

# UC Merced

## UC Merced Previously Published Works

### Title

scMicrobe PTA: near complete genomes from single bacterial cells

### Permalink

<https://escholarship.org/uc/item/1tc5h5jn>

### Journal

ISME Communications, 4(1)

### ISSN

2730-6151

### Authors

Bowers, Robert M

Gonzalez-Pena, Veronica

Wardhani, Kartika

et al.

### Publication Date

2024-01-08

### DOI

10.1093/ismeco/ycae085

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

# scMicrobe PTA: near complete genomes from single bacterial cells

Robert M. Bowers<sup>1,†</sup>, Veronica Gonzalez-Pena<sup>2,†</sup>, Kartika Wardhani<sup>2</sup>, Danielle Goudeau<sup>1</sup>, Matthew James Blow<sup>1</sup>, Daniel Udvary<sup>1</sup>, David Klein<sup>2</sup>, Albert C. Vill<sup>3</sup>, Ilana L. Brito<sup>3</sup>, Tanja Woyke<sup>1</sup>, Rex R. Malmstrom<sup>1,\*,†</sup>, Charles Gawad<sup>2,3,\*,†</sup>

<sup>1</sup>DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, United States

<sup>2</sup>Department of Pediatrics, Stanford University, Stanford, CA, United States

<sup>3</sup>Chan Zuckerberg Biohub, San Francisco, CA, United States

<sup>4</sup>Meinig School of Biomedical Engineering, Cornell University, Ithaca, NY, United States

\*Corresponding author: Charles Gawad, Stanford University, Chan Zuckerberg Biohub Investigator, Biomedical Innovation, BMI-2200, 240 Pasteur Drive, Stanford, CA 94305, United States. Email: cgawad@stanford.edu; Rex Malmstrom, DOE Joint Genome Institute, Microscale Applications Group Lead, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Bldg 91, Berkeley, CA 94720, United States. Email: rrmalmstrom@lbl.gov

†Robert Bowers, Veronica Gonzalez-Pena, Charles Gawad and Rex Malmstrom contributed equally to this work.

## Abstract

Microbial genomes produced by standard single-cell amplification methods are largely incomplete. Here, we show that primary template-directed amplification (PTA), a novel single-cell amplification technique, generated nearly complete genomes from three bacterial isolate species. Furthermore, taxonomically diverse genomes recovered from aquatic and soil microbiomes using PTA had a median completeness of 81%, whereas genomes from standard multiple displacement amplification-based approaches were usually <30% complete. PTA-derived genomes also included more associated viruses and biosynthetic gene clusters.

**Keywords:** single cell genomics, unculturable bacteria, microbiome, microbial ecology, primary template-directed amplification, multiple displacement amplification, whole genome amplification, microbiome sequencing, mobile genetic elements, biosynthetic gene clusters

## Introduction

Difficulties in cultivating most bacterial and archaeal species presents a barrier to exploring the genetic make-up of the Earth's microbiomes. To access the genomes of most microorganisms, culture-independent methods such as shotgun metagenomic sequencing [1–3] and single-cell sequencing [4–8] can be employed. While metagenomics has led to unprecedented insights into the metabolic potential of uncultured microorganisms [9–12], the approach has some limitations. For example, it is difficult to connect mobile genetic elements such as plasmids and phages to metagenome-assembled genomes (MAGs) [13]. Generating MAGs from heterogeneous or low abundance populations is also challenging [14, 15]. Single-cell sequencing, in contrast, does not share these same limitations [5], and the approach has provided insights into microbial dark matter [4, 7], experimentally linked phages to their hosts [16, 17], and dissected natural populations [13, 18, 19]. However, multiple displacement amplification (MDA)—the predominant single-cell genome amplification method [20]—is limited by the poor uniformity and completeness of the genomes it produces [21]. Single-cell amplified genomes (SAGs) typically have genome completeness  $\leq 40\%$  [4].

Different variations on genome amplification chemistry [22–24] and sample processing strategies [25–31] have improved

genome recovery in some situations, but an approach for consistently generating complete or nearly complete genomes from single microbial cells is still lacking. We recently developed primary template-directed amplification (PTA), which significantly improves amplified genome uniformity and variant calling in single human cells [32]. Here, we investigated whether PTA could also improve the quality of genomes recovered from single bacterial cells.

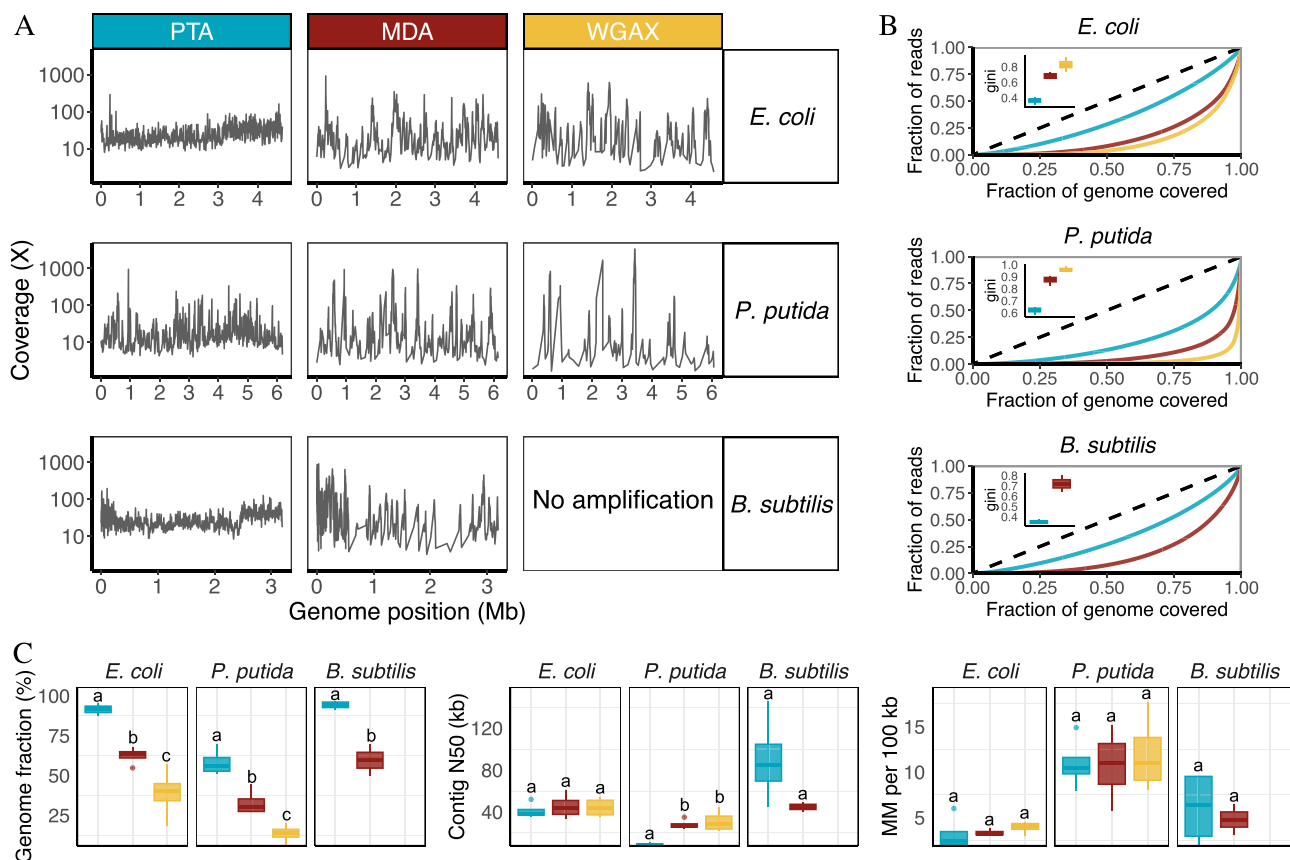
To benchmark PTA performance against the genome amplification chemistries commonly used in microbiome studies, we first sequenced the genomes of three bacterial isolate species: *Escherichia coli* (Gram-, GC%=51%), *Pseudomonas putida* (Gram-, GC%=62%), and *Bacillus subtilis* (Gram+, GC=43%). Individual cells were sorted into 96-well plates using fluorescence activated cell sorting (FACS), and replicate plates were subjected to genome amplification using PTA, MDA, and whole-genome amplification (WGA)-X, a modified version of MDA that uses a more thermostable variant of phi29 polymerase [23] (Supplementary Table S1). Sequencing reads were mapped to reference genomes to measure coverage uniformity, and later assembled de novo using SPAdes [33]. All libraries were subsampled to 1 M reads (150 Mbp) prior to these analyses to ensure comparable sequencing effort among SAGs.

In every case, genome coverages from PTA reactions were significantly more uniform than MDA and WGA-X reactions based

Received: 4 March 2024. Revised: 5 June 2024

© The Author(s) 2024. Published by Oxford University Press on behalf of the International Society for Microbial Ecology.

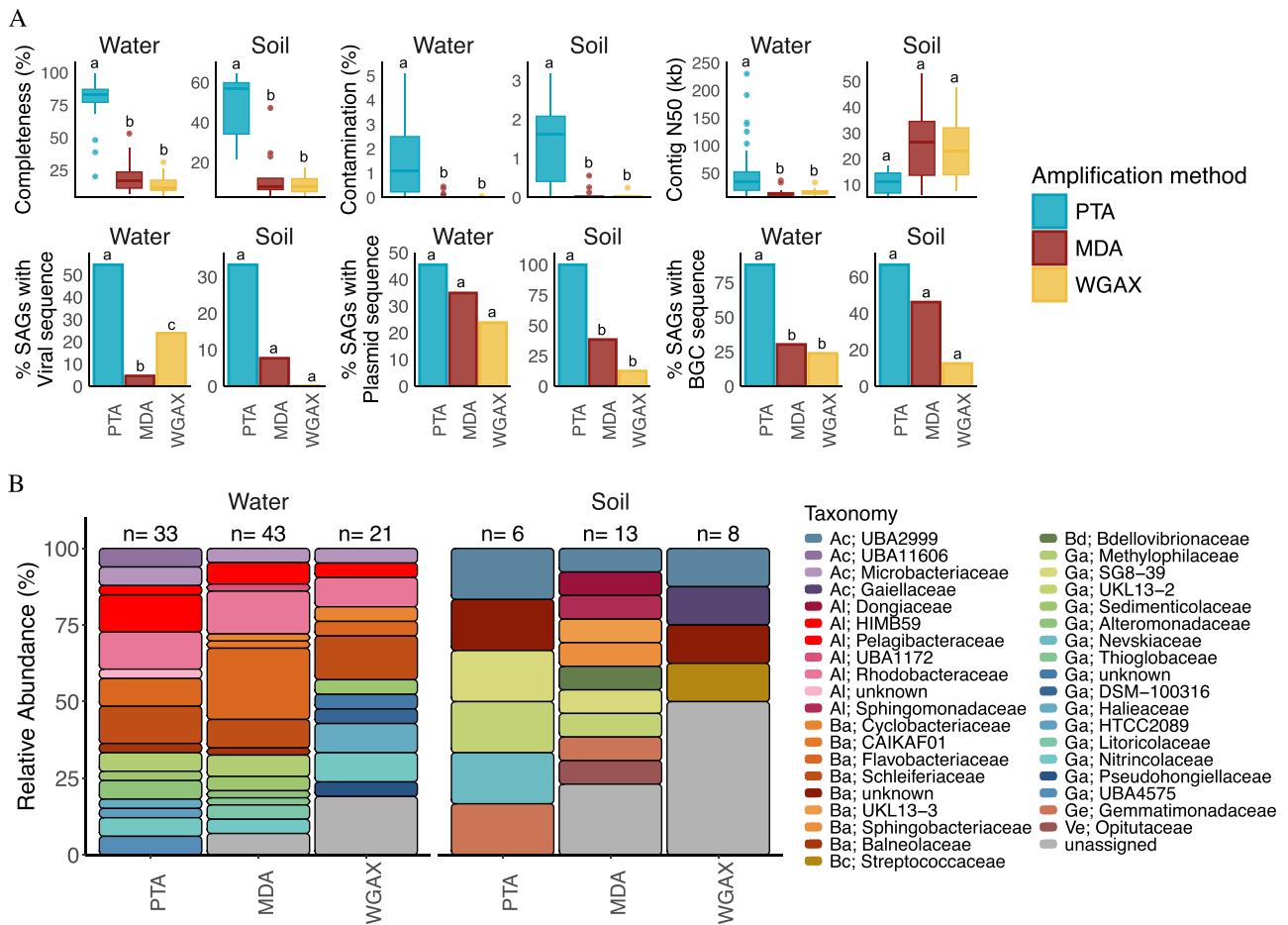
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.



**Figure 1. Genome quality of *E. coli*, *P. putida*, and *B. subtilis* SAGs amplified using PTA (blue), MDA (red), and WGA-X (yellow). A)** Genome coverage of 500 bp windows from one representative replicate of each species amplified with each chemistry. WGA-X amplification reactions of *B. subtilis* failed and were not repeated. Refer to [Supplementary Fig. 1](#) for genome coverage plots of all replicates. **B)** Uniformity of genome coverage illustrated by Lorenz curves and Gini coefficients. The dotted line represents the expected pattern of perfect uniform coverage, and solid lines illustrate the observed coverage for representative cells. **C)** Key summary statistics of de novo genome assemblies including completeness, contig N50, and the number of mismatches (MM) per 100 kb. The letters a, b, and c above the boxplots denote significance at the alpha 0.05 level. Sample sizes are  $n = 4$  for all species and chemistries except for MDA amplified *B. subtilis*, which had  $n = 2$ . The boxplot dots represent outliers that are beyond the 1.5-fold interquartile range. Additional summary statistics are reported in [Supplementary Fig. 2](#) and [Supplementary Tables S2](#) and [S3](#).

on Lorenz curves and Gini coefficients (Fig. 1 and [Supplementary Table S2](#);  $P < .01$  one way Analysis of Variance (ANOVA) and Tukey HSD for *E. coli* and *P. putida*;  $P < .01$  one way t-test for *B. subtilis*). In addition, PTA amplification resulted in significantly greater genome completeness than did MDA and WGA-X for all three species (Fig. 1C and [Supplementary Table S2](#);  $P < .01$  one way ANOVA and Tukey HSD for *E. coli* and *P. putida*;  $P < .01$  one way t-test for *B. subtilis*). For example, *B. subtilis* and *E. coli* SAGs assembled de novo had an average completeness of 94% and 91%, respectively, whereas genomes generated by MDA recovered only 60% and 62% on average ([Supplementary Table S3](#)). *P. putida* SAGs were less complete for all chemistries, but genomes generated by PTA were nearly 2-fold more complete than those generated by MDA and WGA-X. *P. putida* genome completeness improved to 91% for PTA chemistry after increasing the number of input reads to an average of 4 M (600 Mbp); *P. putida* genomes amplified by MDA and WGA-X were not sequenced deeply enough to assess completeness at this higher coverage level. PTA also showed similar fidelity to MDA and WGA-X when copying the genomes, e.g. no significant difference in genome mismatch rates per 100 kilobases among amplification chemistries (Fig. 1C and [Supplementary Table S2](#);  $P > .05$  one-way ANOVA). Overall, these results mirror the superior performance of PTA versus MDA and other genome amplification strategies observed previously using human cells [32].

After performing these benchmarking experiments with bacterial isolates, we sought to determine if the improved performance of PTA could be extended to environmental samples. To accomplish this, we utilized the same comparison strategy to amplify and sequence single cells recovered by FACS from aquatic and soil samples. We again found that PTA resulted in significantly greater genome completeness than MDA and WGA-X (Fig. 2A and [Supplementary Table S4](#);  $P < .01$  one way ANOVA and Tukey HSD). For example, PTA reactions from aquatic samples had median genome completeness of 83%, while completeness from MDA and WGA-X reactions had medians of 17% and 11%, respectively (Fig. 2A and [Supplementary Table S5](#)). Deeper sequencing of MDA and WGA-X libraries to  $\sim 20$  M reads (3 Gbp) increased median completeness estimates to 30% and 23%, respectively ([Supplementary Table S5](#)), but these genomes were still far less complete than those derived from PTA reactions ( $P < .01$  one way ANOVA and Tukey HSD). Similar patterns were observed from a smaller soil microbiome dataset where PTA produced genomes with much greater completeness than MDA and WGA-X (Fig. 2A and [Supplementary Table S4](#);  $P < .01$  one way ANOVA and Tukey HSD). Additionally, a larger fraction of PTA genomes recovered from the aquatic system had virus and biosynthetic gene clusters (BGC) sequences, and a larger fraction of PTA genomes from soil had plasmid sequences (Fig. 2A and [Supplementary Table S4](#);



**Figure 2. Comparison of SAGs from aquatic and soil microbiomes amplified with PTA (blue), MDA (red), and WGA-X (yellow). A)** Estimated genome completeness and contamination using CheckM2 [34], contig N50, and the percentage of SAGs containing at least one predicted plasmid (> 5 kb), virus (> 5 kb), or BGC. The letters a, b, and c denote significance at the alpha 0.05 level. **B)** Family level taxonomic assignment of SAGs assembled from  $\leq 20$  Million reads. Phylum / class abbreviations are as follows: AC: Acidobacteria, AL: Alphaproteobacteria, BA: Bacteroidota, BD: Bdellovibrionota, BC: Bacillota, GA: Gammaproteobacteria, GE: Gemmatimonadota, VE: Verrucomicrobiota.

$P < .05$  Fisher's exact test), presumably because PTA genomes were more complete than MDA and WGA-X genomes. Finally, phylogenetic analysis revealed successful PTA reactions on cells belonging to 20 families spread across 6 phyla (Fig. 2B), suggesting that PTA is amenable to a wide variety of microorganisms and produces substantially more near-complete genomes than standard amplification chemistries used in microbiome studies (Fig. 2 and Supplementary Fig. 3).

For single-cell genomes, overall genome quality is measured by a combination of completeness and contamination, with "high-quality" genomes defined as having >90% completeness and < 5% contamination [35]. Environmental genomes generated by PTA had a median contamination of 1.5% after applying an informatic decontamination procedure. Some contaminating sequences were found across SAGs and in no-template controls, suggesting that the PTA reagents contained trace levels of contaminating DNA. Single-cell whole genome amplification chemistries use short random primers to amplify a few femtograms of DNA, so even trace amounts of contaminating DNA can appear in assemblies. To decrease contaminating DNA, MDA, and WGA-X reagents underwent secondary treatment with UV prior to genome amplification [36]. PTA reagents underwent UV decontamination during manufacturing, but no secondary UV treatment of reagents or consumables was done prior to PTA reactions which may explain the higher contamination levels. Treating PTA reagents with an additional dose of UV may reduce

contamination in future studies [28, 37]. Reducing reaction volume could also lower nonspecific synthesis from contaminant DNA and would lower reagent consumption and costs [28, 31].

In summary, our data shows that 78% of environmental SAGs produced with PTA were medium- or high-quality genome drafts (median completeness of 81% and 1.5% contamination), while >90% of the SAG produced with MDA and WGAX were classified as low-quality (median completeness <30%, and < 0.1% contamination) as defined by the Minimum Information about Single Amplified Genome standards [35] (Supplementary Fig. 3). Although SAGs from all chemistries were fragmented into many contigs—an on-going challenge for single-cell and metagenomic assembly—PTA often enabled recovery of nearly complete genomes from individual bacteria. We believe that these results set the stage for a renaissance in single-cell-based environmental genomics by offering a more comprehensive insight into the population structure of the microbial dark matter that accounts for a large fraction of the Earth's biomass.

## Materials and methods

### Sample collection and processing

Fresh cultures of *E. coli* MG1655, *P. putida* KT2440, and *B. subtilis* pDR244 were grown overnight in LB at 37°C, then used immediately for cell sorting as described below. In addition, a 50 ml aquatic sample was collected from the surface

waters of Mountain View Slough (latitude 37.432400, longitude -122.086632). The sample was vortexed for 15 s to release cells attached to sediment, filtered using a 15  $\mu\text{m}$  cell strainer (pluriStrainer from pluriSelect, Germany) to remove large particles, and stored in 25% glycerol at  $-80^{\circ}\text{C}$  until sorting. Finally, a soil microbiome sample was collected at Lawrence Berkeley National Laboratory (latitude 37.877382, longitude -122.250410). Approximately 5 g of soil was vortexed for 1 minute in 30 ml of water to release cells attached to sediment before centrifuging a 2 ml aliquot at 500 RCF for 5 minutes to pellet large particles. The supernatant was passed through a 5  $\mu\text{m}$  filter and immediately used for cell sorting.

### Fluorescence activated cell sorting

Immediately before cell sorting, environmental bacteria and bacterial isolates were filtered through a 35  $\mu\text{m}$  cell strainer to remove any remaining large particles and diluted to  $\sim 10^6$  cells/ml in filter-sterilized 1 $\times$  PBS containing 1 $\times$  SYBR-Green DNA stain (ThermoFisher, USA). Individual cells were sorted using an Influx FACS machine (BD Biosciences) into LoBind 96-well plates (Eppendorf, Germany) containing either 3  $\mu\text{l}$  of BioSkryb SL1-B Solution for PTA reactions or 1.2  $\mu\text{l}$  of TE for MDA and WGA-X reactions. Plates were treated for 10 minutes in a UV crosslinker before sorting to remove any contaminating DNA. Cells were discriminated based on a combination of forward scatter characteristics and SYBR Green fluorescence. A single-cell sort mask with extra droplet discrimination was used to ensure only one cell was sorted into each well.

### Whole genome amplification

PTA was performed using the ResolveDNA Bacteria kit (BioSkryb Genomics, USA). Briefly, 3  $\mu\text{l}$  of SL-B lysis reagent (BioSkryb Genomics, USA) was deposited in each well of a LoBind PCR plate (Eppendorf, Germany) prior to sorting. Plates containing sorted cells were film-sealed, briefly spun, mixed at 1400 rpm for 1 minute, and briefly spun again. The plates were then incubated at room temperature for 30 minutes to lyse the cells. Plates were stored at  $-80^{\circ}\text{C}$  until ready to undergo whole genome amplification. PTA DNA amplification was carried as per the ResolveDNA Bacteria protocol (BioSkryb Genomics) unmonitored on a standard thermocycler for 12 hours at  $30^{\circ}\text{C}$  and in total reaction volumes of 20  $\mu\text{l}$ , followed by 3 minutes at  $65^{\circ}\text{C}$ . Amplification kinetics were not monitored for PTA reactions, and all PTA reactions were selected for library creation and sequencing.

MDA was performed using Phi29 DNA Polymerase (Watchmaker Genomics, USA) as described previously [5] with 20  $\mu\text{l}$  reaction volumes to match PTA reaction volumes. Similarly, 20  $\mu\text{l}$  WGA-X [23] reactions were performed with EquiPhi29<sup>TM</sup> DNA Polymerase (Thermo Fisher). Alkaline cell lysis was performed after sorting and before genome amplification as described previously [5, 23]. In addition, a subset of MDA reactions received an additional Ready-Lyse (LGC Biosearch Technologies) lysozyme treatment of 50 U/ $\mu\text{l}$  for 15 minutes prior to alkaline lysis (Supplementary Tables S6 and S7). The kinetics of all MDA and WGA-X reactions were monitored by qPCR for background amplification. Reactions that amplified before no-template negative controls were considered successful and a subset was selected for sequencing library preparation. Supplementary Table S1 summarizes the number of total amplification reactions, successful reactions, and libraries sequenced.

Amplified DNA from PTA reactions and the selected MDA and WGA-X reactions were cleaned using SeraMagSelect beads at

a 2 $\times$  beads-to-sample ratio (Cytiva Life Sciences, USA) prior to sequencing library preparation.

### Library preparation and genome sequencing

Sequencing libraries were prepared from 10 to 100 ng of amplified DNA using the Nextera DNA flex library prep (Illumina, USA) using IDT for Illumina – Nextera DNA UD Indexes Sets A-D (Illumina, USA). Fragmentation times and amplification cycles were performed according to the ranges recommended by the manufacturer. Amplification reactions were cleaned using SPRI beads (Beckman Coulter, USA) at a 2 $\times$  beads-to-sample ratio. Library concentrations and sizes were analyzed by TapeStation 2200 using D1000 ScreenTapes (Agilent, USA), and library concentration was determined using a Qubit fluorometer with DNA High Sensitivity reagents (ThermoFisher, USA). Bacterial isolates and a subset of the aquatic environmental cells were sequenced on the NextSeq 2000 (Illumina), while the remaining libraries from aquatic and soil bacteria were sequenced on the Novaseq 6000 (Illumina) (Supplementary Tables S6 and S7). All libraries were sequenced using a 2 $\times$ 150bp read format.

### Read processing and genome assembly

Sequencing reads were filtered for quality using the `rqc.filter2.sh` script from BBTools Version 39.01 (<https://bbtools.jgi.doe.gov>) with following parameters: `rna=f trimfragadapter=t qtrim=r trimq=6 maxns=1 maq=10 minlen=49 mlf=0.33 phix=t removehuman=t removedog=t removecat=t removemouse=t khist=t removemicrobes=t sketch kapa=t clumpify=t rqcfilterdata=/clusterfs/jgi/groups/gentech/genome_analysis/ref/RQCFilterData barcodefilter=f trimpolyg=5`.

To generate assemblies from high and low levels of sequencing effort, each library was first subsampled to a maximum of 20 M and 1 M quality filtered reads (3 Gbp and 150 Mbp, respectively). Each subsampled library version was then normalized using `bbtools.bbnorm` with parameters: `bits=32 min=2 target=100 pigz unpigz ow=t`. This normalization reduces the massive redundancy of reads from highly covered genome regions. Error correction was done on the normalized fastq using `bbtools.tadpole` with parameters: `mode=correct pigz unpigz ow=t`. Normalized reads were assembled using SPAdes v3.15.3 [33] using parameters: `—phred-offset 33 -t 16 -m 64 —sc -k 25,55,95`.

Assembled contigs were trimmed to remove 200 bp from each end, and contigs <2000 bp were removed to mitigate the impact of spurious short contigs [38–41].

### Genome quality assessment and taxonomic classification

The quality of SAGs derived from isolates was determined using QUAST version 5.2.0 [42]. Because sequencing effort varied substantially among bacterial isolate SAGs, assemblies made with 1 M reads were compared so that all replicates had equivalent sequencing depths. Genome coverage levels were determined by mapping each of the isolate SAGs against its corresponding reference genome: *E. coli* (IMG taxon ID: 2600254969), *P. putida* (IMG taxon ID: 2667527229) and *B. subtilis* (IMG taxon ID: 643886132). The `bbmap` parameters used in the analysis were `bbmap.sh -Xmx100g fast=t 32bit=t`. The resulting bam files were passed `bedtools` (v2.31.0) [43] to generate coverage files using the `genomecov` function. Lorenz curves and Gini coefficients were calculated from `genomecov` files using the R package `gglorenz` (v0.0.2). The Gini coefficient quantifies the observed deviation from perfect uniformity for each replicate cell, with smaller coefficients indicating more uniform coverage [44].

Environmental SAG assemblies were screened for contamination using a stepwise approach. First, we removed any human contigs. Next, we applied MAGpurify [45] in two sequential stages to remove contaminant contigs based on GC content and phylogenetic markers (stage 1) and tetranucleotide signatures (stage 2). Following the MAGpurify cleanup, we mapped reads generated from negative control reactions that lacked sorted cells and removed contigs with coverage >5X. Finally, we ran megablast against the NCBI non-redundant database and removed contigs with top hits to a set of organisms consistently found in the negative control reactions (Supplementary Table S8). Informatic decontamination reduced median contamination estimates for PTA SAGs from roughly 3% to 1.5% in genome versions assembled from 1 M reads (150 Mbp). Decontamination had little to no impact on MDA and WGA-X SAGs whose contamination levels were <0.1% before treatment. Following contaminant removal, the quality of the environmental SAGs was assessed with CheckM2 (v1.0.1) [34] and MDMcleaner [46].

Assemblies from MDA and WGAX reactions were excluded from subsequent analyses if the total assembly size using all available reads was <0.2 Mbp and genome completeness was <2%. PTA assemblies entirely dominated by contaminant contigs and indistinguishable from assemblies of reagent-only, no-template controls were also excluded.

Statistical analysis of proportional results such as Gini coefficients, genome completeness, and genome contamination were performed on arcsine transformed data.

Taxonomic assignments of environmental SAGs were made with GTDB-tk (v2.3.2) [47]. SAGs derived from 20 M reads (3 Gbp) were used, when available, for taxonomic analysis because GTDB-tk struggled to make assignments to the less complete MDA and WGA-X genomes generated with 1 M reads (150 Mbp).

## Identification of viruses, plasmids, and biosynthetic gene clusters

Putative virus and plasmid contigs were identified by screening genomes with geNomad [48] using the end-to-end analysis parameter. Only hits >5 kb were included in downstream analyses. Of the viral contigs, 85% were predicted to be Caudoviricetes. Biosynthetic gene clusters were predicted using the JGI Secondary Metabolites Collaboratory pipeline which primarily uses anti-SMASH v7.0 for prediction [49].

## Acknowledgements

The work conducted by the U.S. Department of Energy Joint Genome Institute (<https://ror.org/04xm1d337>), a DOE Office of Science User Facility, was supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

## Author contributions

C.G., R.R.M., T.W., I.B., and V.G.P. designed the study. V.G.P., D.G., and K.W. performed the experiments. R.M.B., M.B., D.W.U., C.G., I.B., and D.K. analyzed the data. R.R.M., R.M.B., T.W., C.G., and V.G.P. wrote the manuscript. All authors read and approved the final manuscript.

## Supplementary material

Supplementary material is available at ISME Communications online.

## Conflicts of interest

CG and VGP are co-inventors on a patent related to this work. CG and VGP are equity holders of BioSkrbyb Genomics. CG is a co-founder and Board member of BioSkrbyb Genomics. The remaining authors declare no competing interests.

## Funding

Dr. Charles Gawad has been supported for this work by a Chan Zuckerberg Biohub Investigator Award, NIH Director's New Innovator Award (7DP2CA239145), and Burroughs Wellcome Career Award for Medical Scientists.

## Data availability

Raw sequencing reads were deposited in NCBI's SRA (<https://www.ncbi.nlm.nih.gov/sra>), and annotated assemblies of environmental SAGs based on 1 M reads were deposited in the JGI's Integrated Microbial Genomes and Microbiomes database (<https://img.jgi.doe.gov/>). Bioproject, biosample, and IMG genome IDs can be found in Supplementary Tables S6 and S7.

## References

1. Tyson GW, Chapman J, Hugenholtz P et al. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 2004;**428**:37–43. <https://doi.org/10.1038/nature02340>
2. Wrighton KC, Thomas BC, Sharon I et al. Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 2012;**337**:1661–5. <https://doi.org/10.1126/science.1224041>
3. Albertsen M, Hugenholtz P, Skarshewski A et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* 2013;**31**:533–8. <https://doi.org/10.1038/nbt.2579>
4. Rinke C, Schwientek P, Sczyrba A et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 2013;**499**:431–7. <https://doi.org/10.1038/nature12352>
5. Rinke C, Lee J, Nath N et al. Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat Protoc* 2014;**9**:1038–48. <https://doi.org/10.1038/nprot.2014.067>
6. Woyke T, Doud DFR, Schulz F. The trajectory of microbial single-cell sequencing. *Nat Methods* 2017;**14**:1045–54. <https://doi.org/10.1038/nmeth.4469>
7. Marcy Y, Ouverney C, Bik EM et al. Dissecting biological 'dark matter' with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci USA* 2007;**104**:11889–94. <https://doi.org/10.1073/pnas.0704662104>
8. Pachiadaki MG, Brown JM, Brown J et al. Charting the complexity of the marine microbiome through single-cell genomics. *Cell* 2019;**179**:1623–1635.e11. <https://doi.org/10.1016/j.cell.2019.11.017>
9. Nayfach S, Roux S, Seshadri R et al. A genomic catalog of Earth's microbiomes. *Nat Biotechnol* 2020;**39**:499–509. <https://doi.org/10.1038/s41587-020-0718-6>
10. Almeida A, Mitchell AL, Boland M et al. A new genomic blueprint of the human gut microbiota. *Nature* 2019;**568**:499–504. <https://doi.org/10.1038/s41586-019-0965-1>
11. Pasolli E, Asnicar F, Manara S et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes

- from metagenomes spanning age, geography, and lifestyle. *Cell* 2019;**176**:649–662.e20. <https://doi.org/10.1016/j.cell.2019.01.001>
12. Pavlopoulos GA, Baltoumas FA, Liu S et al. Unraveling the functional dark matter through global metagenomics. *Nature* 2023;**622**:594–602. <https://doi.org/10.1038/s41586-023-06583-7>
  13. Bowers RM, Nayfach S, Schulz F et al. Dissecting the dominant hot spring microbial populations based on community-wide sampling at single-cell genomic resolution. *ISME J* 2021;**16**:1337–47. <https://doi.org/10.1038/s41396-021-01178-4>
  14. Meziti A, Tsementzi D, Rodriguez-R LM et al. Quantifying the changes in genetic diversity within sequence-discrete bacterial populations across a spatial and temporal riverine gradient. *ISME J* 2019;**13**:767–79. <https://doi.org/10.1038/s41396-018-0307-6>
  15. Meyer F, Fritz A, Deng ZL et al. Critical assessment of metagenome interpretation: the second round of challenges. *Nat Methods* 2022;**19**:429–40.
  16. Jarett JK, Džunková M, Schulz F et al. Insights into the dynamics between viruses and their hosts in a hot spring microbial mat. *ISME J* 2020;**14**:2527–41. <https://doi.org/10.1038/s41396-020-0705-4>
  17. Roux S, Hawley AK, Torres Beltran M et al. Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *elife* 2014;**3**:e03125. <https://doi.org/10.7554/eLife.03125>
  18. Engel P, Stepanauskas R, Moran NA. Hidden diversity in honey bee gut symbionts detected by single-cell genomics. *PLoS Genet* 2014;**10**:e1004596. <https://doi.org/10.1371/journal.pgen.1004596>
  19. Kashtan N, Roggensack SE, Rodrigue S et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* 2014;**344**:416–20. <https://doi.org/10.1126/science.1248575>
  20. Ishoey T, Woyke T, Stepanauskas R et al. Genomic sequencing of single microbial cells from environmental samples. *Curr Opin Microbiol* 2008;**11**:198–204. <https://doi.org/10.1016/j.mib.2008.05.006>
  21. Clingenpeel S, Clum A, Schwientek P et al. Reconstructing each cell's genome within complex microbial communities - dream or reality? *Front Microbiol* 2014;**5**:771. <https://doi.org/10.3389/fmicb.2014.00771>
  22. Zong C, Lu S, Chapman AR et al. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 2012;**338**:1622–6. <https://doi.org/10.1126/science.1229164>
  23. Stepanauskas R, Fergusson EA, Brown J et al. Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat Commun* 2017;**8**:84. <https://doi.org/10.1038/s41467-017-00128-z>
  24. Chen C, Xing D, Tan L et al. Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). *Science* 2017;**356**:189–94. <https://doi.org/10.1126/science.aak9787>
  25. Aoki H, Masahiro Y, Shimizu M et al. Agarose gel microcapsules enable easy-to-prepare, picolitre-scale, single-cell genomics, yielding high-coverage genome sequences. *Sci Rep* 2022;**12**:17014. <https://doi.org/10.1038/s41598-022-20923-z>
  26. Zheng W, Zhao S, Yin Y et al. High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome. *Science* 2022;**376**:eabm1483. <https://doi.org/10.1126/science.abm1483>
  27. Hosokawa M, Nishikawa Y, Kogawa M et al. Massively parallel whole genome amplification for single-cell sequencing using droplet microfluidics. *Sci Rep* 2017;**7**:5199. <https://doi.org/10.1038/s41598-017-05436-4>
  28. Marcy Y, Ishoey T, Lasken RS et al. Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS Genet* 2007;**3**:1702–8. <https://doi.org/10.1371/journal.pgen.0030155>
  29. Chijiwa R, Hosokawa M, Kogawa M et al. Single-cell genomics of uncultured bacteria reveals dietary fiber responders in the mouse gut microbiota. *Microbiome* 2020;**8**:5. <https://doi.org/10.1186/s40168-019-0779-2>
  30. Xu L, Brito IL, Alm EJ et al. Virtual microfluidics for digital quantification and single-cell sequencing. *Nat Methods* 2016;**13**:759–62. <https://doi.org/10.1038/nmeth.3955>
  31. Sobol MS, Kaster A-K. Back to basics: a simplified improvement to multiple displacement amplification for microbial single-cell genomics. *Int J Mol Sci* 2023;**24**:4270. <https://doi.org/10.3390/ijms24054270>
  32. Gonzalez-Pena V, Natarajan S, Xia Y et al. Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc Natl Acad Sci USA* 2021;**118**:e2024176118. <https://doi.org/10.1073/pnas.2024176118>
  33. Bankevich A, Nurk S, Antipov D et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**:455–77. <https://doi.org/10.1089/cmb.2012.0021>
  34. Chklovski A, Parks DH, Woodcroft BJ et al. CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat Methods* 2023;**20**:1203–12. <https://doi.org/10.1038/s41592-023-01940-w>
  35. Bowers RM, Kyrpides NC, Stepanauskas R et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017;**35**:725–31. <https://doi.org/10.1038/nbt.3893>
  36. Woyke T, Sczyrba A, Lee J et al. Decontamination of MDA reagents for single cell whole genome amplification. *PLoS One* 2011;**6**:e26161. <https://doi.org/10.1371/journal.pone.0026161>
  37. Hutchison CA, Smith HO, Pfannkoch C et al. Cell-free cloning using  $\phi$ 29 DNA polymerase. *Proc Natl Acad Sci* 2005;**102**:17332–6. <https://doi.org/10.1073/pnas.0508809102>
  38. Kang DD, Li F, Kirton E et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019;**7**:e7359. <https://doi.org/10.7717/peerj.7359>
  39. Sczyrba A, Hofmann P, Belmann P et al. Critical assessment of metagenome interpretation - a benchmark of metagenomics software. *Nat Methods* 2017;**14**:1063–71. <https://doi.org/10.1038/nmeth.4458>
  40. Li D, Liu C-M, Luo R et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015;**31**:1674–6. <https://doi.org/10.1093/bioinformatics/btv033>
  41. Kang DD, Froula J, Egan R et al. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 2015;**3**:e1165. <https://doi.org/10.7717/peerj.1165>
  42. Mikheenko A, Prjibelski A, Saveliev V et al. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 2018;**34**:i142–50. <https://doi.org/10.1093/bioinformatics/bty266>
  43. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2. <https://doi.org/10.1093/bioinformatics/btq033>

44. Motley ST, Picuri JM, Crowder CD *et al.* Improved multiple displacement amplification (iMDA) and ultraclean reagents. *BMC Genomics* 2014;**15**:443. <https://doi.org/10.1186/1471-2164-15-443>
45. Nayfach S, Shi ZJ, Seshadri R *et al.* New insights from uncultivated genomes of the global human gut microbiome. *Nature* 2019;**568**:505–10. <https://doi.org/10.1038/s41586-019-1058-x>
46. Vollmers J, Wiegand S, Lenk F *et al.* How clear is our current view on microbial dark matter? (Re-)assessing public MAG & SAG datasets with MDMcleaner. *Nucleic Acids Res* 2022;**50**:e76. <https://doi.org/10.1093/nar/gkac294>
47. Chaumeil PA, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* 2020;**36**:1925–7. <https://doi.org/10.1093/bioinformatics/btz848>
48. Camargo AP, Roux S, Schulz F *et al.* Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 2023. <https://doi.org/10.1038/s41587-023-01953-y>
49. Blin K, Shaw S, Augustijn HE *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res* 2023;**51**:W46–50. <https://doi.org/10.1093/nar/gkad344>