomes contained the *D. acidovorans* strain ANG1 genome, PileupProfile.py (source code available at https://github.com/banfieldlab/mattolm-public-scripts) Was used to calculate the ANI of reads mapping to the ANG1 genome. ANI was at least 99.9% in all cases where the median coverage was ≥2 (see Table S1). To compare strains present in 46 public metagenomic projects sequenced at the JGI that had been flagged as possibly containing *D. acidovorans* contamination, reads were mapped to the *D. acidovorans* SPH-1 genome with Bowtie 2 (51). Projects with at least 20% of the bases having 5× coverage were said to have significant *D. acidovorans* present and were analyzed in more detail. To compare the strains in each metagenome, the custom script ReadComparer.py was used (source code available at https://github.com/banfieldlab/mattolm-public-scripts). Briefly, the program first aligns reads from all projects to the same genome and compares the mutational patterns between projects to determine the relatedness of strains. VarScan (59) was used to create the input files (command, java -jar VarScan.v2.3.8.jar pileup2cns −min-coverage 3), and the script was run with the command, ReadComparer.py −min-_breadth 0.2 −matrix −dend −smart_ignore. The similarity matrix was then plotted into a dendrogram with R, and clusters were determined and colored on the basis of manual inspection of the resulting dendrogram.

Mutation identification and profiling. A number of methods were used to identify differences between *D. acidovorans* strains SPH-1 and ANG1. Mutations were identified with breseq (60) mapping reads from project 13_1_20cm_4 to the *D. acidovorans* SPH-1 genome. To identify larger indels, *D. acidovorans* genomes were aligned with Mauve (53) and the alignment was manually inspected for differences. Finally, VarScan (59) was used on reads mapping from 13_1_20cm_4 to the *D. acidovorans* ANG1 genome. All called mutations were manually verified by inspection of the region with the software package Geneious (26).

To track the frequency of mutations among all projects of the contaminant clade, polymorpher2.py (available at https://github.com/banfieldlab/mattolm-public-scripts) was used to determine the frequency of each base at all of the positions identified. Variants that were <315 bp apart (a 1% chance of occurring assuming a random distribution of variants) were excluded from mutation rate calculations on the basis of the assumption that they were likely acquired in a recombination event. Positions were required to have some coverage in all projects, which excluded 4.4% of the variants from the analysis.

Polymorphisms with a frequency of at least 50% at the first time point that became nearly fixed at the last time point (≤0.1%) were said to be evolving and were used in rate calculations (including a correction to account for variants without sufficient coverage for analysis). Full calculation details are available in Data Set S1.

ORFs predicted by Prodigal were used to determine the total number of synonymous and nonsynonymous sites in the genome, as well as to classify each variant as synonymous, nonsynonymous, or intergenic. Genome-wide *pN*/*pS* ratios were calculated with the formula (number of nonsynonymous substitutions/number of nonsynonymous sites)/(number of synonymous substitutions/number of synonymous sites).

Data availability. Raw reads for all of the soil and sediment metagenomes in this study are available at the JGI Genome Portal (JGI Proposal ID, 1430). Accession numbers for specific samples are in Table S1. The source code of custom scripts used in data analysis (polymorpher2.py, PileupProfile.py, and ReadComparer.py) is available at https://github.com/banfieldlab/mattolm-public-scripts. The complete *D. acidovorans* ANG1 genome sequence is in the NCBI GenBank database under accession no. CP019171, BioProject no. PRJNA297196.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mBio.01969-16.

FIG S1, PDF file, 0.4 MB.
FIG S2, PDF file, 0.5 MB.
DATA SET S1, TXT file, 0.2 MB.
TABLE S1, XLSX file, 0.04 MB.
TABLE S2, XLSX file, 0.04 MB.
TABLE S3, XLSX file, 0.1 MB.
TABLE S4, XLSX file, 0.03 MB.
TABLE S5, XLSX file, 0.03 MB.
TABLE S6, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

We thank Susan Spaulding for sample collection and for performing the DNA extractions, Alex Spunde for access to JGI metagenomes, Brandon Brooks for help with PCR testing, and Sage Science for providing information needed to complete this study.

This work was supported by the Office of Science, Office of Biological and Environmental Research, U. S. Department of Energy (grant DOE-SC10010566). The sequencing was conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, and Lawrence Berkeley National Laboratory under contract DE-AC02-05CH11231. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under grant no. DGE 1106400.

REFERENCES

1. Greenblum S, Carr R, Borenstein E. 2015. Extensive strain-level copy-number variation across human gut Microbiome species. Cell 160:583–594. https://doi.org/10.1016/j.cell.2014.12.038.
2. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, Kota K, Sunyaev SR, Weinstock GM, Bork P. 2013. Genomic variation landscape of the human gut microbiome. Nature 493:45–50. https://doi.org/10.1038/nature11711.
3. Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. 2015. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol 33:1045–1052. https://doi.org/10.1038/nbt.3319.
4. Hagedorn C, Blanch AR, Harwood VJ. 2011. Microbial source tracking: methods, applications, and case studies. Springer Science and Business Media, New York, NY.
5. Harwood VJ, Staley C, Badgley BD, Borges K, Korajkic A. 2014. Microbial source tracking markers for detection of fecal contamination in environmental waters: relationships between pathogens and human health outcomes. FEMS Microbiol Rev 38:1–40. https://doi.org/10.1111/1574-6976.12031.
6. Snitkin ES, Zelazny AM, Thomas PJ, Stock F, Henderson DK, Palmore TN, Segre JA, NISC Comparative Sequencing Program Group. 2012. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Sci Transl Med 4:148ra116–148ra116. https://doi.org/10.1126/scitranslmed.3004129.
7. Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, Chantratita N, Gardete S, Tavares A, Day N, Lindsay JA, Edgeworth JD, de Lencastre H, Parkhill J, Peacock SJ, Bentley SD. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. Science 327:469–474. https://doi.org/10.1126/science.1182395.
8. Federal Bureau Investigation. 2016. Amerithrax or anthrax investigation. Federal Bureau Investigation, Washington, DC. https://www.fbi.gov/history/famous-cases/amerithrax-or-anthrax-investigation.
9. McGhee GC, Sundin GW. 2012. Erwinia amylovora CRISPR elements provide new tools for evaluating strain diversity and for microbial source tracking. PLoS One 7:e41706. https://doi.org/10.1371/journal.pone.0041706.
10. Budowle B, Connell ND, Bielecka-Oder A, Colwell RR, Corbett CR, Fletcher J, Forsman M, Kadavy DR, Markotic A, Morse SA, Murch RS, Sajantila A, Schmedes SE, Ternus KL, Turner SD, Minot S. 2014. Validation of high throughput sequencing and microbial forensics applications. Invest Genet 5:9.
11. Chan MS, Maiden MC, Spratt BG. 2001. Database-driven multi locus sequence typing (MLST) of bacterial pathogens. Bioinformatics 17:1077–1083. https://doi.org/10.1093/bioinformatics/17.11.1077.
12. Olive DM, Bean P. 1999. Principles and applications of methods for DNA-based typing of microbial organisms. J Clin Microbiol 37:1661–1669.

13. Gilchrist CA, Turner SD, Riley MF, Petri WA, Hewlett EL. 2015. Whole-genome sequencing in outbreak analysis. Clin Microbiol Rev 28:541–563. https://doi.org/10.1128/CMR.00075-13.
14. National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications. 2007. The new science of metagenomics: revealing the secrets of our microbial planet. National Academies Press, Washington, DC.
15. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523:208–211. https://doi.org/10.1038/nature14486.
16. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun 7:13219. https://doi.org/10.1038/ncomms13219.
17. Lee ST, Kahn SA, Delmont TO, Hubert NA, Morrison HG, Antonopoulos DA, Rubin DT, Eren AM. 2 December 2016. High-resolution tracking of microbial colonization in fecal microbiota transplantation experiments via metagenome-assembled genomes. bioRxiv. http://biorxiv.org/content/early/2016/12/02/090993.
18. Dombrowski N, Donaho JA, Gutierrez T, Seitz KW, Teske AP, Baker BJ. 2016. Reconstructing metabolic pathways of hydrocarbon-degrading bacteria from the Deepwater Horizon oil spill. Nat Microbiol 1:16057. https://doi.org/10.1038/nmicrobiol.2016.57.
19. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12:87. https://doi.org/10.1186/s12915-014-0087-z.
20. Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, Bushman FD, Knight R, Kelley ST. 2011. Bayesian community-wide culture-independent microbial source tracking. Nat Methods 8:761–763. https://doi.org/10.1038/nmeth.1650.
21. Fierer N, Ladau J, Clemente JC, Leff JW, Owens SM, Pollard KS, Knight R, Gilbert JA, McCulley RL. 2013. Reconstructing the microbial diversity and function of Pre-Agricultural Tallgrass prairie soils in the United States. Science 342:621–624. https://doi.org/10.1126/science.1243768.
22. Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, Brown CT. 2014. Tackling soil diversity with the assembly of large, complex metagenomes. Proc Natl Acad Sci U S A 111:4904–4909. https://doi.org/10.1073/pnas.1402564111.
23. Williamson KE, Radosevich M, Wommack KE. 2005. Abundance and diversity of viruses in six Delaware soils. Appl Environ Microbiol 71:3119–3125. https://doi.org/10.1128/AEM.71.6.3119-3125.2005.
24. Butterfield C, Zhou L, Andeer P, Spaulding S, Thomas B, Andrea S, Hettich R, Suttle K, Probst A, Tringe S, Northen T, Pan C, Banfield J. 2016. Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. PeerJ 4:e2687.
25. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol 57:81–91. https://doi.org/10.1099/ijs.0.64483-0.
26. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649. https://doi.org/10.1093/bioinformatics/bts199.
27. Korem T, Zeevi D, Suez J, Weinberger A, Avnit-Sagi T, Pompan-Lotan M, Matot E, Jona G, Harmelin A, Cohen N, Sirota-Madi A, Thaiss CA, Pevsner-Fischer M, Sorek R, Xavier RJ, Elinav E, Segal E. 2015. Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. Science 349:1101–1106. https://doi.org/10.1126/science.aac4812.
28. Brown CT, Olm MR, Thomas BC, Banfield JF. 16 June 2016. In situ replication rates for uncultivated bacteria in microbial communities. bioRxiv. http://biorxiv.org/content/biorxiv/early/2016/06/16/057992.full.pdf.
29. Loman NJ, Constantinidou C, Christner M, Rohde H, Chan JZ-m, Quick J, Weir JC, Quince C, Smith GP, Betley JR, Aepfelbacher M, Pallen MJ. 2013. A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. JAMA 309:1502–1510. https://doi.org/10.1001/jama.2013.3231.
30. Hoffmaster AR, Fitzgerald CC, Ribot E, Mayer LW, Popovic T. 2002. Molecular subtyping of Bacillus anthracis and the 2001 bioterrorism-associated anthrax outbreak, United States. Emerg Infect Dis 8:1111–1116. https://doi.org/10.3201/eid0810.020394.
31. Rasko DA, Worsham PL, Abshire TG, Stanley ST, Bannan JD, Wilson MR, Langham RJ, Decker RS, Jiang L, Read TD, Phillippy AM, Salzberg SL, Pop M, Van Ert MN, Kenefic LJ, Keim PS, Fraser-Liggett CM, Ravel J. 2011. Bacillus anthracis comparative genome analysis in support of the Amerithrax investigation. Proc Natl Acad Sci U S A 108:5027–5032. https://doi.org/10.1073/pnas.1016657108.
32. Ochman H, Elwyn S, Moran NA. 1999. Calibrating bacterial evolution. Proc Natl Acad Sci U S A 96:12638–12643. https://doi.org/10.1073/pnas.96.22.12638.
33. Bjedov I, Tenaillon O, Gérard B, Souza V, Denamur E, Radman M, Taddei F, Matic I. 2003. Stress-induced mutagenesis in bacteria. Science 300:1404–1409. https://doi.org/10.1126/science.1082240.
34. Tenaillon O, Skurnik D, Picard B, Denamur E. 2010. The population genetics of commensal Escherichia coli. Nat Rev Microbiol 8:207–217. https://doi.org/10.1038/nrmicro2298.
35. Rosen MJ, Davison M, Bhaya D, Fisher DS. 2015. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science 348:1019–1023. https://doi.org/10.1126/science.aaa4456.
36. Smillie CS, Smith MB, Friedman J, Cordero OX, David LA, Alm EJ. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. Nature 480:241–244. https://doi.org/10.1038/nature10571.
37. Schleheck D, Knepper TP, Fischer K, Cook AM. 2004. Mineralization of individual congeners of linear alkylbenzenesulfonate by defined pairs of heterotrophic bacteria. Appl Environ Microbiol 70:4053–4063. https://doi.org/10.1128/AEM.70.7.4053-4063.2004.
38. Friedman R, Drake JW, Hughes AL. 2004. Genome-wide patterns of nucleotide substitution reveal stringent functional constraints on the protein sequences of thermophiles. Genetics 167:1507–1512. https://doi.org/10.1534/genetics.104.026344.
39. Rocha EPC, Smith JM, Hurst LD, Holden MTG, Cooper JE, Smith NH, Feil EJ. 2006. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol 239:226–235. https://doi.org/10.1016/j.jtbi.2005.08.037.
40. Joshi N, Fass J. 2011. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files. https://github.com/najoshi/sickle.
41. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28:1420–1428. https://doi.org/10.1093/bioinformatics/bts174.
42. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
43. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42:D199–D205. https://doi.org/10.1093/nar/gkt1076.
44. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 23:1282–1288. https://doi.org/10.1093/bioinformatics/btm098.
45. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461. https://doi.org/10.1093/bioinformatics/btq461.
46. Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29:2933–2935. https://doi.org/10.1093/bioinformatics/btt509.
47. Nawrocki E. 2009. Structural RNA homology search and alignment using covariance models. Ph.D. dissertation. Washington University, St. Louis, MO.
48. Lowe TM, Eddy SR. 1997. TRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964. https://doi.org/10.1093/nar/25.5.0955.
49. Raveh-Sadka T, Thomas BC, Singh A, Firek B, Brooks B, Castelle CJ, Sharon I, Baker R, Good M, Morowitz MJ, Banfield JF. 2015. Gut bacteria are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis development. eLife 4:e05477. https://doi.org/10.7554/eLife.05477.
50. Spellerberg IF, Fedor PJ. 2003. A tribute to Claude Shannon (1916–2001)

and a plea for more rigorous use of species richness, species diversity and the "Shannon-Wiener" index. Glob Ecol Biogeogr 12:177–179.
51. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.
52. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. 2009. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25:1968–1969. https://doi.org/10.1093/bioinformatics/btp347.
53. Rissman AI, Mau B, Biehl BS, Darling AE, Glasner JD, Perna NT. 2009. Reordering contigs of draft genomes using the Mauve aligner. Bioinformatics 25:2071–2073. https://doi.org/10.1093/bioinformatics/btp356.
54. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340.
55. Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690. https://doi.org/10.1093/bioinformatics/btl446.
56. Horz HP, Vianna ME, Gomes BPFA, Conrads G. 2005. Evaluation of universal probes and primer sets for assessing total bacterial load in clinical samples: general implications and practical use in endodontic antimicrobial therapy. J Clin Microbiol 43:5332–5337. https://doi.org/10.1128/JCM.43.10.5332-5337.2005.
57. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
58. Broad Institute. 2009. Picard Tools. Broad Institute, Cambridge, MA. http://broadinstitute.github.io/picard/.
59. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics 25:2283–2285. https://doi.org/10.1093/bioinformatics/btp373.
60. Deatherage DE, Barrick JE. 2014. Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. Methods Mol Biol 1151:165–188. https://doi.org/10.1007/978-1-4939-0554-6_12.