

# Droplet-based high-throughput single microbe RNA sequencing by smRandom-seq

Received: 17 November 2022

Accepted: 10 July 2023

Published online: 23 August 2023

 Check for updates

Ziye Xu<sup>1,2,9</sup>, Yuting Wang<sup>2,3,9</sup>, Kuanwei Sheng<sup>4,5,9</sup> , Raoul Rosenthal<sup>6</sup>, Nan Liu<sup>2</sup>, Xiaoting Hua<sup>7</sup>, Tianyu Zhang<sup>2</sup>, Jiaye Chen<sup>8</sup>, Mengdi Song<sup>2</sup>, Yuexiao Lv<sup>2</sup>, Shunji Zhang<sup>3</sup>, Yingjuan Huang<sup>2</sup>, Zhaolun Wang<sup>2</sup>, Ting Cao<sup>1,4,6</sup>, Yifei Shen<sup>1</sup>, Yan Jiang<sup>7</sup>, Yunsong Yu<sup>7</sup>, Yu Chen<sup>1</sup>, Guoji Guo<sup>2</sup>, Peng Yin<sup>4,5</sup> , David A. Weitz<sup>4,6</sup>  & Yongcheng Wang<sup>1,2,3,4</sup> 

Bacteria colonize almost all parts of the human body and can differ significantly. However, the population level transcriptomics measurements can only describe the average bacteria population behaviors, ignoring the heterogeneity among bacteria. Here, we report a droplet-based high-throughput single-microbe RNA-seq assay (smRandom-seq), using random primers for in situ cDNA generation, droplets for single-microbe barcoding, and CRISPR-based rRNA depletion for mRNA enrichment. smRandom-seq showed a high species specificity (99%), a minor doublet rate (1.6%), a reduced rRNA percentage (32%), and a sensitive gene detection (a median of ~1000 genes per single *E. coli*). Furthermore, smRandom-seq successfully captured transcriptome changes of thousands of individual *E. coli* and discovered a few antibiotic resistant subpopulations displaying distinct gene expression patterns of SOS response and metabolic pathways in *E. coli* population upon antibiotic stress. smRandom-seq provides a high-throughput single-microbe transcriptome profiling tool that will facilitate future discoveries in microbial resistance, persistence, microbe-host interaction, and microbiome research.

Only recently have scientists begun to appreciate how much our health and disease depend on the trillions of resident microbial communities<sup>1</sup>. Microbial transcriptomics has been a powerful tool for understanding the complexity, plasticity, and regulatory mechanism of microbes<sup>2,3</sup>. However, these population-level gene expression measurements can only describe the average microbe population behaviors. Individual microbes, even of an isogenic population, can keep evolving and differ significantly to increase the adaptivity to environmental changes<sup>4</sup>. To reveal the microbial states with different gene expression patterns, which is critical for microbial resistance,

persistence, microbe–host interaction, and microbiome research, transcriptome profiling of each microbe is required.

Since its establishment in 2009<sup>5</sup>, multiple single-cell RNA-seq (scRNA-seq) methods have been developed and widely used in eukaryotic biology<sup>6–10</sup>. The key strategy of current prevalent droplet-based scRNA-seq systems, such as 10X Genomics Chromium platform<sup>7</sup>, is using droplet microfluidics to co-encapsulate a single cell with a uniquely DNA barcoded bead in a droplet, where subsequently polyadenylated mRNAs are captured by billions of copies of unique bar-coded poly(T) primers carried on the bead. However, this strategy is

<sup>1</sup>Department of Laboratory Medicine, the First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, China. <sup>2</sup>Liangzhu Laboratory, Zhejiang University, Hangzhou, China. <sup>3</sup>College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China. <sup>4</sup>Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, USA. <sup>5</sup>Department of Systems Biology, Harvard Medical School, Boston, MA, USA. <sup>6</sup>John A. Paulson School of Engineering and Applied Sciences and Department of Physics, Harvard University, Cambridge, MA, USA. <sup>7</sup>Department of Infectious Diseases, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, China. <sup>8</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. <sup>9</sup>These authors contributed equally: Ziye Xu, Yuting Wang, Kuanwei Sheng. ✉ e-mail: [Kuanwei.Sheng@wyss.harvard.edu](mailto:Kuanwei.Sheng@wyss.harvard.edu); [Peng\\_Yin@hms.harvard.edu](mailto:Peng_Yin@hms.harvard.edu); [weitz@seas.harvard.edu](mailto:weitz@seas.harvard.edu); [yongcheng@zju.edu.cn](mailto:yongcheng@zju.edu.cn)

incompatible with the single-cell RNA-seq of bacteria. One major reason is that bacterial mRNAs lack 3'-end poly(A) tails, the capture site for poly(T) primers. In addition, the RNA content of a typical bacterium is about two orders of magnitude lower than a typical mammalian cell<sup>3,11</sup>, and ribosomal RNA (rRNA) occupies >80% of total bacterial RNAs<sup>12</sup>, which makes it challenging to efficiently capture and differentiate those low content mRNAs in single bacterium level. The tough cell wall also makes it hard to lyse bacteria and release the enclosed RNA in the droplet. Some relatively low-throughput single bacterium RNA-seq methods<sup>13,14</sup> have been developed to address these issues, which require single bacterium isolation into multi-well plates using single cell manipulator<sup>13</sup> or FACS<sup>14</sup>. Two previously developed bacterial scRNA-seq methods, PETRI-seq<sup>15</sup>, and microSPLIT<sup>16</sup>, using random reverse transcription (RT) primers, enable researchers to analyze thousands of fixed bacteria simultaneously with a split-pool barcoding strategy<sup>17</sup>, which have made great advances in our ability to study the transcriptome heterogeneity of bacterial communities on a large scale. Nonetheless, this combinatorial split-pool barcoding strategy requires multiple reactions in 96-well plates, which is not a popular or sensitive platform for high-throughput single-cell RNA-seq. Moreover, the majority of mapped reads in both PETRI-seq<sup>15</sup> and microSPLIT<sup>16</sup> were rRNA even with mRNA enrichment, leading to considerable sequencing costs.

Microfluidic has been widely used for high-throughput single-cell analysis<sup>18,19</sup>. Our previous work has established a microfluidic droplet barcoding platform and developed inDrop<sup>20</sup>, a high-throughput method for single-cell RNA sequencing, and Microbe-seq<sup>21</sup>, a high-throughput method for single-microbe genome sequencing.

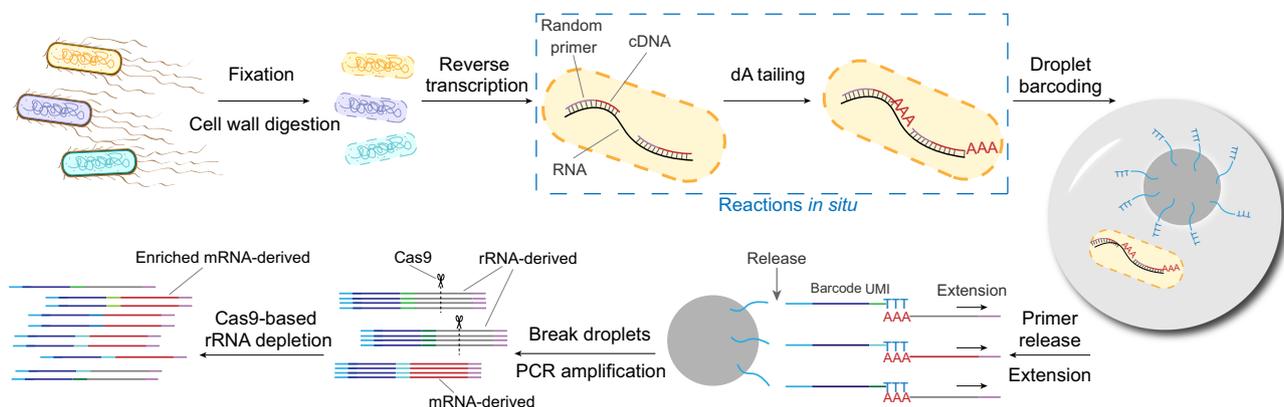
In this work, we advance on recent efforts to develop a high-throughput and high-sensitive single microbe RNA-seq method (smRandom-seq), using random primers for in situ complementary DNA (cDNA) generation, droplets microfluidics for single-microbe barcoding and CRISPR-based rRNA depletion for mRNA enrichment (Fig. 1). The droplet microfluidics platform shows high barcoding efficiency, high species specificity (99%), minor cross-contamination (1.6%), and automation capability. The CRISPR-based rRNA depletion dramatically reduces the rRNA percentage (83–32%). The mapped mRNA reads of *E. coli* are highly enriched with a 4-fold increase (16–63%) due to efficient rRNA depletion. A median of ~1000 genes can be detected per bacterium in a population of ~8000 *E. coli*. We also demonstrate that smRandom-seq could be applied to other common bacterial species, including Gram-negative bacteria (*A. baumannii*, *K. pneumoniae*, and *P. aeruginosa*) and Gram-positive bacteria (*B. subtilis* and *S. aureus*). Moreover, we find that *E. coli* population displays

morphologic heterogeneity upon antibiotic stress and perform smRandom-seq to investigate these transcriptome changes of the individual bacterium. The results of cluster analysis of gene expression by smRandom-seq are consistent with the morphologic features. Three subpopulations of *E. coli* display distinct gene expression patterns of SOS response and metabolic pathways, which might be interesting targets for antibiotic resistance research. By combining the high-sensitive cDNA generation chemistry with the popular droplet barcoding platform, we envision smRandom-seq will have a broad impact and facilitate discoveries in microbiology.

## Results

### Overview of the droplet-based smRandom-seq

The workflow of smRandom-seq is schematically shown in Fig. 1. First, bacteria were fixed overnight with ice-cold 4% paraformaldehyde (PFA) to crosslink the RNAs, DNAs, and proteins inside the bacteria. The fixed bacteria were then permeabilized to facilitate the following in situ reactions. Next, random primers with a GAT 3-letter PCR handle were added to capture total RNAs through multiple temperature cycling to enable maximum bindings of primers on each transcript inside bacteria. cDNAs were then converted in situ by reverse transcription reaction. Subsequently, poly(dA) tails were added to the 3' hydroxyl terminus of the cDNAs in situ by terminal transferase (TdT). Excessive primers, primer dimers, and leftover reagents were washed away by centrifugal washing after each step of in situ reactions. In our experimental pipeline, in situ reactions can be finished in two steps (~3 h). Before proceeding to microfluidic encapsulation, bacteria were imaged to confirm single bacterial morphology and manually counted under a microscope (Supplementary Fig. 1a). Each bacterium was encapsulated into a ~100- $\mu$ m droplet with a poly(T) barcoded bead by using a modified microfluidic device (Supplementary Fig. 1b–d). To efficiently barcode the low content RNAs in single bacterium, we optimized the droplet barcoding platform (Supplementary Figs. 1b, c, 2a) based on our previous inDrop platform<sup>20</sup>, with smaller barcoded beads (~40  $\mu$ m) (Supplementary Fig. 2b), smaller droplets (~100  $\mu$ m) (Supplementary Fig. 1d), a USER enzyme cutting strategy replacing the photocleaving strategy to release primers from barcode beads (Supplementary Fig. 2c), and a 3-step ligation reaction replacing the 2-step extension reaction for barcodes synthesis on the beads (Supplementary Fig. 2c). Shorter and cleaner barcode primers were synthesized by the ligation reaction (Supplementary Fig. 2d). During the barcoding reaction in droplets, the poly(T) primers were released from barcoded beads by the USER enzyme. Simultaneously, the cDNAs were released from bacteria by the RNase H enzyme. Then poly(T) primers bind with



**Fig. 1 | Overview of the droplet-based smRandom-seq.** The workflow of smRandom-seq for microbe samples includes fixation, cell wall digestion, reverse transcription, dA tailing, droplet barcoding, primers releasing and extension, droplets breaking and PCR amplification, and Cas9-based rRNA depletion and

sequencing. Blue dashed box: the two in situ reactions, including reverse transcription and dA tailing. AAA: dA tail in the 3' of cDNA, TTT: poly(dT) in the barcoded primers.

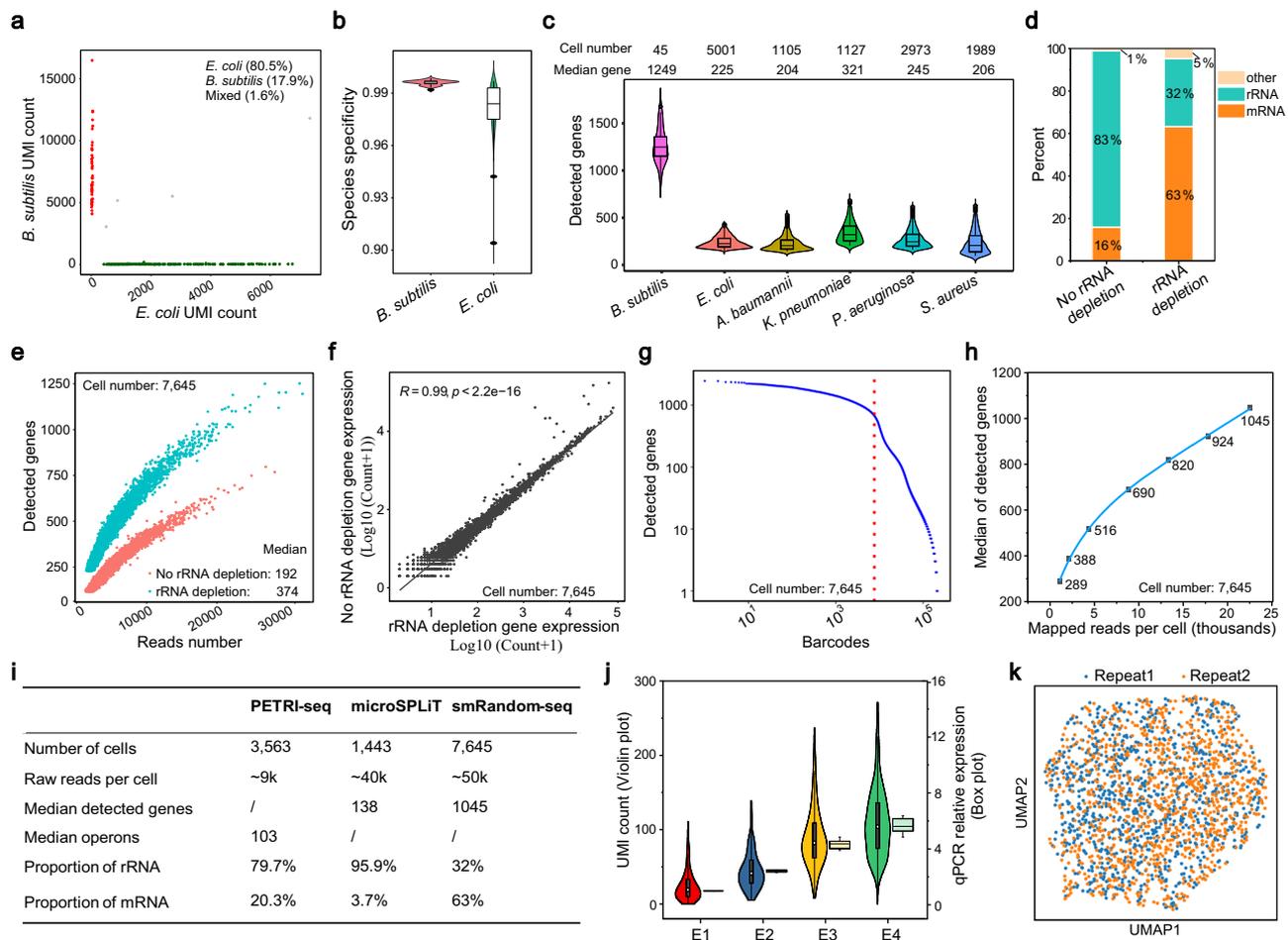
the poly(A) tails on the end of the cDNAs and extend to add a specific barcode to the cDNAs in each droplet and a unique molecular identifier (UMI) to each cDNA. These optimizations increased the barcoding efficiency for a single bacterium and reduced primer-dimers during the barcoding reaction in droplets. The diversity of barcode beads in this study was 442,368 ( $96 \times 96 \times 48$ ). The throughput of current smRandom-seq is estimated from the Poisson distribution described by Zilionis et al.<sup>22</sup>. About 10,000 cells can be processed per experiment using the current smRandom-seq.

After barcoding, we broke the droplets and amplified the bar-coded cDNAs. CRISPR-based rRNA depletion was performed on the cDNA library before next-generation sequencing. As a result of the optimized first- and second-strand cDNAs synthesis, the cDNA library prepared by our method ranged from 200 to 500 bp (Supplementary Fig. 1e), an optimal size range with no need to fragment library for next-generation sequencing platforms. After cutting the total length of the

adaptors (123 nt), the actual length of the mRNA-derived insert in libraries of smRandom-seq is about 100–400 bp.

### Validation of smRandom-seq method using reference bacteria species

We first performed the smRandom-seq assay on a two-species-mixing experiment of *E. coli* (Gram-negative bacteria) and *B. subtilis* (Gram-positive bacteria) to assess the purity of libraries (Fig. 2a). After sequencing and data processing, we used the molecular barcodes for unique cDNAs (UMI) to quantify smRandom-seq. A minor inter-species doublet rate (1.6%) and a very high species specificity of UMI (*B. subtilis*: 99.6%, *E. coli*: 98.4%) were calculated from the alignment results (Supplementary Fig. 3a, Fig. 2b). We detected a median of 6564 and 1785 UMI counts per cell for *B. subtilis* and *E. coli*, as well as a median of 1249 and 429 detected genes per cell, respectively (Supplementary Fig. 3b–e). There were 5 overlapped gene IDs between the top 10



**Fig. 2 | Validation of smRandom-seq using reference bacteria samples.** **a** Scatter plot of UMI counts per cell barcode in the mixture. *E. coli* number: 202. *B. subtilis* number: 45. Mixed cell number: 4. **b** Species specificity of UMI in the mixture. *E. coli* number: 202. *B. subtilis* number: 45. Data in the box plot in **b**, **c**, **j** corresponded to the first (lower hinges) quartiles, third quartiles (upper hinges), and median (center). **c** Distribution of detected genes of different bacterial species. The identified cell numbers were as follows: 45 *B. subtilis*, 5001 *E. coli*, 1105 *A. baumannii*, 1127 *K. pneumoniae*, 2973 *P. aeruginosa*, and 1989 *S. aureus*. **d** Representative proportions of transcript categories of *E. coli* samples with and without rRNA depletion. The cell number of the *E. coli* samples in **d–f** was 7645. **e** Scatter plot showing detected genes and the number of reads per barcode of *E. coli* samples with (blue) and without (red) rRNA depletion. **f** Correlation of gene expression (Log<sub>10</sub>(Counts + 1)) of *E. coli* samples with and without rRNA depletion. The *E. coli* sample with rRNA depletion was sequenced more deeply for the next performance evaluations. **g** The

barcode rank plot of log<sub>10</sub>(total number of detected genes per barcode) versus log<sub>10</sub>(Barcode number) for each possible barcode. The knee point (red dashed line) indicated the threshold for putative *E. coli* (cell number: 7645). **h** Saturation curve for the rRNA depleted and deeper sequenced *E. coli* sample (Cell number: 7645). **i** Comparison of performance with other two reported high-throughput single bacterium RNA-seq methods (PETRI-seq<sup>15</sup> exponential-phase *E. coli* dataset from experiment 2.01, microSPLiT<sup>16</sup>: *E. coli* dataset from the *E. coli* and *B. subtilis* species-mixing experiment). **j** Comparison of qPCR relative expression ( $n = 3$  biologically independent samples) and UMI count of *E. coli* samples (E1  $n = 252$  cells, E2  $n = 254$  cells, E3  $n = 254$  cells, E4  $n = 253$  cells) with a progressive increase in expression of GFP. E1–E4 represents the GFP-*E. coli* samples induced with 0, 6.25, 25, and 100 mM propionate. **k** UMAP projection of two repeated *E. coli* samples (Repeat 1 and Repeat 2) applied with smRandom-seq separately. Source data are provided as a Source Data file.

highest expression genes in the *B. subtilis* datasets of the species mixture by smRandom-seq and the public bulk RNA-seq dataset<sup>23</sup> (Supplementary Fig. 3f). We also performed a three-species mixture experiment of *A. baumannii*, *K. pneumoniae*, and *E. coli* by smRandom-seq (Supplementary Fig. 4a). All these species clustered separately, resulting from uniform manifold approximation and projection (UMAP)'s reduction (Supplementary Fig. 4b). The inter-species double rate of the three-species mixture experiment was 2.8%. However, the current smRandom-seq used in species-mixing experiments showed bias in cell number, UMI count, and detected genes due to variations in bacterial size, RNA content, and cell wall composition. We next tested other common clinical pathogenic bacteria (including Gram-negative bacteria: *E. coli*, *A. baumannii*, *K. pneumoniae*, *P. aeruginosa*, and Gram-positive bacteria: *S. aureus*) to confirm the general applicability of smRandom-seq. smRandom-seq method displayed desirable performance in these different bacterial species. A median of 225 genes and 428 UMIs was detected in 5001 *E. coli*; a median of 204 genes and 307 UMIs was detected in 1105 *A. baumannii*; a median of 321 genes and 610 UMIs was detected in 1127 *K. pneumoniae*; a median of 245 genes and 324 UMIs was detected in 2973 *P. aeruginosa*; a median of 206 genes and 393 UMIs was detected in 1989 *S. aureus* (Fig. 2c, Supplementary Fig. 5a). As shown on the UMAP plot, these bacterial species are also clustered separately (Supplementary Fig. 5b). As expected, smRandom-seq captured all biotypes of bacterial RNA, including mRNA, rRNA, tRNA, and ncRNA (Source data Fig. 2d). However, the mapped reads of bacterial libraries were dominated by rRNA (~80%) (Fig. 2d). To reduce the sequencing cost, we applied depletion of abundant sequences by hybridization (DASH) to deplete rRNA-derived cDNAs by CRISPR/Cas9<sup>24</sup>. We designed single guide RNAs (sgRNA) according to the *E. coli* rRNA sequencing results obtained above and optimized experimental conditions as described<sup>24,25</sup>. The rRNA-depletion treatment reduced the rRNA proportion from 83% to 32% (Fig. 2d). Correspondingly, mRNA proportion was multiplied four times (16–63%) (Fig. 2d). As a result, the sequencing sensitivity of smRandom-seq was increased by rRNA-depletion (Fig. 2e, Supplementary Fig. 6a). The medians of UMI count and detected genes of *E. coli* with rRNA depletion were 685 and 374, respectively. While the medians of UMI count and detected genes of *E. coli* without rRNA depletion were 332 and 192, respectively. CRISPR-based rRNA depletion did not change the gene expression pattern compared with the control ( $R = 0.99$ ,  $p < 2.2e-16$ ) (Fig. 2f). We also performed the Cas9-based rRNA depletion in other bacterial species, including *A. baumannii*, *K. pneumoniae*, *E. coli*, and *S. aureus*, and a mixed population of *A. baumannii*, *K. pneumoniae*, and *E. coli*. The rRNA-depletion treatment reduced the rRNA proportion from 71% to 10%, 88% to 29%, 74% to 4%, 91% to 45%, 83% to 5%, respectively (Supplementary Fig. 6b–f), which proved the broad applicability of our method. This rRNA-depleted *E. coli* library was further sequenced and displayed an obvious drop-off point (red dashed line) determining the threshold for actual bacteria with noise (Fig. 2g). Saturation analysis showed that the median of detected genes per bacterium continued increasing after 10k uniquely aligned reads per bacterium (Fig. 2h). The median of detected genes per bacterium could achieve ~1000 at the sequencing depth of ~20k mapped reads per bacterium (Fig. 2h). The performance of our smRandom-seq method is well above the published high-throughput single-bacterium RNA-seq methods, PETRI-seq<sup>15</sup> and microSPLIT<sup>16</sup> (Fig. 2i). The average coverage across all transcripts at each percentile of length gradually decreases from 5' to 3' end (Supplementary Fig. 7), which can be explained by the 5'–3' direction of cDNA synthesis during RNA capturing.

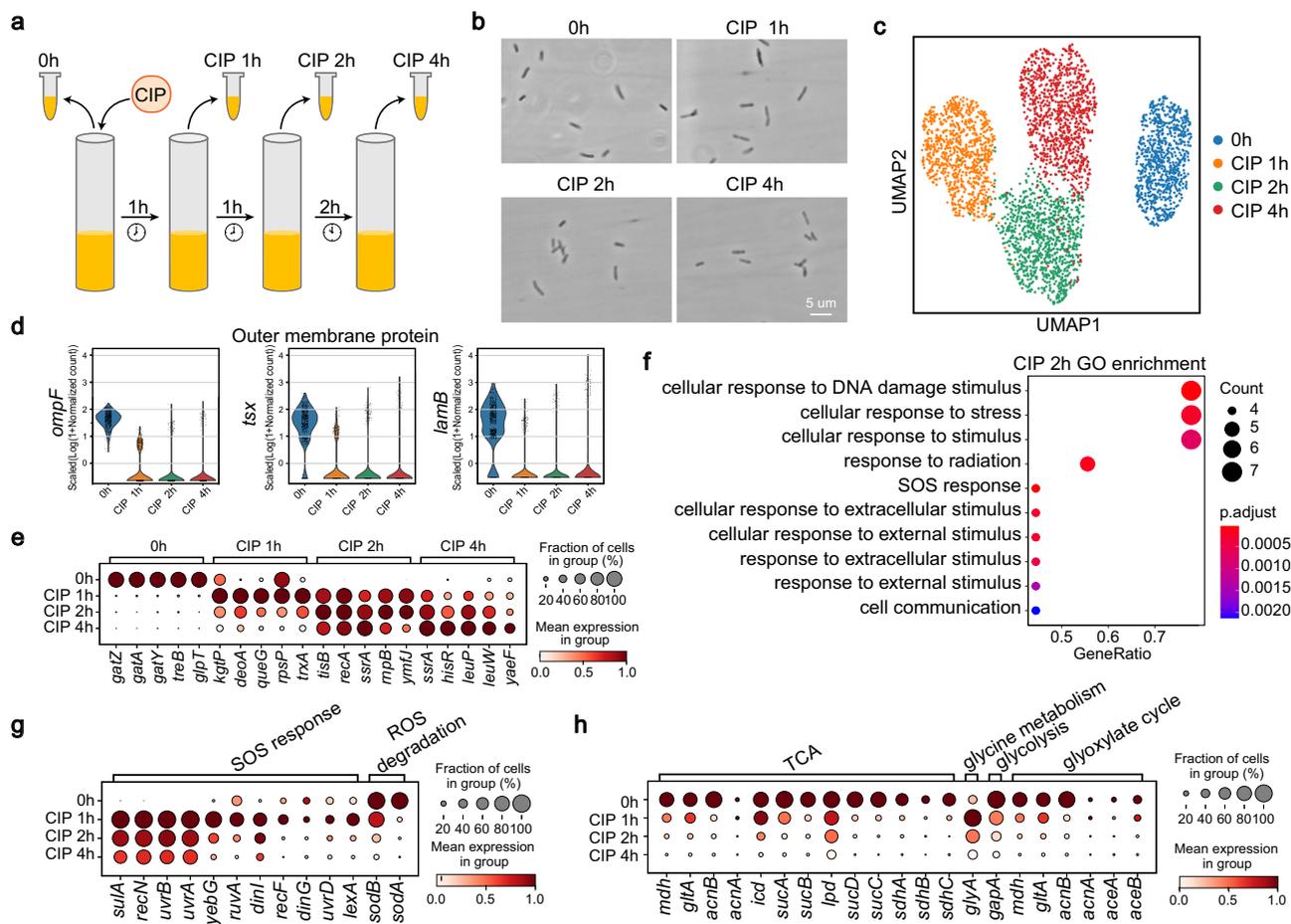
To evaluate the accuracy of gene expressions by smRandom-seq, green fluorescence protein (GFP), as a quantitative reporter, was induced by propionate in *E. coli* with a GFP plasmid (E1–E4 samples: 0, 6.25, 25, and 100 mM propionate, respectively). GFP fluorescence increased, and Ct values of the GFP coding gene in qPCR measurement

decreased as the amount of propionate increased (Supplementary Fig. 8a, b). Consistent with the GFP fluorescence values and GFP mRNA expression levels, the median UMI counts of the GFP gene detected by smRandom-seq also increased along with the increased amount of propionate and correlated well with the relative GFP gene expression levels measured by fluorescent measurement ( $R = 0.97$ ,  $p = 0.033$ ) and qPCR ( $R = 1$ ,  $p = 0.0018$ ) (Fig. 2j, Supplementary Fig. 8c, d), suggesting that the gene expression by smRandom-seq is consistent with that by qPCR.

To prove the technical reproducibility of smRandom-seq, two repeated exponential-phase *E. coli* samples from the same culture tube were applied with smRandom-seq separately (Supplementary Fig. 9a). These two independent results displayed proper alignment visualizing with UMAP dimensionality reduction (Fig. 2k) and had a high correlation on gene expression ( $R = 0.95$ ,  $p < 2.2e-16$ ) (Supplementary Fig. 9b). These results verified that smRandom-seq could efficiently and accurately capture total RNAs in a single bacterium and is applicable to multiple bacteria species.

### smRandom-seq revealed transcriptome changes of single *E. coli* upon antibiotic stress

smRandom-seq was next applied to study the transcriptome programs in single *E. coli* under antibiotic stress. We chose ciprofloxacin (CIP), a fluoroquinolone antibiotic commonly used to treat *E. coli* infections. CIP induces antibacterial activity primarily via induction of double-stranded DNA breaks, as well as induction of oxidative stress<sup>26</sup>. Exponential stage *E. coli* was treated with a 15  $\mu\text{g}/\text{mL}$  concentration of CIP, which could induce a decline of the growth curve upon antibiotic stress (Supplementary Fig. 10). The survived *E. coli* were sampled after 0, 1, 2, and 4 h, and applied with smRandom-seq, respectively (Fig. 3a). The morphology of bacteria in each group was homogeneous (Fig. 3b). The UMI count and detected genes in these CIP-treated *E. coli* samples gradually decreased after 1, 2, and 4 h (Supplementary Fig. 11a, b). There were 5 overlapped gene IDs among the top 10 highest expression genes in the *E. coli* datasets of the CIP-T0 sample, the species-mixture by smRandom-seq, and the public bulk RNA-seq dataset<sup>27</sup> (Supplementary Fig. 11c). Unsupervised graph-based clustering resulted in four distinct clusters, while the gene expression in each cluster was continuous (Fig. 3c), which is consistent with the morphologic feature. The strongly decreased expression of *ompF* was consistent with the reported CIP resistance mechanism through the decreased amount of porin outer membrane protein OmpF<sup>28</sup> (Fig. 3d). Gene expressions of other two outer membrane protein-encoding genes, *tsx* and *lamB*, were also reduced (Fig. 3d). We ranked marker genes in each group and identified the top 5 differentially expressed genes (DEGs) to characterize the transcriptional profile of each *E. coli* sample (Fig. 3e). Among the top 5 DEGs for 2 h time point upon CIP treatment, *recA* was reported to promote DNA repair and recombination, contributing the ciprofloxacin resistance<sup>29</sup>, and *tisB* was reported to be a persister gene that induces reversible dormancy by shutting down cell metabolism<sup>30</sup>. Consistently, the top 20 DEGs of group CIP 2 h were enriched in the cellular response to DNA damage stimulus (Fig. 3f). We next analyzed the expression patterns of genes involved in SOS response and metabolic pathways (Fig. 3g, h). Genes involved in SOS response displayed an acute increase upon CIP treatment, especially at 1 h, which is consistent with the fact that fluoroquinolones are potent inducers of the SOS response<sup>31</sup> (Fig. 3f). The inhibited reactive oxygen species (ROS) degradation genes upon CIP treatment were consistent with reported ROS accumulation induced by CIP<sup>32</sup> (Fig. 3g). The expression levels of metabolic genes were significantly reduced with treatment time (Fig. 3h), suggesting a metabolic shutdown in CIP-treated *E. coli*, avoiding metabolic toxicity, and minimizing drug lethality<sup>33</sup>, which is also consistent with the reduced UMI count and detected genes in those samples showed above (Supplementary Fig. 11a, b). These results suggested that smRandom-seq could sensitively capture the



**Fig. 3 | smRandom-seq captured transcriptome changes of single *E. coli* upon antibiotic stress.** **a** Experimental design for CIP (ciprofloxacin, 15  $\mu\text{g}/\text{mL}$ )-treated *E. coli* samples. **b** Images of *E. coli* treated with CIP for 0, 1, 2, and 4 h.  $n = 4$  independent experiments. **c** UMAP projection of all the bacteria collected at the different time points, based on their gene expression colored by time point. **d** Violin plot showing the expression levels of outer membrane protein-encoding genes (*ompF*, *tsx*, and *lamb*) in samples. 0 h  $n = 771$  cells, CIP 1 h  $n = 850$  cells, CIP 2 h  $n = 950$  cells, CIP 4 h  $n = 949$  cells. The expression levels of each gene in each cell were normalized, log-transformed ( $\log(1+x)$ ), and scaled data. The violin plots showed the kernel density estimate of the underlying data. **e** Mean expression levels of top 5 DEGs among different samples. The color of each dot represents the mean expression within each sample and the size of each dot represents the

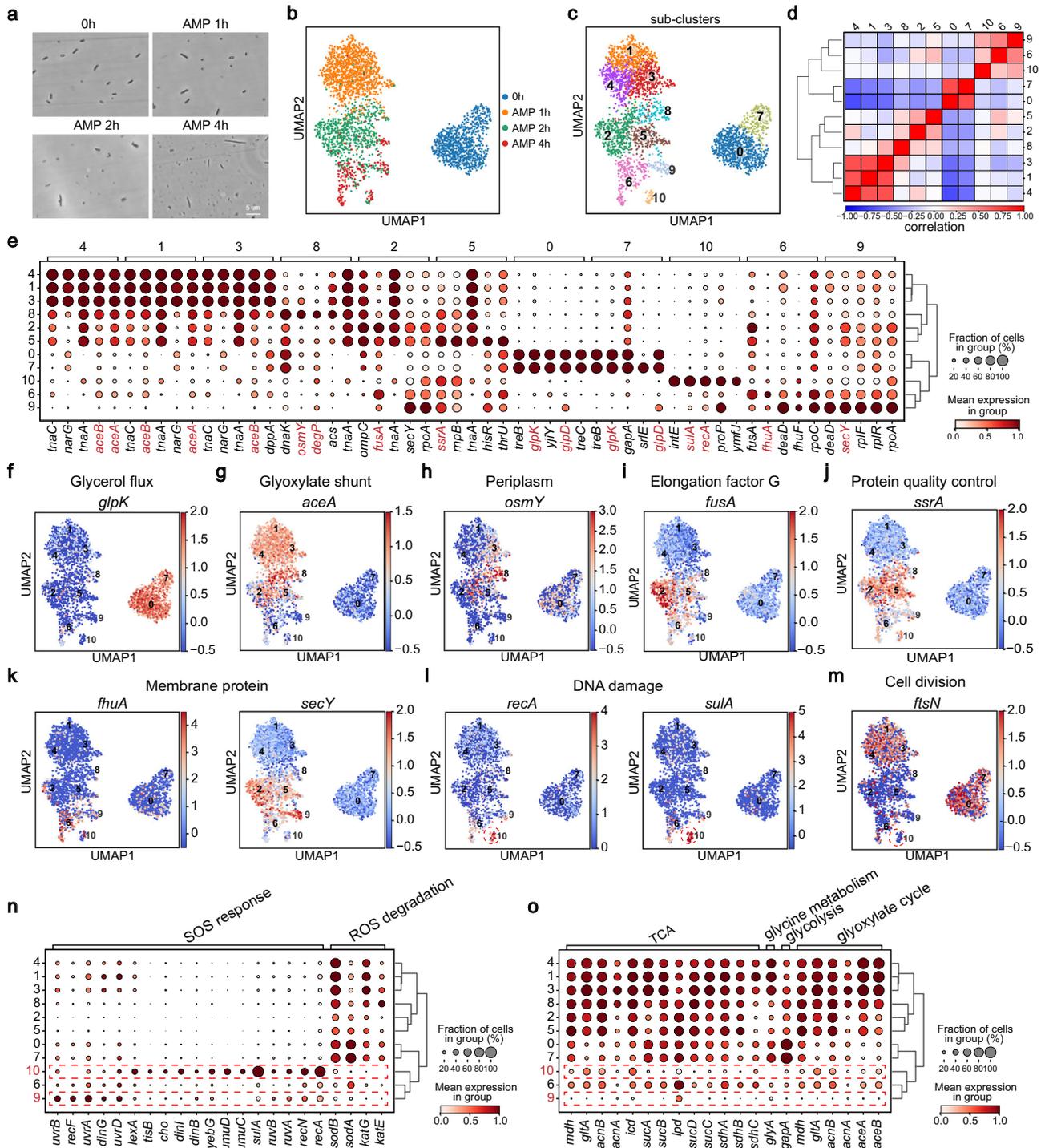
fraction of cells expressing the DEGs in different samples. **f** Bubble plots of GO enrichment analysis of top 20 DEGs in CIP 2 h sample. Gene count: the number of DEGs enriched in a GO term. Gene ratio: the number of observed DEGs divided by the number of expected genes from each GO term. The cutoff value of the  $p$ -value of GO enrichment was 0.05 and the cutoff value of the  $q$ -value of GO enrichment was 0.05. **g, h** Mean expression levels of genes involved in SOS-response and reactive oxygen species (ROS) degradation (**g**) and different metabolic pathways (**h**) in different time point samples. The color of each dot represents the mean expression within each sample and the size of each dot represents the fraction of cells expressing the specific genes in different samples. TCA tricarboxylic acid cycle. Source data are provided as a Source Data file.

transcriptome changes of *E. coli* upon antibiotic stress at the single-bacterium level.

### smRandom-seq discovered antibiotic-resistant subpopulations in *E. coli* upon antibiotic stress

To further characterize the heterogeneous response of bacteria upon another antibiotic that is widely used in human and livestock *E. coli* infection, we first treated exponential-stage *E. coli* with ampicillin (AMP), a semi-synthetic  $\beta$ -lactam antibiotic, at a concentration (5  $\mu\text{g}/\text{mL}$ ) inducing apparent bacteria death according to the growth curve (Supplementary Fig. 12a). All the *E. coli* showed a clear elongation after 4 h treatment of AMP at 5  $\mu\text{g}/\text{mL}$  (Supplementary Fig. 12b). We next increased the concentration of AMP (12.5  $\mu\text{g}/\text{mL}$ ) and found that the morphology of the survived *E. coli* displayed an increasing heterogeneity towards the 4-h time point (Fig. 4a). To profile the transcriptome of these survived subpopulations, we performed smRandom-seq on these *E. coli* samples treated with 12.5  $\mu\text{g}/\text{mL}$  AMP for 0, 1, 2, and 4 h. The UMI count and detected genes of AMP-treated *E. coli* decreased after 2 and 4 h (Supplementary Fig. 13a, b). The

separated clusters (Fig. 4b) and the top 10 DEGs (Supplementary Fig. 14a) of the four-time points suggested that AMP treatment induced significant changes in gene expression patterns over time. For example, the DNA damage-related genes, *recA* and *sulA*, were significantly raised at the 4 h time point (Supplementary Fig. 14a), which is consistent with the increased proportions of dead cells and the decreased cell viabilities after AMP treatment (Supplementary Fig. 15a, b). Furthermore, different subpopulations of these AMP-treated samples were algorithmically clustered using the Leiden algorithm (Fig. 4c). Consistent with the bacterial morphology heterogeneity, gene expression patterns of these 12.5  $\mu\text{g}/\text{mL}$  AMP-treated bacteria showed distinct subclusters at 2 and 4 h time points as visualized by a UMAP dimensionality reduction (Fig. 4c). The subclusters mainly from 2 h (subclusters 2, 5, and 8) and 4 h (subclusters 6, 9, and 10) time points showed low correlations of gene expression patterns (Fig. 4d), which suggested that the gene expression patterns diverge as the bacteria were exposed to AMP over time. Therefore, we ranked genes for characterizing these subclusters and identified the top 5 DEGs to further understand the heterogeneity of *E. coli* upon AMP treatment



**Fig. 4 | smRandom-seq identified subpopulations reacting differently to ampicillin.** **a** Images of *E. coli* treated with AMP (ampicillin, 12.5  $\mu\text{g}/\text{mL}$ ) for 0, 1, 2, and 4 h.  $n = 4$  independent experiments. **b, c** UMAP projection of all the cells collected at the different time points (b) or sub-clusters (c). **d** The correlation matrix showed the correlation coefficients between sub-clusters. **e** The mean expression levels of the top 5 DEGs among different subclusters. The color of each dot represents the mean expression within each subcluster and the size of each dot represents the fraction of cells expressing the DEGs in different subclusters. **f–l** Expression levels of the glycerol flux-related marker gene of subclusters 0 and 7 (*glpK*) (f), the glyoxylates shunt-related marker gene of subclusters 1, 3 and 4 (*aceA*) (g), the periplasm-related marker gene of subcluster 8 (*osmY*) (h), the elongation factor G related marker gene of subcluster 2 (*fusA*) (i), the protein quality control related marker gene of subcluster 5 (*ssrA*) (j), the membrane related marker genes of subclusters 6 and 9 (*fhuA*, *secY*, respectively) (k), the two DNA damage related marker genes of subcluster 10

(*recA*, *sulA*) (l) in each cell overlaid on the UMAP plot. The expression levels of each gene in each cell were normalized, log-transformed ( $\log(1 + x)$ ), and scaled (zero mean and unit variance) data. Dashed red circle: the distinct sub-cluster 10. **m** Expression of cell division-related gene (*ftsN*) overlaid on the UMAP plot. Dashed red circle: the distinct sub-cluster 10. **n** Mean expression levels of genes involved in SOS-response and ROS degradation in different sub-clusters. Dashed red rectangle: the distinct gene expression patterns of sub-clusters 9 and 10, which are associated with a 4 h time point. **o** Mean expression levels of genes involved in different metabolic pathways in different subclusters. The color of each dot represents the mean expression within each subcluster and the size of each dot represents the fraction of cells expressing the genes involved in different metabolic pathways in different subclusters. TCA tricarboxylic acid cycle. Dashed red rectangle: the distinct gene expression patterns of sub-clusters 9 and 10, which are associated with 4 h time point.

(Fig. 4e). We selected these antibiotic response-related DEGs (including *aceA*, *aceB*, *osmY*, *degP*, *fusA*, *ssrA*, *glpD*, *glpK*, *sulA*, *recA*, *fhuA*, and *secY*) representing each subcluster (Fig. 4e). Further, we investigated the feature expression in each bacterium in low-dimensional space (UMAP) (Fig. 4f–l, Supplementary Fig. 14b–d). The aerobic fermentation of glycerol-associated genes (*glpD* and *glpK*) highly and specifically expressed at 0 h time points (subclusters 0 and 7) (Fig. 4f, Supplementary Fig. 14b), followed by general high expression of the genes (*aceA* and *aceB*) that encode enzymes of the glyoxylate shunt at the 1 h time point (subclusters 1, 3 and 4) (Fig. 4g, Supplementary Fig. 14c). These results suggested that the aerobic fermentation of *E. coli* run at high growth rates at the 1 h time point. As time went on, *aceA* and *aceB* showed distinct expression patterns among subclusters, especially the highest upregulation in subcluster 8 of the 2 h time point (Fig. 4g, Supplementary Fig. 14c). Subcluster 8 displayed remarkable expressions of genes (*osmY* and *degP*) related with periplasm (Fig. 4h, Supplementary Fig. 14d), which are critical for bacterial stress resistance<sup>34,35</sup>. Subcluster 2 displayed a remarkable expression of *fusA* (Fig. 4i), encoding elongation factor G, a critical target during oxidative damage to the translation system of *E. coli*<sup>36</sup>. Subcluster 5 was active in protein-quality control of translation with a high expression of *ssrA* (Fig. 4j) that mediates degradation of aberrant, unfinished proteins when *E. coli* ribosomes stall during translation<sup>37</sup>. These results suggested that subpopulations of *E. coli* undergo different degrees of membrane integrity and translation damage upon 2-hour AMP treatment. Subclusters 6 displayed a high expression level of *fhuA* which encodes *E. coli* ferric hydroxamate protein uptake component A, and subclusters 9 displayed a high expression level of *secY* which encodes nonspecific sugar transporters in *E. coli*, respectively (Fig. 4k). The membrane protein-encoding gene, *secY*, also had high expression level in subcluster 2 of 2-h time point, suggesting that subcluster 9 might be developed from subcluster 2. Gene Ontology (GO) enrichment analysis revealed that the marker genes (*rpoC*, *rplF*, *rplR*, *rpoA*) of subclusters 6 and 9 also enriched in ribosome assembly (Fig. 4e, Supplementary Fig. 14e, f). Severe DNA damage of bacteria in subcluster 10 could be inferred, as *recA* and *sulA* both showed an apparent enrichment in subcluster 10, where the former is essential in DNA repair<sup>38</sup> and the latter is inducible by DNA damaging agents<sup>39</sup> (Fig. 4l). The top 20 DEGs of subcluster 10 were mainly enriched in cellular response to stimulus and SOS response (Supplementary Fig. 14g), which is supported by the fact that both *recA*<sup>40</sup> and *sulA*<sup>39</sup> are SOS response regulators. Consistent with that SOS response aids bacterial propagation by inhibiting cell division during repair of DNA damage<sup>40</sup>, cell division gene *ftsN* was inhibited at the corresponding point on the UMAP plot (Fig. 4m). More SOS response- and ROS degradation-related genes displayed different expression patterns in subclusters 6, 9, and 10, which were associated with the 4-h time point (Fig. 4n). In addition, GO enrichment analysis revealed enrichment in metabolic process pathways in subclusters 6 and 9 (Supplementary Fig. 14e, g). The expression levels of metabolic-related genes were increased after 1 h of antibiotic treatment and then decreased at 2- and 4-h time points (Supplementary Fig. 14h), which is consistent with the trend of the average respiratory activity of *E. coli* first increasing and then decreasing after antibiotic treatment (Supplementary Fig. 16a, b). At the 4-h time point, we observed strong downregulation of metabolic-related genes in sub-clusters 9 and 10 but upregulation in sub-cluster 6 (Fig. 4o). These findings suggest a low metabolic activity that can generate ROS, which is consistent with the low expression of ROS degradation genes and high expression of SOS response genes in sub-clusters 9 and 10 (Fig. 4n). These findings were also consistent with the observation that a sub-population of bacteria at the 4-h time point displayed high respiratory activity, as evidenced by high levels of CTC fluorescence intensity (Supplementary Fig. 16b). These results showed that *E. coli* population responded differently to high-concentration AMP. This heterogeneity is ignored by traditional bulk RNA-seq methods but well characterized by smRandom-seq.

## Discussion

In this study, we developed a droplet-based, high-throughput, and high-sensitivity single-microbe RNA-seq method, named smRandom-seq, with the pipeline of bacteria preparation, reactions in situ, droplet barcoding, and library preparation. We provided sufficient and systematic evidence to prove the efficiency and technical reproducibility of smRandom-seq and validated the performance of smRandom-seq by practical application.

We used a GAT three-letter code primer with seven random nucleotides as the RT primer, which can evenly and multiply hybridize to total RNA templates at low temperature<sup>41</sup>. The above results proved that this adapted RT primer is efficient for the reverse transcription of fixed bacterial RNA with a complex structure inside a single bacterium. Otherwise, the current popular scRNA-seq platform 10X Genomics<sup>7</sup>, as well as the single microbe RNA-seq method PETRI-seq<sup>15</sup> and microSPLIT<sup>16</sup>, used a template-switching strategy for the second strand cDNA synthesis. The template-switching strategy could be affected by the complex structure of RNAs and fail to capture interrupted cDNAs<sup>42</sup>. Thus, we adapted a dA tailing strategy for efficient second-strand cDNA synthesis, as non-template dA can be added to the 3' OH terminus of all cDNA fragments, including interrupted products, by terminal deoxynucleotidyl transferase. Unlike the multiple split-pool ligation-based barcoding in PETRI-seq<sup>15</sup> and microSPLIT<sup>16</sup>, the barcoding strategy of smRandom-seq uses an efficient and stable droplet microfluidics platform to co-encapsulate single bacterium with uniquely barcoded beads. Recently, three bacterial single-cell RNA-seq methods, ProBac-seq<sup>43</sup>, M3-Seq<sup>44</sup>, and BacDrop<sup>45</sup> were reported using the commercially 10X Genomics microfluidic single-cell sequencing platform. To efficiently barcode the low-content RNAs in the single bacterium, we optimized our previous droplet barcoding platform designed for single-cell barcoding<sup>20</sup>, by generating smaller barcoded beads and droplets and modifying the primers releasing and the barcodes synthesis strategy. These optimized strategies in the smRandom-seq method come out high-quality cDNA library with 200–500 bp size from a single bacterium, which is not needed to fragment but is just suitable for the next generation sequencing (NGS). A pre-index strategy could be easily combined with the smRandom-seq by adding indices to RT primers according to the published scifi-RNA-seq<sup>46</sup>, where the bacteria were split into different tubes for RT with pre-indexed random primers, and then those pre-indexed samples could be combined into one sample for the following barcoding procedures. These double-end barcoded cDNAs need to be directly paired-end sequenced without fragmentation. Pre-indexed smRandom-seq would not only extend the throughput of our method but also reduce spurious biological conclusions of doublets arising during microfluidics encapsulation. In addition, this pre-indexed smRandom-seq will be suitable for processing multiple samples simultaneously.

The efficiency of smRandom-seq is mainly affected by an overwhelming quantity of rRNAs (>80% of mapped reads) which are usually not of interest. To avoid the cost of sequencing non-messenger transcripts in such large-scale bacterial transcriptomics studies, here we incorporated an rRNA depletion step at the library level according to the reported DASH method<sup>24,25</sup>. Our results indicated that this CRISPR-based technique markedly improved the performance of smRandom-seq on single *E. coli* RNA sequencing with a median of ~1000 genes detected per bacterium by efficiently removing rRNA-derived cDNAs. Recently, Homberger et al. also posted successful Cas9-based rRNA depletion for the libraries of *Salmonella* by a low-throughput bacterial single cell RNA-seq method<sup>47</sup>. Wang et al. reported another post-hoc rRNA depletion scheme for M3-Seq<sup>44</sup>, which used a duplex-specific nuclease to remove the rRNAs hybridized with DNA probes targeting rRNAs of specific bacterial species. BacDrop depleted rRNAs in cells prior to reverse transcription also using the duplex-specific nuclease<sup>45</sup>. In order to apply smRandom-seq to other microbial species and even

more complex natural communities, a universal microbiome rRNA depletion method<sup>48</sup> is needed to be incorporated into smRandom-seq.

We also applied smRandom-seq to profile transcriptome changes of single *E. coli* under antibiotic treatment. These results found a few antibiotic-resistant subpopulations in antibiotic-treated *E. coli* displaying distinct gene expression patterns of SOS response and metabolic pathways. Our study provides a way to tease out how individual bacterium adapts and interacts with each other under environmental stress. We can then efficiently predict which subpopulations will evolve resistance and survive to impact health, as well as figure out the mechanisms of bacterial resistance and persistence, which would be valuable for precision diagnosis and treatment of bacterial infection. Moreover, combined with the Combi-Seq<sup>49</sup> which uses a combinatorial DNA barcoding approach to encode treatment conditions in droplets, smRandom-seq would probably allow monitoring the bacterial transcriptomic changes at a single bacterium level for high-throughput drug screening. We also anticipate that smRandom-seq would be broadly applied in revealing heterogeneous transcriptome profiles of microbiota, identifying rare species, and deciphering cross-species interactions in more varied environments, such as gut microbiota and soil microbiota.

## Methods

### Bacterial culture

*Escherichia coli* BW25113 (*E. coli*), *Bacillus subtilis* NCBI3610 (*B. subtilis*), *Acinetobacter baumannii* ATCC17978 (*A. baumannii*), *Klebsiella pneumoniae* XH209 (*K. pneumoniae*), *Pseudomonas aeruginosa* PAO1 (*P. aeruginosa*) and *Staphylococcus aureus* subsp. *Aureus* SA268 (*S. aureus*) was provided by Sir Run Run Shaw Hospital, Zhejiang University School of Medicine. Overnight cultured bacteria were inoculated into fresh LB Lennox medium at 1:1000, respectively, and grown at 37 °C with shaking at 250 rpm.

For the two-species-mixing experiment, cultures of *E. coli* (Gram-negative bacteria) and *B. subtilis* (Gram-positive bacteria) were sampled upon reaching the OD<sub>600</sub> - 0.5, immediately centrifuged at 4 °C, 2000×g for 2 min, next washed twice by PBS, and mixed before applied with smRandom-seq.

For the three-species-mixing experiment, cultures of *E. coli*, *A. baumannii* and *K. pneumoniae* were sampled upon reaching the OD<sub>600</sub> - 0.5, immediately centrifuged at 4 °C, 2000×g for 2 min, next washed twice by PBS, and mixed before applied with smRandom-seq.

For single-species experiments, cultures of Gram-negative bacteria: *E. coli*, *B. subtilis*, *A. baumannii*, *K. pneumoniae*, *P. aeruginosa*, and Gram-positive bacteria: *S. aureus* were sampled upon reaching the OD<sub>600</sub> - 0.5, immediately centrifuged at 4 °C, 2000×g for 2 min, next washed twice by PBS, and applied with smRandom-seq, respectively.

For the GFP quantitation experiment, *E. coli* harboring pPro24-gfp<sup>50</sup> (GFP-*E. coli*) was inoculated into fresh LB medium supplemented with ampicillin (100 µg/mL) and grown at 37 °C. When the OD<sub>600</sub> reached 0.5, GFP-*E. coli* samples were induced with 0, 6.25, 25, and 100 mM propionate (E1-E4 samples, respectively), respectively. GFP fluorescence was measured in a SpectraFluor Plus plate reader (Tecan) using an excitation wavelength of 405 nm and an emission wavelength of 535 nm. GFP fluorescence was normalized for cell density (GFP fluorescence/OD<sub>600</sub>).

The bactericidal effect experiments for CIP and AMP were prepared by inoculating overnight cultured *E. coli* into fresh LB medium and growing until OD<sub>600</sub> - 0.1 was reached. CIP was added to a final concentration of 0, 6.25, 12.5, 25, and 50 µg/mL, respectively. AMP was added to a final concentration of 0 and 5 µg/mL, respectively. The optical density at 600 nm (OD<sub>600</sub>) of bacterial cell suspensions at different stages was measured with an xMark™ microplate absorbance spectrophotometer (Bio-Rad).

The antibiotic-treated samples for smRandom-seq were prepared by inoculating overnight cultured *E. coli* into fresh LB medium and growing until OD 0.5 was reached. 0 h sample was taken, and 15 µg/mL CIP, 5 µg/mL AMP, or 12.5 µg/mL AMP was added to the rest *E. coli*, respectively. The following samples were taken from the culture at 1, 2, and 4 h time points, respectively. Samples were centrifuged at 4 °C, 2000×g for 2 min, washed twice with PBS, and applied with smRandom-seq, respectively.

### Fixation and permeabilization

Bacteria were fixed in 4% PFA (Beyotime Biotechnology, Cat. P0099) at 4 °C overnight, then washed and permeabilized by 0.04% Tween-20 (Sangon Biotech, Cat. 9005-64-5) in PBS. The cell wall was digested with lysozyme by incubating at 37 °C for 15 min. The lysozyme was ordered from Thermo Fisher (Cat. 90082). Following the cell wall digestion step, bacteria were immediately washed and resuspended in PBS with RNase inhibitor (Thermo Fisher Scientific, Cat. N8080119) for the next step.

### In situ reverse transcription

In situ, reverse transcription of bacteria was performed according to the procedures of the reverse transcription kit. The reverse transcription kit (Cat. R20114124), including reverse transcriptase (50 U/µL), 5× reverse transcription buffer, and dNTP Mix (100 mM), was ordered from M20 Genomics. The RT reaction mix (per 50-µL reaction) was prepared on ice, including 1–10 million bacteria in 27.5 µL PBS, 10 µL 5× reverse transcription buffer, 5 µL 10 µM random primer (sequence in Supplementary Table 1), 2.5 µL 100 mM dNTP, 2.5 µL RNase inhibitor, 2.5 µL reverse transcriptase (50 U/µL), and incubated with twelve cycles of multiple annealing ramping from 8 to 42 °C, followed by a 42 °C 30 min, in a thermal cycler. After reverse transcription, bacteria were washed with PBST (PBS with 0.05% Tween-20) five times.

### In situ dA tailing

For In situ dA tailing of bacteria, the following reaction mix was prepared, including a few million bacteria in 39 µL PBS, 5 µL 10X TdT buffer, 5 µL 2.5 mM CoCl<sub>2</sub>, 0.5 µL 100 mM dATP, 0.5 µL TdT enzyme, and incubated at 37 °C for 30 min, and then washed with PBST three times. The TdT reaction kit (including TdT enzyme, 10X TdT buffer and CoCl<sub>2</sub>) was ordered from New England Biolabs (Cat. M0315S).

### Microfluidic device fabrication

PDMS-based microfluidic devices were designed for hydrogel beads synthesis and single bacterium barcoding, and fabricated<sup>22,51</sup>. The channel depth of microfluidic devices for hydrogel beads synthesis is 30 µm (Supplementary Fig. 2a). The channel depth of microfluidic devices for cell encapsulation is 50 µm (Supplementary Fig. 1b). Molds for a microfluidic device were made using a photolithographic approach, consisting of centrifugally coating and modeling the SU-8. The silicon molds were cast with PDMS (Sylgard-184) to fabricate microfluidic devices.

### Microfluidic platform

The microfluidic platform for hydrogel beads synthesis and single bacterium barcoding was established<sup>22,51,52</sup>, which includes microfluidic devices, two or three syringe pumps, some syringes (1 mL) and tubing, and an inverted bright-field microscope with a fast speed camera and a computer.

### Barcoded beads synthesis

The hydrogel barcoded beads for single bacterium barcoding were designed<sup>22,51,52</sup> (Supplementary Table 1) and customized by M20 Genomics company. Hydrogel beads were synthesized by the microfluidic emulsification and polymerization of an acrylamide-primer mix.

The acrylamide-primer mix contains acrylamide:bis-acrylamide solution (1×) (Invitrogen, Cat. AM9022), acrydite-modified oligonucleotides (50 μM) that is covalently incorporated into the hydrogel mesh, ammonium persulfate (APS, 10% wt/vol, Sangon Biotech Cat. A100486-0025) that initiates the free-radical polymerization, and Tris-buffered saline-EDTA-Triton (TBSET) buffer (1×). The photocleavable moiety contained in the acrydite-modified oligonucleotide on previously reported hydrogel beads<sup>22</sup> was replaced by a deoxy Uridine base in our customized protocol. A carrier oil-TEMED mix (1 mL:4 μL) was applied in microfluidic. The acrylamide-primer mix and carrier oil-TEMED mix were transferred to syringes, respectively. Then, the syringes were connected to corresponding inlets of the 30-μm-deep hydrogel bead synthesis device (Supplementary Fig. 2a) via tubing, respectively. The flow rates for the acrylamide-primer mix and carrier oil-TEMED mix were 900 and 1800 μL/h, respectively. The size of generated hydrogel beads should be 40 μm (Supplementary Fig. 2b). Next, the DNA primers on the generated hydrogel beads are barcoded using a combination of a split-and-pool method and a 3-step primer ligation reaction (Supplementary Fig. 2c), instead of the previous reported 2-step extension reaction<sup>22</sup>. The unique barcoded primers were provided in Supplementary Table 2. Hydrogel beads mix, including hydrogel bead, DNA ligase (350 U/mL, New England Biolabs, Cat. M0202S), dNTP (10 mM each), 1× isothermal amplification buffer (New England Biolabs, Cat. B0537S), nuclease-free water was prepared, then split into a round-bottom 96-well plate. The hydrogel beads mix in 96-well plates were mixed with 96 annealed unique barcode primers in 96-well plates, respectively, then incubated at 37 °C for 30 min. All hydrogel beads were combined in a single 50-mL tube and washed with STOP-25 buffer<sup>22</sup>, then performed with the second (96 unique barcoded primers) and third (48 unique barcoded primers) split-and-pool rounds. For quality control of the generated hydrogel barcoded beads, barcoded primers were released by USER (New England Biolabs, Cat. M5505S) enzymic digestion and analyzed with gel electrophoresis (Supplementary Fig. 2d). The highest-molecular-weight peak (that typically appears at -100 bp) in the electropherogram represents the full-length barcoding primer (actual size 96 nt), and lower-molecular-weight peaks are synthesis intermediates (actual size 27, 44, 58 nt) (Supplementary Tables 1, 2). All the required reagents for hydrogel barcoded beads synthesis and the ready-to-use hydrogel barcoded beads can be ordered from the M20 Genomics company.

### Droplet barcoding

The modified droplet barcoding for a single bacterium was performed according to previous work<sup>21,52</sup>. The qualified single bacteria were counted under an optical microscope and diluted with a 15% density gradient solution. Bacteria, 2X DNA extension reaction mix, and hydrogel barcoded beads were encapsulated using the microfluidic platform. The density gradient solution and 2X DNA extension reaction mix were ordered from M20 Genomics. The collected droplets were incubated at 37 °C for 1 h, 50 °C for 30 min, 60 °C for 30 min, and 75 °C for 20 min, and then broken by mixing with PFO buffer. The oil phase was discarded, and the aqueous phase containing cDNAs was purified by Ampure XP beads. The purified cDNAs were amplified by PCR with PCR1 and PCR2 primers (Supplementary Table 1), purified by Ampure XP beads, and quantified by Qubit 2.0 (Thermo Fisher Scientific).

### Ribosomal RNA depletion

We used Cas9-based depletion of abundant sequences by hybridization (DASH) treatment to remove the majority of ribosomal cDNA<sup>24,25,53</sup>. Target cDNA sequences from *E. coli* rRNA genes were identified, and sgRNAs were designed to target these sequences<sup>25,53</sup> (Supplementary Data 1). DNA templates for sgRNAs based on an optimized scaffold were made<sup>24</sup> and transcribed in vitro into sgRNAs using T7 RNA polymerase. The sgRNA pools were purified with VAHTS

RNA Clean Beads, aliquoted, and stored at -80 °C. T7 RNA polymerase was ordered from ThermoScientific (Cat. no. EP0111). VAHTS RNA Clean Beads were ordered from Vazyme (Cat. no. N412-01). Cas9 protein was ordered from NEW ENGLAND BIOLABS (Cat. no. M0386S). A reaction mixture of Cas9 and the sgRNA pool were preincubated at 37 °C for 15 min. Next, 1 ng purified PCR product was added to the Cas9-sgRNA complex. The ratio of the cDNA:Cas9 enzyme:sgRNA is 1:100:1000. The cDNA-Cas9-sgRNA complex incubating at 37 °C for 2 h. After the digestion, Cas9 was inactivated by treatment with 0.8 U (-20 μg) proteinase K for 15 min at 37 °C, followed by heat-inactivation (15 min at 95 °C). sgRNA was digested by RNase If incubating at 37 °C for 15 min and 70 °C for 15 min. Proteinase K (Cat. no. P8107S) and RNase If (Cat. no. M0243S) were ordered from NEW ENGLAND BIOLABS. The resulting DASH-treated samples were PCR amplified for 10–20 cycles to select for non-ribosomal, undigested cDNAs and purified with 1X AMPure XP beads.

### Library preparation

In this study, we used VAHTS Universal DNA Library Prep Kit for Illumina V3 (Vazyme, Cat. no. ND607-03/04) to construct the library. The amplified and purified cDNAs were quantified by Qubit 2.0 and measured by Qsep100™ DNA Fragment Analyzer (Bioptic). Then the qualified cDNAs were performed with end-repair and adenylation. The reaction mixture, containing 50 ng fragmented DNA, end repair enzymes, end repair buffer, and nuclease-free water, was prepared, incubated at 30 °C for 30 min, and then inactivated at 65 °C for 30 min. Working adaptor and ligation enzymes were added to the finished end-prep reaction mixture and then incubated at 20 °C for 15 min. The ligated DNA was purified and selected with AMPure XP beads. Then, the library was amplified by PCR and purified with AMPure XP beads. The final cDNA library was quantified by Qubit 2.0 and measured with Qsep100™ DNA Fragment Analyzer. Finally, the qualified library was sequenced by the NovaSeq 6000 and S4 Reagent Kit with paired-end reads of 150.

### Processing of sequencing data

The raw sequencing data were trimmed using Cutadapt (version 3.7)<sup>54</sup>. Primer sequences were removed, and extra bases generated by the dA-tailing step were trimmed. For each Read1, UMI (8 nts) and cell-specific barcode (30 nts) were extracted. Sequenced barcodes that can be uniquely assigned to an accepted barcode with a Hamming distance of 2 nt or less were merged. Read2 was mapped to the genome of the species of interest using STAR (2.7.10a)<sup>55</sup> with default parameters. Only uniquely mapped reads were used for downstream analyses. The GenBank assembly accession of reference genomes and annotation files used in this study are as follows: GCF\_902728005.1 for *A. baumannii*; GCF\_000750555.1 for *E. coli*, GCF\_000775955.1 for *K. pneumoniae*, GCA\_000006765 for *P. aeruginosa*, GCF\_000737615.1 for *S. aureus*, GCF\_002055965.1 for *Bacillus subtilis*. Read summarization was performed using the featureCounts (2.0.1)<sup>56</sup>. FeatureCounts takes the GTF format annotation file and the alignment file (a.bam file) as input to assign mapped reads to genomic features and then counts reads. The GenBank assembly accession of GTF format annotation files was provided above. The UMI count for each gene was determined using the UMI tools (1.1.2)<sup>57</sup>.

### Identification of the putative cells

To determine the number of putative bacteria in each sample, first, the total number of detected genes was determined for each possible barcode. Barcodes are ordered from the largest to the smallest total number of detected genes and numbered in order. We plotted the log10(total number of detected genes per barcode) versus log10 (Barcode number) rank plot for each possible barcode. On the rank plot, the knee point (red dashed line) indicated the threshold for putative cells. The knee point was identified according to the

calculation results by STAR solo module in STAR (2.7.10a)<sup>55</sup>, which used the EmptyDrop-like type of filtering, and the 10 numeric parameters: nExpectedCells (1500), maxPercentile (0.99), maxMinRatio (10), indMin (45000), indMax (90000), umiMin (300), umiMinFracMedian (0.01), candMaxN (20000), FDR (0.01), and simN (10000). Barcodes on the left of the threshold were identified as bacteria, but the rest was identified as noise. Only the barcodes identified as bacteria were used for downstream analysis.

### Cell quality control

The total number of UMI counts and detected genes in every single cell were counted. A violin plot combined with a box plot was used to visualize the distribution of UMI count and detected genes of every single cell. Statistical medians of UMI count and detected genes were used to describe the datasets. The most extreme outliers were filtered. A summary of the datasets by smRandom-seq was provided including the cell numbers, thresholds, reads numbers, UMI counts, and detected genes for different biotypes (Supplementary Data 2).

### Analysis of gene expression matrix

The single-bacterium gene expression matrix of filtered data was then analyzed by the Scanpy toolkit (version 1.8.2)<sup>58</sup>, including preprocessing, visualization, clustering, and differential expression testing. We kept genes that were expressed in at least three cells and cells that have at least 10 genes detected. We used “pandas.merge” function to create a data frame of different samples, “scanpy.pp.normalize\_total” function (target\_sum = 1e4) to normalize each cell by total counts over all genes, “scanpy.pp.log1p” function with default parameters to logarithmize the data matrix, “sc.pp.highly\_variable\_genes” function (min\_mean = 0.0125, max\_mean = 3, min\_disp = 0.5) to identify highly variable genes, “sc.pp.scale” function (max\_value = 10) to scale each gene to unit variance and zero mean, “sc.tl.pca” function (svd\_solver = ‘arpack’) to perform principal component analysis (PCA), “scanpy.pl.pca\_variance\_ratio” function to plot the variance ratio and identify the number of PCs, “sc.pp.neighbors” function for clustering at a resolution of 1, “sc.tl.umap” function to reduce the dimensionality with UMAP and bin the cells into cell populations using Leiden clustering, “sc.tl.rank\_genes\_groups” function to find marker genes enriched in each cluster with a statistical test (*t*-test). The “sc.tl.rank\_genes\_groups” function returned scores underlying the computation of a *p*-value for each gene for each group. The marker gene was set to a *p*-value < 0.05 and ordered according to scores. The top 20 marker genes were retained as the DEGs. The expression of the top 5 or 10 DEGs was visualized using a dot plot. The “enrichGO” function of clusterProfiler R package (version 3.16.0) was used for GO enrichment analysis (*p*-value cutoff value = 0.05 and *q*-value cutoff value = 0.05) of the top 20 DEGs of each group in R studio (version 4.0.1). The ggpubr R package (version 0.4.0) was used to perform a correlation test and plot the correlation coefficient and the significance level in R studio (version 4.0.1). OriginPro 2021 (9.8.0.200) was used for the data analysis and graphing (Fig. 2d, h, j).

### Saturation analysis

The rRNA-depleted and deeper sequenced *E. coli* samples were used to generate the saturation curve. The saturation curve plot was generated by randomly selecting the corresponding number of raw reads from the sample library and then using the same alignment pipeline to calculate the median numbers of genes detected per cell. Sequencing reads were randomly subsampled to 22k, 18k, 13k, 9k, 4k, 2k, and 1k reads per cell using the Seqtk tool.

### Gene body coverage analysis

RSeQC86<sup>59</sup> geneBody\_coverage.py program was used to generate the gene body coverage plots. The BAM file containing aligned,

deduplicated reads and the reference BED file generated by converting the comprehensive gene annotation GFF3 file were input to the RSeQC86.

### Cell viability analysis

At 0, 1, 2, 4 h time points after 12.5 µg/mL AMP treatment, *E. coli* samples were sampled and added to fresh LB medium supplemented with 0.5 µM of propidium iodide (PI) dye. After a 5-min incubation at 37 °C, cells are washed by PBS for three times, followed by fixation with 4% PFA for 15 min. We visualized and quantified the numbers of PI-positive cells and total cells, respectively. We captured images using Leica DMI8 fluorescent microscope.

### Cell respiratory activity analysis

At 0, 1, 2, 4 h time point after 12.5 µg/mL AMP treatment, *E. coli* samples were sampled and added with fresh LB medium supplemented with 2.5 mM of 5-cyano-2,3-ditolyltetrazolium chloride (CTC) dye. After a 30-min incubation at 37 °C, cells are washed with PBS three times, followed by fixation with 4% PFA for 15 min. Total CTC fluorescence intensity was detected using Tecan SPARK fluorescence microplate reader (Ex/Em = 480/630 nm). CTC fluorescence intensity was normalized for cell density. We captured images using Leica DMI8 fluorescent microscope. The duplicate numbers of samples are indicated in figure legends. Results of *E. coli* samples used in cell viability and respiratory activity analysis experiments are presented as means ± standard deviation.

### Statistics and reproducibility

Statistical details for each experiment are provided in the figure legends. Two-tailed Student's *t*-tests were used for the calculation of *p* values. Statistical significance was established at a 95% confidence level (*p*-value < 0.05). The microfluidic encapsulation experiment, beads synthesis experiment, and micrographs experiment were repeated at least four times independently with similar results. Validation experiments of smRandom-seq were repeated three times independently. No statistical method was used to predetermine the sample size. No data were excluded from the analyses. The experiments were not randomized. The Investigators were not blinded to allocation during experiments and outcome assessment.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The smRandom-seq data generated in this study have been deposited in the Genome Sequence Archive database under accession code CRA011274. The public *B. subtilis* data used in this study is available in the Gene Expression Omnibus database under accession code GSE217916. The public *E. coli* data used in this study is available in the Gene Expression Omnibus database under accession code GSE168963. Source data are provided with this paper.

### Code availability

All script files used in the analysis in this manuscript can be downloaded from GitHub at are available at <https://github.com/wanglab2023/smRandom-seq><sup>60</sup>.

### References

1. Silverman, J. D., Washburne, A. D., Mukherjee, S. & David, L. A. A phylogenetic transform enhances analysis of compositional microbiota data. *Elife* **6**, e21887 (2017).
2. Sorek, R. & Cossart, P. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat. Rev. Genet.* **11**, 9–16 (2010).

3. Westermann, A. J. & Vogel, J. Cross-species RNA-seq for deciphering host-microbe interactions. *Nat. Rev. Genet.* **22**, 361–378 (2021).
4. Martins, B. M. C. & Locke, J. C. W. Microbial individuality: how single-cell heterogeneity enables population level strategies. *Curr. Opin. Microbiol.* **24**, 104–112 (2015).
5. Tang, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
6. [No authors listed]. Method of the year 2013. *Nat. Methods* **11**, 1 (2014).
7. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
8. Papalexis, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat. Rev. Immunol.* **18**, 35–45 (2018).
9. Suvà, M. L. & Tirosh, I. Single-cell RNA sequencing in cancer: lessons learned and emerging challenges. *Mol. Cell* **75**, 7–12 (2019).
10. Armand, E. J., Li, J., Xie, F., Luo, C. & Mukamel, E. A. Single-cell sequencing of brain cell transcriptomes and epigenomes. *Neuron* **109**, 11–26 (2021).
11. Taniguchi, Y. et al. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).
12. Giannoukos, G. et al. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* **13**, R23 (2012).
13. Wang, J., Chen, L., Chen, Z. & Zhang, W. RNA-seq based transcriptomic analysis of single bacterial cells. *Integr. Biol.* **7**, 1466–1476 (2015).
14. Imdahl, F., Vafadarnejad, E., Homberger, C., Saliba, A. E. & Vogel, J. Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria. *Nat. Microbiol.* **5**, 1202–1206 (2020).
15. Blattman, S. B., Jiang, W., Oikonomou, P. & Tavazoie, S. Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing. *Nat. Microbiol.* **5**, 1192–1201 (2020).
16. Kuchina, A. et al. Microbial single-cell RNA sequencing by split-pool barcoding. *Science* **371**, eaba5257 (2021).
17. Rosenberg, A. B. et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
18. Guo, M. T., Rotem, A., Heyman, J. A. & Weitz, D. A. Droplet microfluidics for high-throughput biological assays. *Lab Chip* **12**, 2146–2155 (2012).
19. Prakadan, S. M., Shalek, A. K. & Weitz, D. A. Scaling by shrinking: empowering single-cell ‘omics’ with microfluidic devices. *Nat. Rev. Genet.* **18**, 345–361 (2017).
20. Klein, A.M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
21. Zheng, W. et al. High-throughput, single-microbe genomics with strain resolution, applied to a human gut microbiome. *Science* **376**, eabm1483 (2022).
22. Zilionis, R. et al. Single-cell barcoding and sequencing using droplet microfluidics. *Nat. Protoc.* **12**, 44–73 (2017).
23. Wang, B., Kloet, F.v.D. & Hamoen, W. Induction of the CtsR regulon improves Xylanase production in *Bacillus subtilis*. Preprint from Research Square. <https://doi.org/10.21203/rs.3.rs-2800723/v1>. PPR: PPR646005 (2023).
24. Gu, W. et al. Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* **17**, 41 (2016).
25. Prezza, G. et al. Improved bacterial RNA-seq by Cas9-based depletion of ribosomal RNA reads. *RNA* **26**, 1069–1078 (2020).
26. van der Putten, B. C. L. et al. Quantifying the contribution of four resistance mechanisms to ciprofloxacin MIC in *Escherichia coli*: a systematic review. *J. Antimicrob. Chemother.* **74**, 298–310 (2019).
27. Ojo, O., Scott, D., Iwalokun, B., Odetoyn, B. & Grove, A. Transcriptome RNA sequencing data set of differential gene expression in *Escherichia coli* BW25113 wild-type and slyA mutant strains. *Microbiol. Resour. Announc.* **10**, e00294-21 (2021).
28. Hooper, D. C., Wolfson, J. S., Ng, E. Y. & Swartz, M. N. Mechanisms of action of and resistance to ciprofloxacin. *Am. J. Med.* **82**, 12–20 (1987).
29. Diaz-Diaz, S. et al. Effect of RecA inactivation and detoxification systems on the evolution of ciprofloxacin resistance in *Escherichia coli*. *J. Antimicrob. Chemother.* **77**, 641–645 (2022).
30. Dörr, T., Vulić, M. & Lewis, K. Ciprofloxacin causes persister formation by inducing the TisB toxin in *Escherichia coli*. *PLoS Biol.* **8**, e1000317 (2010).
31. Blázquez, J., Rodríguez-Beltrán, J. & Matic, I. Antibiotic-induced genetic variation: how it arises and how it can be prevented. *Annu. Rev. Microbiol.* **72**, 209–230 (2018).
32. Umezawa, N. et al. Participation of reactive oxygen species in phototoxicity induced by quinolone antibacterial agents. *Arch. Biochem. Biophys.* **342**, 275–281 (1997).
33. Lopatkin, A. J. et al. Clinically relevant mutations in core metabolic genes confer antibiotic resistance. *Science* **371**, eaba0862 (2021).
34. Yim, H. H. & Villarejo, M. osmY, a new hyperosmotically inducible gene, encodes a periplasmic protein in *Escherichia coli*. *J. Bacteriol.* **174**, 3637–3644 (1992).
35. Zhang, S. et al. Degp degrades a wide range of substrate proteins in *Escherichia coli* under stress conditions. *Biochem. J.* **476**, 3549–3564 (2019).
36. Nagano, T. et al. Elongation factor G is a critical target during oxidative damage to the translation system of *Escherichia coli*. *J. Biol. Chem.* **287**, 28697–28704 (2012).
37. Farrell, C. M., Grossman, A. D. & Sauer, R. T. Cytoplasmic degradation of ssrA-tagged proteins. *Mol. Microbiol.* **57**, 1750–1761 (2005).
38. McGrew, D. A. & Knight, K. L. Molecular design and functional organization of the RecA protein. *Crit. Rev. Biochem. Mol. Biol.* **38**, 385–432 (2003).
39. Cole, S. T. Characterisation of the promoter for the LexA regulated sulA gene of *Escherichia coli*. *Mol. Gen. Genet.* **189**, 400–404 (1983).
40. Miller, C. et al. SOS response induction by beta-lactams and bacterial defense against antibiotic lethality. *Science* **305**, 1629–1631 (2004).
41. Chapman, A. R. et al. Single cell transcriptome amplification with MALBAC. *PLoS ONE* **10**, e0120889 (2015).
42. Sheng, K., Cao, W., Niu, Y., Deng, Q. & Zong, C. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat. Methods* **14**, 267–270 (2017).
43. McNulty, R. et al. Probe-based bacterial single-cell RNA sequencing predicts toxin regulation. *Nat. Microbiol.* **8**, 934–945 (2023).
44. Wang, B. et al. Massively-parallel Microbial mRNA Sequencing (M3-Seq) reveals heterogeneous behaviors in bacteria at single-cell resolution. Preprint at *bioRxiv* <https://doi.org/10.1101/2022.09.21.508688> (2022).
45. Ma, P. et al. Bacterial droplet-based single-cell RNA-seq reveals antibiotic-associated heterogeneous cellular states. *Cell* **186**, 877–891.e814 (2023).
46. Datlinger, P. et al. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat. Methods* **18**, 635–642 (2021).
47. Homberger, C., Hayward, R. J., Barquist, L. & Vogel, J. Improved bacterial single-cell RNA-Seq through automated MATQ-Seq and

- Cas9-based removal of rRNA reads. *mBio* **14**, e03557–03522 (2023).
48. Huang, Y., Sheth, R. U., Kaufman, A. & Wang, H. H. Scalable and cost-effective ribonuclease-based rRNA depletion for transcriptomics. *Nucleic Acids Res.* **48**, e20 (2020).
  49. Mathur, L. et al. Combi-seq for multiplexed transcriptome-based profiling of drug combinations using deterministic barcoding in single-cell droplets. *Nat. Commun.* **13**, 4450 (2022).
  50. Lee, S. K. & Keasling, J. D. A propionate-inducible expression system for enteric bacteria. *Appl. Environ. Microbiol.* **71**, 6856–6862 (2005).
  51. Ko, J., Wang, Y., Sheng, K., Weitz, D. A. & Weissleder, R. Sequencing-based protein analysis of single extracellular vesicles. *ACS Nano* **15**, 5631–5638 (2021).
  52. Wang, Y. et al. Dissolvable polyacrylamide beads for high-throughput droplet DNA barcoding. *Adv. Sci. (Weinh.)* **7**, 1903463 (2020).
  53. Chen, B. et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).
  54. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 3 (2011).
  55. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
  56. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2013).
  57. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499 (2017).
  58. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
  59. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).
  60. Wang, Y. Droplet-based high-throughput single microbe RNA sequencing by smRandom-seq. *wanglab2023/smRandom-seq*. <https://doi.org/10.1101/2021.03.10.434868> (2023).

## Acknowledgements

The project was supported by the National Natural Science Foundation of China (No. 32200073, Y.W., No. 82200977, Z.X.), Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (No. 2021R01012, Y.W.), and Liangzhu Laboratory. This work was supported by Harvard MRSEC Program under NSF award DMR 14–20570. This work was also supported by the Wyss Institute for Biologically Inspired Engineering at Harvard University. Thanks for the technical support by the core facilities and computing platform of Zhejiang University Medical Center and Liangzhu Laboratory. We thank the helpful comments from Peter Lu.

## Author contributions

Y.W. and D.W. conceived the study. Y.W., Z.X., K.S., Y.T.W., N.L., and Y.H. developed the chemistry for smRandom-seq. Y.W., Y.L., S.Z., and T.C. constructed the microfluidics platform. Y.T.W., R.R., X.H., M.S., Y.J., and Y.Y. contributed to the antibiotic-treated bacteria experiments. Z.X., T.Z., J.C, Y.S., Y.C., G.G., and Z.W. analyzed the data. Z.X. and Y.W. wrote the manuscript. Y.W., D.W., and P.Y. supervised this project. All authors have revised and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-023-40137-9>.

**Correspondence** and requests for materials should be addressed to Kuanwei Sheng, Peng Yin, David A. Weitz or Yongcheng Wang.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023