The ISME Journal (2017), 1–4

www.nature.com/ismej

SHORT COMMUNICATION

Single-cell transcriptomics of small microbial eukaryotes: limitations and potential

Zhenfeng Liu, Sarah K Hu, Victoria Campbell, Avery O Tatters, Karla B Heidelberg and David A Caron
Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA

Single-cell transcriptomics is an emerging research tool that has huge untapped potential in the study of microbial eukaryotes. Its application has been tested in microbial eukaryotes 50 μm or larger, and it generated transcriptomes similar to those obtained from culture-based RNA-seq. However, microbial eukaryotes have a wide range of sizes and can be as small as 1 μm. Single-cell RNA-seq was tested in two smaller protists (8 and 15 μm). Transcript recovery rate was much lower and randomness in observed gene expression levels was much higher in single-cell transcriptomes than those derived from bulk cultures of cells. We found that the reason of such observation is that the smaller organisms had much lower mRNA copy numbers. We discuss the application of single-cell RNA-seq in studying smaller microbial eukaryotes in the context of these limitations.
The ISME Journal advance online publication, 6 January 2017; doi:10.1038/ismej.2016.190

Single-cell transcriptomics has emerged in recent years as a powerful tool in medical research to study cell-to-cell variability (Saliba *et al.*, 2014). This technology is very appealing in the study of the ecophysiology of microbial eukaryotes. Many organisms of interest are not in culture, so large numbers of cells are not available for transcriptomic analyses. Even for those in culture, it would be interesting to learn their gene expression *in situ*. Single-cell transcriptomics offers the ability to target organisms of interest from environmental samples, therefore not wasting sequencing capacity on non-target taxa. It also provides a means of obtaining genetic information of several co-occurring organisms in microbial communities without the need to bin sequences or align to reference genomes like in metatranscriptome studies. Kolisko *et al.* (2014) described the first successful test of single-cell RNA-seq for microbial eukaryotes. They reported that transcriptome coverage from single cells was comparable to those of culture-based transcriptomes for five different ciliates with sizes ranging from 50 to 500 μm. However, microbial eukaryotes have a wide range of sizes and can be as small as 1 μm (Caron *et al.*, 2009). The feasibility of single-cell RNA-seq in smaller microbial eukaryotes remains unknown. Here we describe results that transcript recovery rate using single-cell RNA-seq was significantly limited in two small microbial eukaryotic organisms. We estimated that these smaller organisms contained only thousands to tens of thousands of total mRNA molecules per cell. We discuss the application of single-cell RNA-seq in small microbial eukaryotes in the context of these limitations.

Single-cell and culture-based transcriptomes of two microbial eukaryotes, the dinoflagellate *Karlodium veneficum* (cell length of ~15 μm) and the haptophyte *Prymnesium parvum* (cell length of ~8 μm), were sequenced, assembled and compared. The assembled transcriptomes contained 63 184 and 38 704 transcripts for *K. veneficum* and *P. parvum*, respectively. Most of these transcripts were detected in the culture-based transcriptomes. In comparison, only ~15% of the transcripts were detected in the transcriptomes of single cells of *K. veneficum* on average, while the average transcript recovery rate was ~3% for smaller *P. parvum* single cells (Table 1). These rates were much lower than those documented for ciliate species of larger size (80–100%; Kolisko *et al.*, 2014). When single-cell data were combined, transcriptomes summed from 10 *K. veneficum* cells recovered two-thirds of the transcripts observed in the cultured-based metatranscriptome, while transcripts summed from 18 *P. parvum* cells recovered less than one-third (Figure 1). Lower gene recovery rate was also reported by Kolisko *et al.* (2014) for their smallest cell (*Tetrahymena thermophile*, ~50 μm) mainly because >90% of its reads were from one rRNA contig. No such bias was observed in this study as reads from all rRNA contigs combined, or the most represented contig never accounted for >12% of total reads in any single cell sample.

Correspondence: Z Liu, Department of Biological Sciences, University of Southern California, 3616 Trousdale Parkway, Los Angeles, CA 90089-0371, USA.
Email: zhenfenl@usc.edu
Received 4 May 2016; revised 7 November 2016; accepted 21 November 2016

Table 1 Summary of single cell and culture based transcriptomes of *K. veneficum* and *P. parvum*, and estimations of mRNA molecules per cell in the two species

Species (cell length)	*Transcriptome assembly (batch culture and single cells combined)*		*Single-cell transcriptomes. No. of transcripts detected*[a]		*Estimation of mRNA molecules per cell*[b]	
	No. of transcripts	*Size*	*Single cells (average)*	*All cells combined*	*Based on amounts of RNA extracted*[c]	*Based on RNA spike-in*
Karlodinium veneficum (~15μm)	63 184	50.9 Mbp	1532–19 001 (9334)	42 360	17 500–87 600	51 000
Prymnesium parvum (~8μm)	38 704	41.9 Mbp	394–2304 (1298)	10 672	3500–17 400	4880

[a]FPKM $\geqslant$ 1. [b]Assume RNA extraction efficiency is 50%. [c]Assume 1–5% of total RNA is mRNA, average transcript length is 1000 nt.

In addition to low transcript recovery rates, we also observed much larger variability among single-cell transcriptomes than typically observed in mammalian studies. Approximately half of the transcripts detected in single-cell transcriptomes were only detected in one cell. Very few transcripts (220 *K. venificum* and 18 *P. parvum* transcripts), usually those with highest expression levels in the culture-based transcriptomes, were detected in all single cells. Among transcripts detected in multiple cells, expression levels often varied markedly between different cells (Figure 1). Many known housekeeping genes such as those encoding ribosomal proteins were not detected in many cells. Despite the large cell-to-cell variability on the gene level, the collective expression levels of major pathways and functions were very similar across different cells, except in one cell (cell #9) with extremely low transcript recovery rate (Supplementary Figure 1). These results suggested that the observed differences in single-cell transcriptomes were unlikely the reflection of physiological differences among cells, but rather of elevated stochasticity on the level of individual genes.

Both low transcript recovery rate and high gene-level variability could result from relatively low RNA content per cell. Single-cell RNA-seq has been applied successfully in human cells, which are estimated to contain 50 000–300 000 mRNA molecules per cell (Marinov *et al*., 2014). On the other hand, it is considered not suitable for bacteria (Taniguchi *et al*., 2010), which have only 200–2000 mRNA molecules per cell (Moran *et al*., 2013). Numbers of mRNA molecules per cell in *K. veneficum* and *P. parvum* were estimated using two methods, based on either total RNA extraction amounts or RNA spike-in standards. Results from both methods were similar. *K. veneficum* and *P. parvum* contained ~51 000 and ~4 800 mRNA molecules per cell on average, respectively (Table 1). These mRNA copy numbers limited the inventory of transcripts that these two organisms could possibly carry at any particular time. Our *in silico* simulations showed that mRNA copy numbers fewer than 100 000 could significantly limit transcript recovery rate in these two organisms (Supplementary Figure 2A). In microbial eukaryotes of similar sizes, which probably have similar mRNA copy numbers, low transcript recovery rate per cell can be expected, unless they have very small genomes.

Gene transcription generally occurs in stochastic bursts (Golding *et al*., 2005; Suter *et al*., 2011), and single-cell transcriptomes of cells with relatively few mRNA molecules are much more susceptible to biological and technical stochasticity (Marinov *et al*., 2014). Because of the small mRNA copy numbers in the two species examined in this study, it was doubtful that the single-cell gene expression levels and transcript presence/absence in different cells were reliable. Average expression levels of transcripts in single cells had no correlation with those in cultures in both organisms, except for transcripts with extremely high expression levels (Supplementary Figure 2C and D). In human cells, 30–100 single cells are needed to reliably measure gene expression levels (Marinov *et al*., 2014). In small microbial eukaryotes, many more cells would be needed to achieve the same goal. Caution is warranted when interpreting single-cell transcriptome comparisons of different samples, especially if the cells are small.

Our data illustrated a simple but important concept: when using single-cell transcriptomics with microbes, size matters. Less efficient gene discovery and higher stochasticity in gene expression levels should be expected when designing experiments using single-cell transcriptomics on smaller microbial eukaryotes. However, such limitations should in no way discourage the application of the technology in studying these organisms. A simple solution exists: combining multiple single-cell transcriptomes of the same organism. Our simulations showed that, in cells with mRNA copy numbers similar to *K. veneficum*, 25 cells combined should recover most transcripts. In smaller protists such as *P. parvum*, more than 100 cells are likely needed (Supplementary Figure 2B). With this in mind, we tested microfluidic single-cell RNA-seq of *P. parvum*

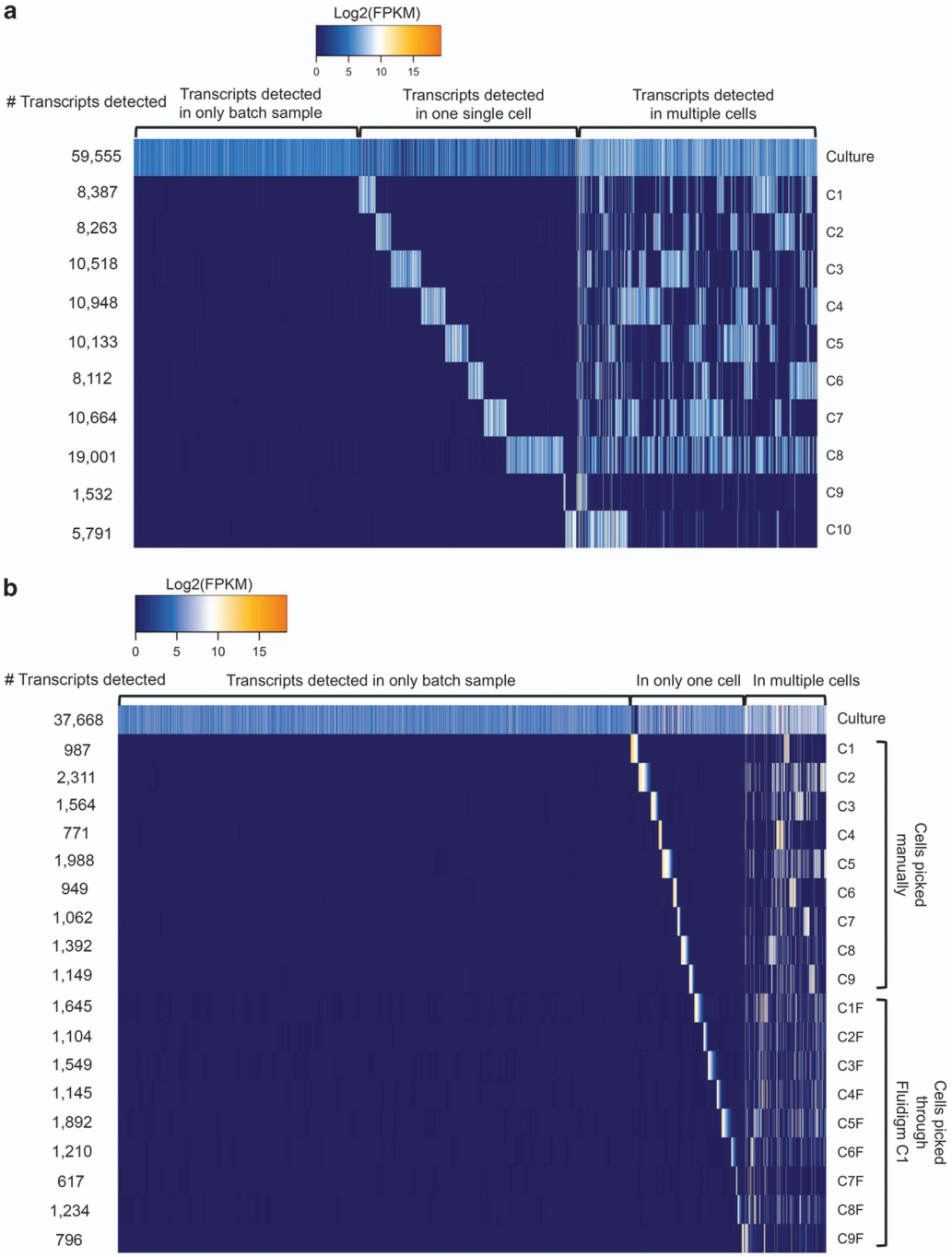

Figure 1 Heatmap of expression levels (in the form of Log2 of FPKM values) of *K. veneficum* (**a**) and *P. parvum* (**b**) transcripts in culture and single cells. Transcripts were grouped by presence/absence in single cells. Transcripts detected in multiple cells were arranged by hierarchical clustering of expression patterns among all samples.

because of its demonstrated ability to capture a large number of single cells quickly (Wu *et al*., 2014). However, our test was less successful than anticipated for this species (we obtained nine single-cell transcriptomes out of 96 wells), presumably because *P. parvum* was smaller than smallest designed cell size (10 μm) of any chip available at the time. The transcriptomes obtained were similar to those obtained from manually isolated cells (Figure 1b). We believe that with some optimization, high-throughput single-cell transcriptomics of microbial eukaryotes should be achievable in the near future.

Single-cell transcriptomics has already been used to advance our knowledge of microbial eukaryotes (for example, Balzano *et al*., 2015 and Gravelis *et al*., 2015). Undoubtedly, it will continue to shine as a powerful tool in studying microbial eukaryotes in nature, large and small, especially when gene discovery is still one of the main goals in the field (Keeling *et al*., 2014).

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work was supported by the Gordon and Betty Moore Foundation through Grant GBMF3299 to DAC and KBH.

References

Balzano S, Corre E, Decelle J, Sierra R, Wincker P, Da Silva C *et al.* (2015). Transcriptome analysis to investigate symbiotic relationships between marine protists. *Front Microbiol* **6**: 98.

Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB. (2009). Protists are microbes too: a perspective. *ISME J* **3**: 4–12.

Golding I, Paulsson J, Zawilski SM, Cox EC. (2005). Real-time kinetics of gene activity in individual bacteria. *Cell* **123**: 1025–1036.

Gravelis GS, White RA, Suttle CA, Keeling PJ, Leander BS. (2015). Single-cell transcriptomics using spliced leader PCR: evidence for multiple losses of photosynthesis in polykrikoid dinoflagellates. *BMC Genomics* **16**: 528.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA *et al.* (2014). The marine microbial eukaryote transcriptome sequencing project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* **12**: e1001889.

Kolisko M, Boscaro V, Burki F, Lynn DH, Keeling PJ. (2014). Single-cell transcriptomics for microbial eukaryotes. *Curr Biol* **24**: R1081–R1082.

Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM *et al.* (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res* **24**: 496–510.

Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan LK *et al.* (2013). Sizing up metatranscriptomics. *ISME J* **7**: 237–243.

Saliba AE, Westermann AJ, Gorski SA, Vogel J. (2014). Single-cell RNA-seq: advances and future challenges. *Nucleic Acids Res* **42**: 8845–8860.

Suter DM, Molina N, Gatfield D, Schneider K, Schibler U, Naef F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**: 472–474.

Taniguchi Y, Choi PJ, Li GW, Chen HY, Babu M, Hearn J *et al.* (2010). Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**: 533–538.

Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME *et al.* (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nat Methods* **11**: 41–46.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)

Methods and Materials

Cell culture

Prymnesium parvum strain UOBS-LP0109 (Texoma1) was isolated from Lake Texoma, Oklahoma, USA. *Karlodinium veneficum* strain K2 was isolated from Coronado Island, California, USA. *P. parvum* was grown in low phosphate (P/100) L1 media minus silica at 18 ppt salinity and *K. veneficum* was grown in L1/2 media at 34 ppt salinity. All cultures were grown in 12:12 hours light:dark conditions at 15°C. Light intensity for *P. parvum* was $75\mu Em^{-2}s^{-1}$ and that for *K. veneficum* was $65\mu Em^{-2}s^{-1}$. Cultures were sampled daily to monitor growth by counting cells using a Palmer-Maloney chamber after fixing 1 mL of culture with 1% formalin. Samples for transcriptomes were taken during late exponential phase. Cell density were ~41,000 and ~220,000 cells/mL for *K. veneficum* and *P. parvum*, respectively, at the time of sampling.

Single cell isolation, RNA extraction and cDNA synthesis

10 single cells of each species were hand picked using a micropipette and gently rinsed twice in filtered seawater and once in culture-grade PBS (Sigma-Aldrich Life Sciences, St. Louis, MO, USA #D1408). Total RNA from single cells was immediately extracted and cDNA was synthesized using the SMART-Seq v3 Ultra Low Input RNA Kit (Clontech Laboratories, Inc. Mountain View, CA, USA, # 634850). cDNA was amplified with 25 cycles of PCR. Both RNA and final cDNA were quality screened using the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA) and Qubit Fluorometer (ThermoFisher Scientific, Waltham, MA, USA). cDNA of one *P. parvum* cell did not pass quality control.

10mL of the same *K. veneficum* and *P. parvum* cultures were taken for batch culture based transcriptome experiments. Cells were collected by centrifugation at 4,000 rpm for 10 minutes at 4°C. Total RNA was immediately extracted using the RNeasy kit (Qiagen, Valencia, CA, USA, #74904) for *K. veneficum* or Direct-zol RNA MiniPrep with TRI reagent (Zymo, Irvine, CA, USA, #R2050S) for *P. parvum*. Total extracted RNA was quality checked using the Agilent 2100 Bioanalyzer and quantified using Qubit Fluorometer. 1 µL of 1:10,000 diluted ERCC RNA spike-in (ThermoFisher Scientific, #4456740) was added to 5 ng of total RNA of each species. cDNA was synthesized using the same SMART-Seq v3 Ultra Low Input RNA Kit but with 10 cycles of PCR.

Microfluidic single cell isolation, RNA extraction, and cDNA synthesis

A second culture of *P. parvum* under the same growth conditions (except for light intensity at 117μE m^{-2} s^{-1}) was used to conduct single-cell transcriptome using the C1 Single-Cell Auto Prep system (Fluidigm, South San Francisco, CA, USA) following the manufacturer protocol (PN 100-7168). After the single cell capture step, the chip was removed from the instrument and examined immediately by microscopy. Single, live cells were observed in 24 of 96 wells. Cell lysis, RNA extraction and cDNA sysnthesis were immediately carried out according to Fluidigm protocol (PN 100-7168). cDNA was quantified using Qubit Fluorometer before library preparation. Only 9 of 24 wells had sufficient cDNA to proceed.

Library preparation and DNA sequencing

Sequence libraries were created with the Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA, USA) using 150 pg of cDNA from manually picked single cells and batch cultures and 1 ng cDNA from Fluidigm single-cell isolations. Final libraries were quality checked using the Agilent 2100 Bioanalyzer before sequencing on a Illumina HiSeq 2500 to obtain 100bp paired-end reads for manually picked single cells and batch cultures and 50bp paired-end reads for *P. parvum* cells captured using the Fluidigm C1. All sequencing was done at University of Southern California UPC Genome and Cytometry Core. Single cell and culture-based transcriptomes were sequenced at the same read depth. On average, about 4 million and 6 million read pairs were generated for *P. parvum* and *K. veneficum* cells and cultures, respectively.

Original sequences are available through the NCBI sequence read archive (SRA) under the accession numbers SRX1430089 for the *P. parvum* batch culture sample, SRX1430091 for the *P. parvum* single cells picked manually, SRX1431799 for *P. parvum* single cells captured using Fluidigm C1, SRX1434822 for the *K. veneficum* batch culture sample, and SRX1434823 for *K. veneficum* single cells picked manually.

Bioinformatic analyses

Adapter sequences of SMART-Seq and Nextera XT kits were trimmed from sequences using Trimmomatic v. 0.32 (Bolger *et al.*, 2014). Sequences were also quality filtered using

Trimmomatic with the options "LEADING:5 TRAILING:5 SLIDINGWINDOW:5:15" for all sequences, plus "MINLEN:50" for 100bp sequences, and "MINLEN:35" for 50bp sequences. Sequences of the two batch culture samples were then aligned to ERCC spike-in sequences using bowtie v. 0.12.7 (Langmead *et al.*, 2009). A custom PERL script was used to count and remove sequences aligned to each spike-in sequence. Sequences from the batch culture sample and manually picked single cell samples of each organism were combined and assembled *de novo* using Trinity v. r20140717 (Grabherr *et al.*, 2011). Sequences of *P. parvum* single cells captured using the Fluidigm C1 were not used to generate the assembly because they were collected from a comparable but different culture.

Sequences from each sample were then aligned back to the assembled transcriptome to estimate transcript abundance in each sample using the script align_and_estimate_abundance.pl included in Trinity toolkit (Haas *et al.*, 2013) with alignment method bowtie2 v. 2.2.3 (Langmead and Salzberg, 2012) and estimation method RSEM v. 1.2.23 (Li and Dewey, 2011). Transcripts with less than 5 total aligned read pairs were removed and not further analyzed. FPKM values of transcripts across different samples of the same organism were then normalized using the script abundance_estimates_to_matrix.pl included in Trinity toolkit (Haas *et al.*, 2013). Normalized FPKM values were used in downstream analyses.

Coding sequences from assembled transcriptomes were predicted using TransDecoder (Haas *et al.*, 2013). KEGG annotation of predicted protein sequences were generated using KAAS annotation server (Moriya *et al.*, 2007).

Numbers of mRNA molecules per cell were estimated using two methods. In both methods, RNA extraction efficiency was assumed to be 50%. The first method was based on total RNA extracted from known numbers of cells. Total numbers of mRNA molecules were calculated assuming that 1-5% of total RNA was mRNA, and average mRNA length was 1000nt. In the second method, FPKM values of spike-in mRNA sequences that were detected in the transcriptome and their numbers of molecules were analyzed with linear regression. R^2 was 0.894 and 0.955 for *P. parvum* and *K. veneficum*, respectively. Numbers of mRNA molecules for all transcripts were then calculated from their FPKM values using the resulting linear predictive model and summed.

In silico simulation of single-cell transcriptomes were carried out as the following. For each transcript t, we assume it is expressed in a proportion p_t ($0 < p_t \leq 1$) of all cells. Its expression level in those cells is $FPKM_t/p_t$, where $FPKM_t$ is the normalized FPKM value of t observed in the batch culture sample. In other cells, t is not expressed at all. However, we have no reliable estimate of the distribution of p_t. We used a strategy as described in (Marinov *et al.*, 2014) to produce a reasonable set of p_t. In short, all transcripts were divided into ten percentile groups in order of their FPKM values. Each of the ten percentile groups from lowest to highest FPKM values was assigned a base probability from 0.1, 0.2, all the way up to 1. p_t of transcripts of each percentile group were randomly generated to follow a normal distribution with a mean equal to the base probability and with the floor of p_t set at 0.01.

For each cell, a random number r_t ($0 < r_t \leq 1$) was generated for each transcript t. t is considered expressed in this cell if $r_t \geq p_t$. For a cell with n mRNA molecules, n rounds of random sampling with replacement of the expressed gene pool, T, with the probability of each gene being sampled equal to $\frac{FPKM_t / p_t}{\sum_{t \in T} FPKM_t / p_t}$ was carried out. Single molecule capture efficiency was assumed to be 0.5. In other words, each mRNA molecule has a 50% chance to be reverse transcribed, amplified, and appear in the final library. Numbers of transcripts with at least one molecule in the final library was tallied for each cell.

References

Bolger AM, Lohse M, Usadel B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* **30**: 2114-2120.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I *et al*. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol* **29**: 644-652.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, *et al.* (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494-1512.

Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357-359.

Langmead B, Trapnell C, Pop M, Salzberg SL. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

Li B, Dewey CN. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**: 323.

Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM *et al*. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Res* **24**: 496-510.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* **35**: W182-W185.

Figure S1. Heatmap of expression levels by KEGG pathways in *K. veneficum* from batch culture and single cells. In each sample, FPKM values of transcripts belonging to the same KEGG pathways were summed, log2-transformed, and plotted.

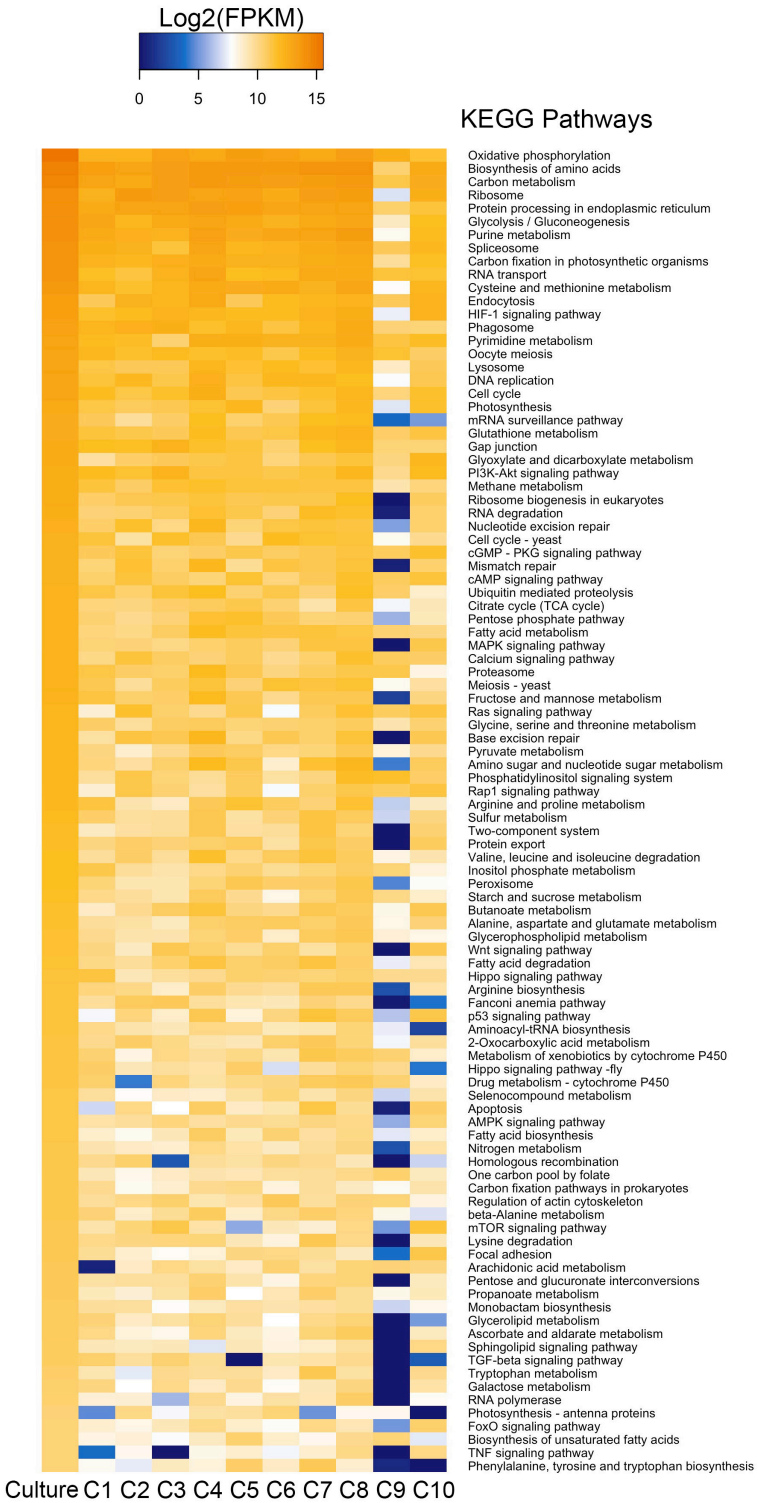

Figure S2. *In silico* simulation showing transcript recovery rate in *K. veneficum* and *P. parvum* with different hypothetical mRNA copy numbers (A) and with multiple single cells combined (B). 50 simulations were carried out in both cases. Simulations were run based on FPKM values of transcripts in the culture-based sample (see Methods and Materials for detail). In (B), estimated mRNA copy numbers (Table 1) were used. The correlation of transcript expression levels in the culture-based samples and average single-cell expression levels of *K. veneficum* (C) and *P. parvum* (D) are shown. Each dot represents a transcript.