Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding

Freeman Lan1,2, Benjamin Demaree1,2, Noorsher Ahmed1 & Adam R Abate1–4

The application of single-cell genome sequencing to large cell populations has been hindered by technical challenges in isolating single cells during genome preparation. Here we present single-cell genomic sequencing (SiC-seq), which uses droplet microfluidics to isolate, fragment, and barcode the genomes of single cells, followed by Illumina sequencing of pooled DNA. We demonstrate ultra-high-throughput sequencing of >50,000 cells per run in a synthetic community of Gram-negative and Gram-positive bacteria and fungi. The sequenced genomes can be sorted in silico based on characteristic sequences. We use this approach to analyze the distributions of antibiotic-resistance genes, virulence factors, and phage sequences in microbial communities from an environmental sample. The ability to routinely sequence large populations of single cells will enable the de-convolution of genetic heterogeneity in diverse cell populations.

We describe a method for single-cell genome sequencing at ultra-high-throughput (SiC-seq) using droplet microfluidics. In SiC-seq, we encapsulate cells in hydrogel microspheres (microgels) that are permeable to molecules with hydraulic diameters smaller than the pore size, including enzymes, detergents, and small molecules, but sterically trap large molecules such as genomic DNA. This allows us to use a series of washes on encapsulated cells, to perform the requisite steps of cell lysis and genome processing, while maintaining compartmentalization of each genome. Using a combination of microgel and microfluidic processing steps, we lyse the cells, fragment the genomes, and attach unique barcodes to all fragments, in a workflow that processes >50,000 cells in a few hours. The barcoded fragments for all cells can then be pooled and sequenced, and the reads grouped by barcode, providing a library of single-cell genomes that can be subjected to additional downstream processing, including demographic characterization and in silico cytometry (Fig. 1).

RESULTS

SiC-seq workflow

The principal strategy of SiC-seq is to label all DNA fragments originating from the same genome with a sequence identifier (barcode) unique to that cell. The resultant products are chimeric, comprising a barcode sequence covalently linked to a random fragment of the cell genome. The barcodes allow all reads belonging to a given cell to be identified through shared sequence. We use libraries of barcode droplets containing the barcode sequences, which we merge with the genome-containing droplets to be barcoded. To prepare a barcode droplet library, we encapsulate into droplets, at limiting dilution, oligonucleotides comprising 15 random bases flanked by constant sequences with PCR reagents and primers complementary to the constant regions of the barcodes with one side containing the Illumina P7

1Department of Bioengineering and Therapeutic Sciences, California Institute for Quantitative Biosciences, University of California, San Francisco, California, USA.
2UC Berkeley-UCSF Graduate Program in Bioengineering, University of California, San Francisco, California, USA.
3Chan Zuckerberg Biohub, San Francisco, California, USA.
4Present address: Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, California, USA. Correspondence should be addressed to A.R.A. (adam@abatelab.org).

Received 23 August 2016; accepted 4 April 2017; published online 29 May 2017; corrected before print 21 June 2017; doi:10.1038/nbt.3880
flow cell adaptor (Fig. 2a)\textsuperscript{16}. The droplets were then thermal-cycled to amplify the barcode sequences via digital droplet PCR, generating \textasciitilde{}10 million barcode droplets in a few hours. Before the single-cell genomes can be barcoded, they must be physically isolated, purified, and fragmented. To accomplish this, we encapsulated single cells in agarose microgels using a two-stream co-flow droplet maker, which merges a cell suspension stream with a molten agarose stream, forming a droplet consisting of an equal volume of both streams (Fig. 2b and Supplementary Fig. 1a). The droplet maker runs at \textasciitilde{}10 kHz, allowing us to generate \textasciitilde{}10 million \textasciitilde{}22-\textmu{}m-diameter droplets in \textasciitilde{}20 min in a total volume of aqueous emulsion of \textasciitilde{}60 \mu{}L. Hence, droplet generation is fast and the total volume consumed small, allowing us to load cells at a rate of 1:10 to minimize multi-cell encapsulation. After solidifying the agarose by cooling, the microgels were then transferred from oil to aqueous carrier phase to be subjected to cell lysis and genome purification. To lyse the cells, we incubated the microgels overnight in a mixture of lytic enzymes, digesting the protective microbial cell walls. We then incubated them in a mixture of detergents and proteases for 30 min, solubilizing lipids and digesting proteins, preserving only the total volume consumed small, allowing us to load cells at a rate of 1:10 to minimize multi-cell encapsulation. After solidifying the agarose by cooling, the microgels were then transferred from oil to aqueous carrier phase to be subjected to cell lysis and genome purification. To lyse the cells, we incubated the microgels overnight in a mixture of lytic enzymes, digesting the protective microbial cell walls. We then incubated them in a mixture of detergents and proteases for 30 min, solubilizing lipids and digesting proteins, preserving only the abundance of corresponding organisms in the data set, we compared abundance estimates calculated by short-read alignment, taxonomic profiling programs, and counting under brightfield microscopy (Fig. 3 and Supplementary Fig. 7). We found that all methods were in reasonable agreement when reads were pooled and analyzed in bulk and when species identities were assigned to each barcode based on the most commonly mapped species in a group. This demonstrates that SiC-seq enables estimation of species abundance in a microbial population consistent with accepted metagenomic methods.

Sequencing the genome of a single cell typically incurs coverage distribution bias\textsuperscript{18} due to uneven amplification of DNA starting from a single genome copy. To investigate coverage distribution bias in SiC-seq, we plotted the normalized coverage distribution for reads aggregated from all barcode groups for each microbe (Fig. 3d,e, and Supplementary Fig. 8). With the exception of coverage gaps due to the low abundance of cells of certain species within the standard microbial community, we observed no substantial coverage bias. This indicates that the sampling of each genome within a barcode group was random, so that when all groups were overlaid, a uniform distribution was obtained. We further inspected the distribution of reads in individual barcode groups and found no substantial bias (Supplementary Fig. 9). We believe that coverage bias was minimal because each genome was amplified in a tiny volume of \textasciitilde{}65 \mu{}L, which has been shown to curtail bias-inducing runaway of exponential amplification\textsuperscript{19}. Since the sequencing library was composed

Validation of SiC-seq on an artificial microbial community

The objective of SiC-seq is to provide single-cell genomic sequences bundled in barcode groups. To validate that SiC-seq generated single-cell barcode groups, we applied it to an artificial microbial community containing three Gram-negative bacteria, five Gram-positive bacteria, and two yeasts, which are typically difficult cell types to lyse. We prepared a single-cell library from this community using SiC-seq and sequenced it on an Illumina MiSeq, yielding \textasciitilde{}6 million single-end reads of 150 bp after quality filtering. We grouped reads by barcode and discarded groups with <50 reads, yielding the final 48,989 barcode groups (Fig. 3a). Each barcode group represented a low-coverage genome of a cell, with a sequencing depth of \textasciitilde{}0.1% to \textasciitilde{}1% (Supplementary Fig. 3).

To determine whether the barcode groups indeed corresponded to single cells, we mapped all reads to the reference genomes of the ten species. If two microbes reside within the same barcode group, reads will map to two genomes. We defined a group purity score as the fraction of reads mapping to the most mapped reference (the ideal barcode group has a purity score of 1.0). The distribution of purity scores was strongly skewed to high values with the majority of purity scores \textasciitilde{}0.95, suggesting that most barcode groups represented single cells; this result was consistent even when we took into account the different genome sizes of the ten species (Fig. 3b and Supplementary Fig. 4) and when we examined purity individually for each species (Supplementary Fig. 5). We further examined the rare barcode groups with low (<0.8) purity scores and determined that the majority of those barcode groups represent rare cases where two cells were encapsulated into one droplet or the occasional coalescence of two single-cell-containing droplets (Supplementary Fig. 6). To determine whether the abundance of SiC-seq barcodes reflected the abundance of corresponding organisms in the data set, we compared abundance estimates calculated by short-read alignment, taxonomic profiling programs, and counting under brightfield microscopy (Fig. 3 and Supplementary Fig. 7). We found that all methods were in reasonable agreement when reads were pooled and analyzed in bulk and when species identities were assigned to each barcode based on the most commonly mapped species in a group. This demonstrates that SiC-seq enables estimation of species abundance in a microbial population consistent with accepted metagenomic methods.

Sequencing the genome of a single cell typically incurs coverage distribution bias\textsuperscript{18} due to uneven amplification of DNA starting from a single genome copy. To investigate coverage distribution bias in SiC-seq, we plotted the normalized coverage distribution for reads aggregated from all barcode groups for each microbe (Fig. 3d,e, and Supplementary Fig. 8). With the exception of coverage gaps due to the low abundance of cells of certain species within the standard microbial community, we observed no substantial coverage bias. This indicates that the sampling of each genome within a barcode group was random, so that when all groups were overlaid, a uniform distribution was obtained. We further inspected the distribution of reads in individual barcode groups and found no substantial bias (Supplementary Fig. 9). We believe that coverage bias was minimal because each genome was amplified in a tiny volume of \textasciitilde{}65 \mu{}L, which has been shown to curtail bias-inducing runaway of exponential amplification\textsuperscript{19}. Since the sequencing library was composed
of ~50,000 amplified genomes, the amplification of each genome was limited by the tiny volume while still producing sufficient total DNA for sequencing.

**SiC-seq data analysis with in silico cytometry**

The genomic sequences generated using SiC-seq are grouped by cell barcodes, which is complementary to the sequences generated from shotgun metagenomic sequencing. Existing computational tools are ill-suited to analyze these data because they do not exploit the single-cell barcode information unique to SiC-seq. To address this, we utilized a sequence analysis pipeline in which reads are organized hierarchically as barcode groups, generating a single-cell-reads database (SiC-Reads) (Supplementary Fig. 10). To build SiC-Reads, we filtered raw sequences by quality, grouped them by barcode, and estimated a taxonomic classification of each group using phylogenetic profilers. We also estimated a purity score equal to the fraction of reads mapping to the dominant taxon within the classifiable set. Additional properties of barcode groups and reads, such as presence of sequences corresponding to antibiotic-resistance genes, can be added to the database as they are discovered during analysis.

The massive set of single-cell genomes present in SiC-Reads provides new opportunities for discovering associations between sequences within single cells, in a process we dub *in silico* cytometry. SiC-Reads comprises a collection of single-cell genomes that can be sorted *in silico*, analogous to what is commonly done with flow cytometry on single cells. The database can be sorted repeatedly to mine for correlations between different genetic sequences and structures. Moreover, as new associations are learned, new sorting parameters can be defined, enabling discoveries without having to repeat the experiment.

**Taxonomic distribution of antibiotic resistance in microbes**

To demonstrate *in silico* cytometry, we used SiC-seq to sequence a microbial community recovered from coastal seawater of San
Antibiotic resistance has become increasingly common and represents a considerable threat to global human health. While antibiotic-resistance genes can be identified in most environments, traditional methods can detect the presence of these genes but not their taxonomic distribution.

To determine the distribution of antibiotic-resistance genes among taxa in our data set, we searched our SiC-Reads database for known antibiotic-resistance genes, finding 1,081 (0.012% of reads), representing 99.8% bacteria, 0.04% archaea, and 0.16% viruses. Barcode groups were assigned a taxonomic classification based on the reads they contained, following the rule that more than 10% of reads in a barcode group must be classified, and the assigned classification is the taxon with the most supporting reads. Most barcode groups were high purity based on the classifiable sequences (~91% average), in accordance with our control sample (~94% average).

Antibiotic resistance has become increasingly common and represents a considerable threat to global human health. While antibiotic-resistance genes can be identified in most environments by short-read sequencing, scant information on how they are distributed among taxa is available, because obtaining this information usually requires testing or whole genome sequencing of single species; however, culture conditions for most species have not been identified, precluding such analyses.

To determine the distribution of antibiotic-resistance genes among taxa in our data set, we searched our SiC-Reads database for known antibiotic-resistance genes, finding 1,081 (0.012% of reads), representing 108 (0.30%) barcode groups. The taxonomic distribution of antibiotic-resistance genes in our database had a clear structure, although it did not correlate with what is known from genomes in public databases (Fig. 4a and Supplementary Fig. 12a). This is unsurprising as differences are expected in the natural coastline environment compared to the environment of isolated and sequenced strains. The most abundant taxa associated with antibiotic resistance were not the most abundant community members overall, suggesting that in this community, certain taxa tend to associate more with antibiotic-resistance genes.

Association of virulence factors with host bacteria

Virulence factors, like antibiotic-resistance genes, are important genetic factors in determining the threat that specific microbes pose to human health. Many opportunistic pathogens reside in natural communities in the environment and cause outbreaks when transmitted to a suitable host. Monitoring and detecting potentially pathogenic microbes is important for public health. Like antibiotic-resistance genes, traditional methods can detect the presence of these genes but not their taxonomic distribution.

To examine the taxonomic distribution of virulence factors in our data set, we searched our coastal microbial community database for known virulence factor genes, yielding matches in 1,949 (0.022%) reads in 101 (0.28%) barcode groups, consisting of 29 prevalent virulence factors distributed among 13 microbial genera. The abundance of each taxon where virulence factors were found did not reflect their abundance in the total population, suggesting that certain genera tend to carry more virulence factors than others. To quantify this, we calculated the virulence factor ratio, the ratio between the number of barcode groups containing virulence factors and the total number of barcodes in the community for that species, and normalized the results to the highest virulence factor ratio for comparison (Fig. 4b). Haemophilus and Escherichia stand out among all species; both are known opportunistic human pathogens.
can modify the genomes of their hosts, leaving a copy of their own.

**Determining transduction potential between bacteria**

Many virulent bacterial strains are thought to arise from horizontal gene transfer aided by cross-infection of bacteriophages. Phages can modify the genomes of their hosts, leaving a copy of their own genome behind or transporting fragments of one species to another in a process known as transduction\(^2\). Characterizing the distribution of these mobile elements is challenging in an ecological context because confident identification of foreign genomic fragments within a specific host requires sequencing large numbers of cultures of single species or single cells. Nevertheless, this information is valuable for understanding how bacteria transfer genetic material in general, and how virulent new strains may emerge through this mechanism.

To explore transduction in the microbial community, we searched the SiC-Reads database of the San Francisco coastal community for barcode sequences that could be linked to transduction events. The barcoded nature of SiC-seq data necessitates additional quality control measures for the data, in addition to the quality control measures used in standard sequencing. First, the barcode reads themselves must be of high quality, thus any reads containing low-quality barcode sequences are eliminated, regardless of the quality of the genomic sequences. Second, barcode groups must be quality-controlled to remove small barcode groups, which are the result of mutations in the barcode sequences and background contamination of free DNA. These quality control measures together result in a typical yield of ~55% of raw reads contributing to the SiC-Reads database. Improvements in yield can be made by, for example, computationally identifying reads with mutated barcodes and ‘correcting’ their sequence, but we have found only modest improvements in yields using this method alone\(^10\).

The taxonomic classification of microbes remains an integral part of studying community dynamics, from ecosystems on Earth to those residing in and on our bodies\(^24,25\). However, the taxonomic classification of short reads is error prone, due to the diversity of microbes in most communities and the high degree of horizontal gene transfer that mixes genomic elements in unpredictable ways. SiC-seq improves upon traditional metagenomics sequencing in addressing this challenge because taxonomic identification can be made on the basis of hundreds of reads within a barcode group. Advanced strategies can be applied to infer the taxonomy of a barcode group, including Bayesian probabilistic ones based on classification of each read in the group, or ones weighted toward specific taxonomic markers. Even with this improvement, accurate classification is difficult because the vast majority of sequences remain unclassifiable and the classification of sequences is biased toward well-sequenced taxa in the databases. As genome coverage improves in future iterations of SiC-seq, taxonomic classification of barcode groups should become more confident and precise, potentially arriving at classifications down to the strain under certain circumstances. It is worth noting that taxonomic classification with SiC-seq is also subject to the fundamental limitations of reference-based classification paradigms, where the classification is only as accurate as the match between the sample and the references. Hence, like traditional methods, SiC-seq phylogenetic profiling will become more reliable and complete with the expanding database of reference genomes.

**DISCUSSION**

SiC-seq generates a metagenomic database grouped by single-cell genomes amenable to repeated mining by *in silico* cytometry, for rapid hypothesis generation and testing. We demonstrated its use in measuring the distributions of antibiotic-resistance genes, virulence factors, and transduction potential in microbial communities. The ability of SiC-seq to sequence all cells in a sample without the need to culture them should help us to characterize the microbial dark matter.

The barcoded nature of SiC-seq data necessitates additional quality control measures for the data, in addition to the quality control measures used in standard sequencing. First, the barcode reads themselves must be of high quality, thus any reads containing low-quality barcode sequences are eliminated, regardless of the quality of the genomic sequences. Second, barcode groups must be quality-controlled to remove small barcode groups, which are the result of mutations in the barcode sequences and background contamination of free DNA. These quality control measures together result in a typical yield of ~55% of raw reads contributing to the SiC-Reads database. Improvements in yield can be made by, for example, computationally identifying reads with mutated barcodes and ‘correcting’ their sequence, but we have found only modest improvements in yields using this method alone\(^10\).

The taxonomic classification of microbes remains an integral part of studying community dynamics, from ecosystems on Earth to those residing in and on our bodies\(^24,25\). However, the taxonomic classification of short reads is error prone, due to the diversity of microbes in most communities and the high degree of horizontal gene transfer that mixes genomic elements in unpredictable ways. SiC-seq improves upon traditional metagenomics sequencing in addressing this challenge because taxonomic identification can be made on the basis of hundreds of reads within a barcode group. Advanced strategies can be applied to infer the taxonomy of a barcode group, including Bayesian probabilistic ones based on classification of each read in the group, or ones weighted toward specific taxonomic markers. Even with this improvement, accurate classification is difficult because the vast majority of sequences remain unclassifiable and the classification of sequences is biased toward well-sequenced taxa in the databases. As genome coverage improves in future iterations of SiC-seq, taxonomic classification of barcode groups should become more confident and precise, potentially arriving at classifications down to the strain under certain circumstances. It is worth noting that taxonomic classification with SiC-seq is also subject to the fundamental limitations of reference-based classification paradigms, where the classification is only as accurate as the match between the sample and the references. Hence, like traditional methods, SiC-seq phylogenetic profiling will become more reliable and complete with the expanding database of reference genomes.
The degree of genome coverage affects the usefulness of single-cell data, including the ability to generate assemblies or identify characteristic sequences for in silico cytometry. A limitation of SiC-seq is that, while the number of cells sequenced far exceeds currently described methods, the coverage per cell is substantially lower. Therefore, dropouts in coverage and false negatives can be expected in in silico cytometry analysis. For abundant organisms with a random distribution of coverage in each barcode group, the system is robust to dropouts because results are averaged over many barcode groups. For example, approximately 7,000 Alteromonas barcode groups were taken into account to determine the antibiotic-resistance profile for Alteromonas bacteria. However, for less abundant species, such as Haemophilus, more dropouts can be expected because there may not be enough total sequence information to detect a specific genetic factor. For this reason, the analysis of SiC-Read databases should be limited to relative comparisons of species within the database, and the abundance of target genes within subpopulations should be normalized to the number of barcode groups in the subpopulation. It is worth noting that the dropout phenomenon is not unique to SiC-seq data, but applies to all metagenomic sequencing data where the subpopulation to be analyzed represents a very small fraction of the whole.

Although coverage can be increased by sequencing more reads, the coverage per cell per barcode group will be <100%. This is because the method begins with a single genome copy without amplification and losses incurred during enzymatic and microfluidic processing are irreparable, thus limiting the maximum coverage attainable. In future iterations of SiC-seq, coverage may be increased by amplifying genomes before processing, for example, with multiple displacement amplification in droplets2. Additionally, different strategies for barcoding genomes may yield higher coverage, such as recently described combinatorial indexing via transposase libraries17, which should be applicable to single-cell genomes encapsulated in microgels.

The de novo assembly of whole genomes from metagenomics sequences is a common goal in the field of metagenomics. Mate-paired sequencing can be used to bridge contigs in a metagenomics sequencing data set and potentially assemble whole genomes given sufficient coverage26. Though powerful, the method is limited by the required micrograms of starting DNA, which can be difficult to obtain from microbial ecosystems. Furthermore, many mate-paired reads are required to assemble a whole genome, since each mate-pair bridges only two contigs. SiC-seq data improve on mate-paired sequencing in this respect by requiring minimal amounts of sample as well as enabling the bridging of multiple contigs per barcode group. Consequently, SiC-seq should allow generation of draft genomes from shotgun metagenomic data with far less DNA input and sequencing effort.

While we focused on microbial communities, SiC-seq is also applicable to populations of mammalian cells, where it can have a more direct impact on human health. The grouped reads provided by SiC-seq should afford the information required to determine copy-number variations within the genome, which is relevant to cancer27. The enormous size of mammalian genomes, however, limits the number of cells that can be sequenced for a target level of coverage. Nevertheless, as the cost of sequencing continues to decrease, more cells can be sequenced to greater depth, creating opportunities for characterizing mammalian tissues, cell by cell.

SiC-seq method is a means to isolate and barcode large DNA molecules, irrespective of the entity from which they originate. While we have focused on cells, similar approaches can be applied to any entities whose genomes can be trapped and processed within the gel matrix. SiC-seq’s ability to build and mine large databases of genomes grouped by single cells should contribute to the characterization of heterogeneity across biology.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful for K. Stedman, R. Malmstrom, R. Andino, K. Pollard, and M. Fischbach for very helpful discussion of and advice on the manuscript. We thank C. O’Loughlin at UCSF for providing microbial strains. This work was supported by the National Science Foundation through a CAREER Award (grant number DBI-1253293); the National Institutes of Health (NIH) (grant numbers HG007233-01, R01-EB019453-01, R21HG007233, DP2-AR068129-01, R01-HG008978); and the Defense Advanced Research Projects Agency Living Foundries Program (contract numbers HR0011-12-C-0065, N66001-12-C-4211, HR0011-12-C-0066). Funding for open access charge: (NIH grant number DP2-AR068129-01). F.L. is supported by a PGS-D grant from the National Science and Engineering Research Council of Canada (NSERC).

AUTHOR CONTRIBUTIONS

F.L. and A.R.A. conceived of the SiC-seq method. F.L., B.D., and N.A. designed and performed the experiments, and analyzed data. F.L. and A.R.A. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.
18. de Bourcy, C.F.A. et al. A quantitative comparison of single-cell whole genome amplification methods. *PLoS One* **9**, e105585 (2014).
19. Gole, J. et al. Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat. Biotechnol.* **31**, 1126–1132 (2013).
20. Nathan, C. & Cars, O. Antibiotic resistance—problems, progress, and prospects. *N. Engl. J. Med.* **371**, 1761–1763 (2014).
21. Yildiz, F.H. Processes controlling the transmission of bacterial pathogens in the environment. *Res. Microbiol.* **158**, 195–202 (2007).
22. Jiang, S.C. & Paul, J.H. Gene transfer by transduction in the marine environment. *Appl. Environ. Microbiol.* **64**, 2780–2787 (1998).
23. Ochman, H. & Moran, N.A. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292**, 1096–1099 (2001).
24. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
25. Afshinnekoo, E. et al. Geospatial resolution of human and bacterial diversity with city-scale metagenomics. *Cell Syst.* **1**, 72–87 (2015).
26. Iverson, V. et al. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* **335**, 587–590 (2012).
27. Ni, X. et al. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc. Natl. Acad. Sci. USA* **110**, 21083–21088 (2013).
ONLINE METHODS

Microfluidic devices. The microfluidic devices are fabricated by pouring poly(dimethylsiloxane) (Dow Corning, Sylgard 184) over a negative photomask (MicroChem, catalog no. SU-8 3025) patterned on a silicon wafer (University Wafer) using UV photolithography. The PDMS devices are cured in an oven for 1 h, extracted with a metal scalpel, and punched with a 0.75-mm biopsy core (World Precision Instruments, catalog no. 504529) to create inlets and outlets. Devices are bonded to a glass slide using an oxygen plasma cleaner (Harrick Plasma), and the channels treated with Aquapel (PPG Industries) and baked at 80 °C for 10 min to render them hydrophobic.

Barcode emulsions. Barcode emulsions are prepared through a digital PCR process wherein barcode oligonucleotides are amplified as single molecules in droplets containing PCR reagents. Barcode oligonucleotides (GCAGCCT GCCTATAGCGATACAACCTGCTGTGCGGCGATAGNNNN) are loaded into separate 1-mL syringes (BD) and injected at 300 and 500 µL/h, respectively, into a flow-focusing droplet maker using syringe pumps (New Era, catalog no. NE-501), controlled with a custom Python script (https://github.com/AbateLab/Pump-Control-Program) at 0.1 pM concentration to form 400 µM emulsions with 2% (v/v) Triton X-100 nonionic surfactant for improved thermal stability. The emulsion is thermal cycled (Bio-Rad, T100) with the following program: 98 °C for 3 min, followed by 40 cycles with 2 °C per second ramp rates of 98 °C for 10s, 62 °C for 20s, and 72 °C for 20s, followed by a hold at 12 °C. Fluorescent DNA staining using 10× SYBR Green I (Thermo Fisher Scientific) in HFE-7500 oil is used to quantify barcode encapsulation rate under a fluoroscope (Life Technologies, catalog no. AMAPF000). The emulsion is collected in PCR tubes, and the oil underneath the emulsion removed by pipette and replaced with FC-40 fluorinated oil (Sigma-Aldrich, catalog no. 51142-49-5) with 5% (w/w) PEG-PFPE amphiphilic block copolymer surfactant for improved thermal stability. The emulsion is thermal cycled (Bio-Rad, T100) with the following program: 98 °C for 3 min, followed by 40 cycles with 2 °C per second ramp rates of 98 °C for 10s, 62 °C for 20s, and 72 °C for 20s, followed by a hold at 12 °C. Fluorescent DNA staining using 10× SYBR Green I (Thermo Fisher Scientific) in HFE-7500 oil is used to quantify barcode encapsulation rate under a fluoroscope (Life Technologies, catalog no. AMAPF000).

Water sample collection and filtering. To obtain a natural sample of a microbial community, we collected marine water from Ocean Beach in West San Diego, California, USA (37°44’55.6”N 122°30’33.6”W). Approximately 2 liters of water is obtained by submerging two 1,000-mL glass bottles below the water surface ~20 m from the shoreline. Samples are placed on ice during transport to the laboratory. 100 mL of the sample is passed through a 40-µm cell strainer (Corning, product no. 352340) to remove large debris, including sand. The sample is loaded into a 0.45 µm vacuum filter (Millipore, catalog no. SCHV010RE); this filtering step separates microbes, which are captured on the membrane, and viruses, which are discarded in the filtrate. The membrane is extracted from the apparatus using a scalpel and inserted into a 15-µL centrifuge tube, to which 5 mL of PBS is added. The tube is vortexed at high speed for ~2 min to free the bacterial cells from the membrane. Finally, the cell solution is loaded into a 10-µL syringe and passed through a 5-µm syringe filter (Millipore, catalog no. SLVS025LS) to remove large remaining particulate. The marine cells are counted using the same protocol as the liquid cultures.

Cell encapsulation in agarose microgels. To prepare the artificial community for processing through the SiC-seq workflow, the frozen stock of cells (Zymo Research, catalog no. D6300) is thawed gently in a room-temperature water bath. Cell concentration is determined by manual cell counting under a microscope, and diluted to an appropriate concentration for single-cell encapsulation. The calculated volume of cell solution is transferred to a 1.5-µL centrifuge tube (Fisher Scientific) and washed twice in 1 mL PBS. The cells are re-suspended in a 1-mL solution of PBS containing 17% OptiPrep Density Gradient Medium (Sigma-Aldrich), 0.1 mg/mL BSA (Sigma-Aldrich, catalog no. A9418), and 1% (v/v) Pluronic F-68 (Life Technologies). The cell solution is loaded into a 1-µL syringe and placed on a syringe pump (New Era, catalog no. NE-501). 1 mL of a 3% solution of low gelling temperature agarose (Sigma-Aldrich, catalog no. A9414) and TE buffer (Teknova, catalog no. T0225) is prepared in a 1.5-mL centrifuge tube and heated on a block at 90 °C for approximately 10 min to completely dissolve the agarose powder. The hot agarose is transferred to a 1-mL syringe and placed on a syringe pump. To keep the agarose molten during the microfluidic experiment, a personal space heater is positioned ~5 cm from the agarose syringe and set to run continuously at high heat. HFE-7500 fluorinated oil with 2% (w/w) deprotonated Kryptofix surfactant (DuPont, catalog no. 157FSH) is loaded into a 3-mL syringe. The cell solution, molten agarose, and oil are injected into the co-flow droplet maker at flow rates of 200, 200, and 400 µL/h, respectively, to form the 1.5% agarose microgels. Approximately 500 µL of droplets are collected in a 15-µL centrifuge tube on ice and incubated for 30 min at 4 °C to ensure complete solidification of the microgels.

Resuspending microgels in aqueous buffer. The droplets are centrifuged at 300g for 1 min to maximize separation of the emulsions from the oil. The oil layer is extracted from the tube using a 5-mL syringe and discarded. Emulsions are broken using 2 mL of a 10% (v/v) solution of perfluoro-octanol (Sigma-Aldrich, catalog no. 370533) in HFE-7500; the emulsions are then mixed by pipetting and centrifuged at 300g for 1 min. The oil is removed from the tube using a syringe, and the droplet breaking step is repeated. Following droplet breaking, 2 mL of hexane containing 1% (w/v) Span 80 (Sigma-Aldrich) is added to the microgels to dissolve any remaining oil, and this solution is mixed and centrifuged at 300g for 1 min. The hexane supernatant is removed from the tube and the hexane addition step is repeated. Finally, the microgels are washed three times in 10 mL of aqueous solution TE buffer containing 0.1% (v/v) Triton X-100 nonionic surfactant (Sigma-Aldrich). The microgels are centrifuged at 1,000g for 2 min and the supernatant aspirated between washes. The washed microgels are stored in 5 mL TE buffer at 4 °C before cell lysis.

Cell lysis in microgels. To lyse the cells in the microgels, the particles are submerged in a solution of 2 mL TE buffer solution containing 10 mM DTT (manu), 2.5 mM EDTA (Teknova), and 10 mM NaCl (Sigma-Aldrich). The following quantities of lytic enzymes are also included: 4 U zymolase (Zyme Research), 10 U lysostaphin (Sigma-Aldrich, catalog no. L7386), 100 U mutanolysin (Sigma-Aldrich, catalog no. M9901), and 40 mg lysomye (MP Biomedicals, catalog no. 195303). Cell lysis proceeds overnight in a shaking incubator at 37 °C. The turbid lystate mixture is centrifuged at 1,000g for 1 min, the supernatant removed, and 5 mL of a solution containing 0.5% (w/v) lithium dodecyl sulfate (Sigma-Aldrich) and 10 mM EDTA in TE buffer is added, along with 4 U of Proteinase K (NEB) to solubilize cell debris and digest cellular proteins. The solution is incubated at 50 °C on a heating block for 30 min. Following lysis, the microgels are thoroughly washed to ensure complete removal of detergents and other chemical species, which may inhibit downstream molecular biology reactions. The following washes occur in 10-mL volumes with centrifugation magnitudes of 1,000g between additions of wash solutions: one wash with 2% (v/v) Tween 20 in water; one wash in 100% ethanol (Koptec) to denature any remaining Proteinase K; and five washes with 0.02% (v/v) Tween 20 in water.

Tagmentation of genomic DNA in microgels. Using reagents from a Nextera DNA Library Prep Kit (Illumina, catalog no. FC-121-1030), the washed and lysed gels containing high-molecular-weight genomic DNA are simultaneously fragmented and tagged with a common adaptor sequence. Microgels are re-encapsulated into droplets to minimize cross-contamination during the tagmentation step. A solution of 192 µL DI water, 200 µL tagmentation buffer, and 8 µL Nextera enzyme is prepared and loaded into a 1-mL syringe. Microgels and the tagmentation solution are injected into the re-encapsulation device (Supplementary Fig. 1). The re-encapsulated microgels are incubated in a 1.5-mL tube on a heating block at 50 °C for 1 h.

Microfluidic barcoding of encapsulated cells. Tagmented microgel droplets, barcode droplets, and 500 µL of PCR solution containing 1× Invitrogen Platinum Multiplex PCR Master Mix (Thermo Fisher Scientific, catalog no. 4464268), 400 nM primers FL127 (AATGATACGGCGACCACCGATATCGACTAGCTCACTCTGACGATATCGACTAC) and FL128 (CTGTCTCTTATACACATCTCAGCCTGCGTAATAGCGAGTACAATCTGCTCTGATGCCGCATAGNNNN) were re-encapsulated into droplets to minimize cross-contamination during the tagmentation step. A solution of 192 µL DI water, 200 µL tagmentation buffer, and 8 µL Nextera enzyme is prepared and loaded into a 1-mL syringe. Microgels and the tagmentation solution are injected into the re-encapsulation device (Supplementary Fig. 1). The re-encapsulated microgels are incubated in a 1.5-mL tube on a heating block at 50 °C for 1 h.

DOI:10.1038/nbt.3880

© 2017 Nature America, Inc., part of Springer Nature. All rights reserved.
(CAAGCAGAAGACGGCATACGAGATCCACGTCAATATGC), 50× dilution of NT buffer from the Nextera XT Kit (0.2% SDS) (Illumina, catalog no. FC-131-1024), 1% (v/v) Tween 20, 1% (v/v) PEG–6000, 2.5 U/µL Bst 2.0 WarmStart DNA Polymerase (NEB, catalog no. M0538S) are each loaded into a 1-ml syringe and injected into the sequential merger device as shown in Supplementary Figure 1. HFE-7500 fluorinated oil with 2% (w/w) 008–Fluorouresactant is used as the continuous phase of the emulsion. Merger of the barcode and gel droplet emulsions is achieved using an electrode connected to a cold cathode fluorescent inverter and DC power supply (Mastech). A voltage of 2.0 V at the power supply produces a ~2-kV AC potential at the electrode, which causes touching droplets to merge. The emulsion is collected in a 0.5-ml thin-walled PCR tube (Applied Biosciences), and the HFE-7500 replaced with FC-40 with 5% (w/w) 008–Fluorouresactant before thermal cycling with the following protocol: 65°C for 5 min, 95°C for 2 min, then 30 cycles at 2 °C/s ramp rates of 95°C for 15s, 60°C for 1 min, 72°C for 1 min, and then 72°C for 5 min with optional 12°C overnight hold. After thermal cycling, large (coalesced) droplets are removed using a micropipette, and the emulsion is broken by addition of 20 µL of perfluoro-octanol and brief centrifugation in a micro-centrifuge. The upper aqueous phase is collected and the DNA library is purified using a Zymo DNA Clean & Concentrator-5 kit (Zymo Research). The library is size-selected for DNA fragments in the 200- to 600-bp range using Agencourt AMPure XP beads (Beckman Coulter), quantified with a Bioanalyzer 2100 instrument and High Sensitivity DNA chip (Agilent), and sequenced on an Illumina MiSeq using a custom index primer (GCCCAAGAGACGGCTCCGGGTCGTTTA).

Generating the SiC-Reads database. Raw reads from the MiSeq-generated FASTQ files are filtered by quality and grouped by barcode sequence using the Python script barcodeCleanup.py. A given read is discarded if more than 20% of its bases have a Q-score less than Q20, and all reads associated with a barcode containing less than 50 reads are discarded. This step ensures that all barcode groups, representing single cells, contain a sufficient number of high-quality reads. The resulting reads are exported to a table in a SQLite database with fields containing the barcode sequence, barcode group size, a unique read ID number, and read sequence. When the reference genomes are known, as in the case of the synthetic cell population experiment, the reads are aligned using bowtie2 v2.2.9 with default settings and the SQLite table is updated with relevant alignment information for each read. For environmental samples, the reads are classified by taxonomy using Kraken v0.10.5 with “–quick–min-hits 2” options set, and the output is exported to the SQLite database. Where noted, an alternative taxonomic classifier, metaplank2 v2.6.0, was used with the default marker information file and taxonomic level set to “species only.” krakenAnalysis.py assigns taxonomic identities from the Kraken database to barcode groups by a majority rule, in which a barcode group is classified according to the most common taxonomic label among its classifiable reads. Barcode group purity is calculated from reference alignment data or phylogenetic labels using the script purity.py.

In silico cytometry. Reads from the SiC-Reads database are aligned, using bowtie2 v2.2.9 with –very-sensitive and –end-to-end settings to reference sequences of interest (antibiotic resistance database obtained from28, virulence factor database obtained from core virulence factor genes at the virulence factor database (VFDB)29 Phage sequence database obtained from Phage genome database accessed on May 2016 at http://www.ebi.ac.uk/genomes/phage.html. Mapping reads are then filtered for Mapq > 2 in order to remove ambiguously mapping reads. Barcode groups containing reads that map to the databases are annotated as containing the target sequence and are exported for further analysis if they are taxonomically classified with purity >0.8. To generate the heatmap for transduction potential, all reads associated with a phage and a Kraken-classified barcode group were extracted and grouped according to phage type. Duplicate and near-duplicate reads were removed. The heatmap intensities were calculated as follows: for a given pair of bacterial hosts, the total number of host-phage-host connections in the database were counted. To normalize the data by host abundance, this number was divided by the total number of barcode groups associated with the two hosts.

Calculating the virulence factor ratios. The virulence factor ratios calculations in Figure 4b of the main text was reproduced using reference genomes for the genera shown in the figure. The complete genomes of all species associated with these 12 genera were downloaded from the RefSeq database using the Perl script ncbiDownload.pl. Genomes were pooled into FASTA files labeled by genus. From these reference files, a Python script (bargroupGenerator.py) generated simulated barcode groups of 200 reads per group, with each single-end read 150-bp long. The number of simulated barcode groups generated for a given genus was equal to the number of barcode groups identified for this genus in the San Francisco Coast water sample. The simulated barcode group reads were then aligned to the original virulence factor database using bowtie2 v2.2.9 in ‘local’ alignment mode with default sensitivity settings. Unaligned sequences were removed using Samtools v1.3.1 (samtools view -b -F) and the remaining aligned reads were used to produce the data shown in Supplementary Figure 7b.

Generating the antibiotic-resistance network with reference genomes. An antibiotic resistance graph in Supplementary Figure 7a was generated using references for the six genomes most commonly associated with antibiotic resistance in the SiC-Reads database of the San Francisco Coast water microbial community. The following genomes (with accession numbers) were downloaded from the NCBI RefSeq repository: Alteromonas macleodii ATCC 27126 (CP003841.1), Bacillus subtilis subsp. spizizenii strain NRS 231 (CP010434.1), Delfitia acidovorans SPI-1 (CP000884.1), Enterobacter cloacae subsp. cloacae ATCC 13047 (CP001918.1), Neisseria meningitidis MC58 (AE002098.2), Propionibacterium acnes KPA171202 (AE017283.1). These genomes were combined into a single FASTA file and passed to a short-read simulator, wgsim v0.3.2, which generated 10 million single-end reads of 70 bp each with a base error rate of 0. These reads were aligned to the antibiotic-resistance gene using bowtie2 in ‘local’ mode with default sensitivity settings. All unaligned sequences were removed using Samtools (samtools view -b -F). The aligned sequences in SAM format were imported into Cytoscape v3.4.0, and the network shown in Supplementary Figure 7a was generated using the reference genus and antibiotic-resistance genes as the network targets and sources, respectively. The darkness of the graph’s edges scale linearly with the total number of connections in the data, where darker lines have a greater number of associations.

Characterizing diffusion of genomic DNA fragments in agarose microgels. A microgel sample containing encapsulated, lysed bacteria was stained with SYBR Green I and observed under a fluorescence microscope before and after tagmentation at various time points (Supplementary Fig 2a). In another experiment, the concentration of DNA in the supernatant and contents of the gels was measured in a sample incubated at room temperature. After 2 d at room temperature, the beads were pelleted by centrifugation; the DNA was extracted from the beads and from the supernatant, using a DNA gel extraction kit and a DNA clean-up and concentrator kit (Zymo research D4001T). The extracted DNA was quantified using the Qubit dsDNA high sensitivity assay and Bioanalyzer High Sensitivity dsDNA chip (Supplementary Fig. 2b). As an additional experiment, microgels were incubated at 55 °C with and without tagmentation enzyme to demonstrate the corresponding change in genomic DNA fragment size distribution before and after tagmentation (Supplementary Fig. 2c).

Cell culture and counting. To generate an additional artificial community with which to validate the SiC-seq workflow, liquid cultures of Staphylococcus epidermidis, Saccharomyces cerevisiae (strain S288c), and Bacillus subtilis (strain 168) were grown overnight in a shaking incubator. The following culture conditions were used: Staphylococcus epidermidis and Bacillus subtilis were grown in 3-ml LB broth at 37 °C; Saccharomyces cerevisiae was grown in 3 ml YPD broth at 30 °C. Cell concentration is determined by manually counting serial dilutions of the liquid culture on plastic slides (Thermo Fisher Scientific, catalog no. C10228) using a microscope. The cultures were kept at 4 °C before being used in the microfluidic experiment. An analysis of the SiC-seq experiment for this synthetic microbial community is shown in Supplementary Figure 7.
Data availability. Code availability: scripts used to generate the data in this paper can be accessed at the Abatelab Github: https://github.com/AbateLab/SiC-seq. Accession codes: all sequencing data are accessible at the NCBI Sequence Read Archive through the following accession numbers: SRR2516128, SRR2522025, SRR2516129.

28. Liu, B. & Pop, M. ARDB—Antibiotic Resistance Genes Database. *Nucleic Acids Res.* **37** (Suppl. 1), D443–D447 (2009).
29. Chen, L., Xiong, Z., Sun, L., Yang, J. & Jin, Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.* **40**, D641–D645 (2012).
Corrigendum: Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding

Freeman Lan, Benjamin Demaree, Noorsher Ahmed & Adam R Abate

Nat. Biotechnol. doi:10.1038/nbt.3880; corrected online 21 June 2017

In the version of this article initially published, in Figure 3c, bars on the x axis were labeled S. epidermidis rather than S. enterica. The error has been corrected for the print, PDF and HTML versions of this article.