Correspondence

Single-cell transcriptomics for microbial eukaryotes

Martin Kolisko[1], Vittorio Boscaro[2], Fabien Burki[1], Denis H. Lynn[3,4], and Patrick J. Keeling[1,*]

One of the greatest hindrances to a comprehensive understanding of microbial genomics, cell biology, ecology, and evolution is that most microbial life is not in culture. Solutions to this problem have mainly focused on whole-community surveys like metagenomics, but these analyses inevitably loose information and present particular challenges for eukaryotes, which are relatively rare and possess large, gene-sparse genomes [1,2]. Single-cell analyses present an alternative solution that allows for specific species to be targeted, while retaining information on cellular identity, morphology, and partitioning of activities within microbial communities [2]. Single-cell transcriptomics, pioneered in medical research [3], offers particular potential advantages for uncultivated eukaryotes, but the efficiency and biases have not been tested. Here we describe a simple and reproducible method for single-cell transcriptomics using manually isolated cells from five model ciliate species; we examine impacts of amplification bias and contamination, and compare the efficacy of gene discovery to traditional culture-based transcriptomics. Gene discovery using single-cell transcriptomes was found to be comparable to mass-culture methods, suggesting single-cell transcriptomics is an efficient entry point into genomic data from the vast majority of eukaryotic biodiversity.

Ciliates are good models to evaluate single-cell transcriptomics for uncultivated microbial eukaryotes, because they allow many of the potential problems to be examined in a relatively well-controlled fashion. Several species with well-curated whole-genomes or deep-coverage transcriptomes are available to provide points of reference (e.g., [4–6]). Moreover, cell size varies by orders of magnitude between species [7]. Also, because ciliates are obligate heterotrophs [7], cultures are typically 'contaminated' with food bacteria or other eukaryotes, and even isolated cells can retain potentially misleading remnants of partially digested cells of foreign origin. We have selected five species representing a wide variety of cell sizes (from *Condylostoma* at 500 µm to *Tetrahymena* at 50 µm) and derived from different environments. To evaluate contamination, we include species in axenic culture (*Tetrahymena*), with *endosymbiotic* bacteria (*Euplotes harbours Protistobacter heckmanni*), feeding on mixed prey from a natural environment (*Blepharisma*), or feeding on defined eukaryotes (*Condylostoma* feeding on the diatom *Phaeodactylum tricornutum*, and *Euplotes* and *Paramecium* feeding on the green alga *Dunaliella tertiolecta*). From each species, individual cells were manually isolated and washed, and single-cell cDNA libraries were constructed and sequenced (detailed methods are available in Supplemental Information, published with this article online).

We assessed three important characteristics of high-throughput sequence datasets: bias introduced by library construction and sequencing; contamination levels; and the effectiveness of gene discovery. Assembly and annotation resulted in between 12,030 and 39,221 contigs (Figure 1A). Larger cells yielded more contigs, potentially due to differential bias for reads mapping to individual contigs (Figure 1B). For example, in *Condylostoma* (comparatively large cells) the most abundantly represented contig (a cystein protease) accounted for 8% of reads, whereas in *Tetrahymena* (comparatively small cells) 90% of the reads mapped to the LSU rRNA. In contrast, contamination levels were relatively even and uniformly low (2.1%–4.27%; Figure 1A). Putative contaminants were most often related to bacteria, but generally not to a single type. In *Euplotes*, contamination from its *endosymbiotic* bacteria was also low — only 0.4% of sequences were identified as being from *Burkholderiacaea*. Interestingly, no sequences from known eukaryotic prey were found, despite the fact that they should be unaffected by polyA selection. It is possible that prey RNA (perhaps unlike DNA) is quickly cleared from feeding cells, so single-cell transcriptomes may offer a manageable solution to contamination in complex natural communities.

To evaluate the success of gene discovery, we first examined the recovered proportion of two defined collections of housekeeping genes: 20 aa-tRNA synthetases, and 248 core eukaryotic proteins [8]. Recovery rates varied between 90 and 100% for aa-tRNA synthetases and 66 and 94% for the 248 core-gene set (Figure 1A). Because some of the 248-gene set may not be present in ciliates, this may be a slight underestimate. Second, we compared the single-cell transcriptomes to equivalent data from mass-culture, which was done in three different ways depending on the best available comparators. For *Condylostoma*, a direct comparison with a transcriptome from the same strain [6,9] revealed the single-cell transcriptome actually recovered more unique contigs than the mass-culture transcriptome. Both data sets included more than 3,000 unique contigs, but shared ~19,000 contigs in common, suggesting they comparably reflect the expression status of the cell. No transcriptome is available from the same strain/species of *Blepharisma*, *Euplotes*, and *Paramecium*, so their transcriptomes were compared with those of closely related congeners against the most similar available genome or complete transcriptome. Here, the single-cell transcriptome yielded about 90% of the genes recovered by culture-based methods and, except for *Paramecium*, both datasets contained similar sets of orthologues when compared to the reference. For *Tetrahymena*, a genome of the same species is available [4], so single-cell and culture-based transcriptomes were mapped directly to its full gene set. The single-cell transcriptome was less efficient due to the rRNA bias described above, but most reads and contigs nevertheless mapped to the genome (Figure S1), recovering ~11,000 genes, or 33% of the genome (compared with 77% from a comparable number of reads from a culture-based transcriptome).

Single-cell RNAseq is a powerful method to generate large-scale datasets from uncultivated microbial eukaryotes. Comparing data from ciliates revealed some biases, but even in the excessively biased case of *Tetrahymena* (where 90% of

A

	Contamination levels		Estimation of transcriptome completness		Comparison with culture-based transcriptome
Species (cell lenght)	**# Total contigs**	**# Contaminant contigs**	**tRNA synthetase**	**248 gene set**	**Single cell vs. Culture**
Condylostoma magnum (500μm)	**39221**	**1646** (4.2%)	**20/20**	**232/248**	**102%** (19033/3817/3257)
***Blepharisma* sp.** (150μm)	**35221**	**743** (2.12%)	**20/20**	**232/248**	**96%** (3467/774/921)
Euplotes woodruffi (120μm)	**23491**	**1013** (4.31%)	**18/20**	**197/248**	**80%** (2510/986/1873)
Paramecium duboscqui (100μm)	**18005**	**538** (2.99%)	**20/20**	**216/248**	**90%** (3835/3539/4366)
Tetrahymena thermophile (50μm)	**12030**	**351** (2.92%)	**19/20**	**164/248**	Single cell: 33% Culture based: 77%

B

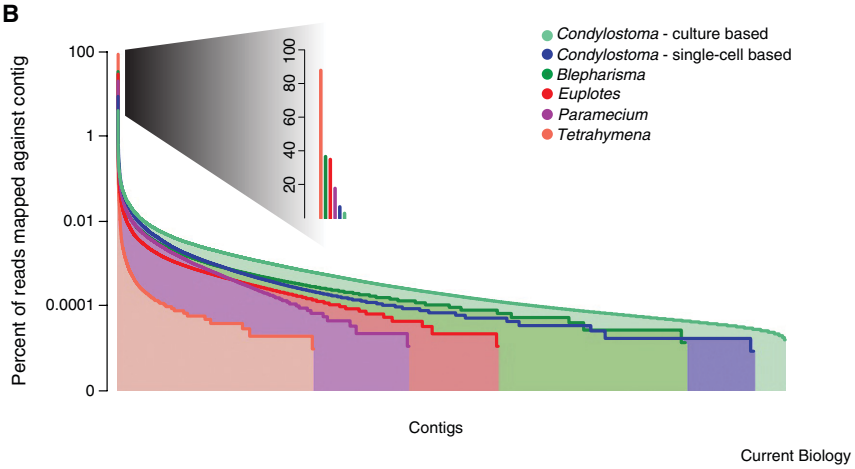

Figure 1. Comparison of single-cell and mass-culture transcriptomes from five species.
(A) Summary of the general characteristics of the single-cell transcriptome data sets, including levels of identifiable contamination (left columns, see also Figure S1), estimation of completeness by comparison with two sets of generally universally present housekeeping genes (middle columns), and a direct comparison of the efficiency of gene discovery with culture-based transcriptome data (right column; the three numbers in brackets represent from left to right: the transcripts shared, those unique to the single-cell data, and those unique to the culture data). (B) Summary of bias (over-representation) of five single-cell transcriptomes (colour coded to the right). The graph shows a log-scale bar chart with the percentage of reads mapping to each contig from each species. Along the X-axis are bars that each represent a contig (colour coded depending on the species). Because there are >10,000 contigs per species, they are packed closely together and are not each visible as discrete bars (except in the blow-up of the top end). The height of each bar (the Y-axis) is a log-scale percentage of reads that map against that particular contig. Contigs are sorted so that moving from left to right corresponds to the largest number of the reads mapped to lowest number of the reads mapped. The blow-up expands the upper portion showing the most over-represented contigs from each species, which vary from as low as 8% in *Condylostoma* to as high as 90% in *Tetrahymena*.

the sequence was uninformative), about one-third of the known genes in the genome were still identified. The majority of the single-cell transcriptomes were comparable to those from mass-culture. Because the total number of genes expressed in one cell at one time must be lower than those expressed collectively in cells in mass-culture, these results suggest the method is a very efficient way to recover transcripts in isolated cells. The parsimonious nature of this approach is also noteworthy: each transcriptome required resources comparable to cloning and sequencing 4–5 protein-coding genes, but instead generated tens of thousands of genes. Single-cell transcriptomes are readily applicable to a wide range of questions, the most obvious being the acquisition of data from species that are uncultivated or in complex culture (e.g., obligate predators), or that have uncultured life cycle stages (e.g., parasites). Enabling expression profiling and analysis of genome-wide data from these abundant but poorly studied systems will be key to advancing our understanding of microbial eukaryotes, their interactions with other microbial life, and the roles they play in natural environments.

Supplemental Information

Supplemental information contains experimental procedures and one figure, and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2014.10.026.

References

1. Keeling, P.J. (2013). Elephants in the room: protists and the importance of morphology and behaviour. Environ. Microbiol. Rep. *5*, 5–6.
2. del Campo, J., Sieracki, M.E., Molestina, R., Keeling, P.J., Massana, R., and Ruiz-Trillo, I. (2014). The others: our biased perspective of eukaryotic genomes. Trends. Ecol. Evol. *29*, 252–259.
3. Saliba, A.E., Westermann, A.J., Gorski, S.A., and Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. *42*, 8845–8860.
4. Eisen, J.A., Coyne, R.S., Wu, M., Wu, D., Thiagarajan, M., Wortman, J.R., Badger, J.H., Ren, Q., Amedeo, P., Jones, K.M., *et al.* (2006). Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. PLoS Biol. *4*, e286.
5. Aury, J.M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., *et al*. (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature *444*, 171–178.
6. Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C., *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.
7. Lynn, D.H. (2008). The Ciliated Protozoa. Characterization, Classification, and Guide to the Literature, Third Edition (Springer).
8. Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinfo. *23*, 1061–1067.
9. Gentekaki, E., Kolisko, M., Boscaro, V., Bright, K.J., Dini, F., Di Giuseppe, G., Gong, Y., Miceli, C., Modeo, L., Molestina, R.E., *et al*. (2014). Large-scale phylogenomic analysis reveals the phylogenetic position of the problematic taxon *Protocruzia* and unravels the deep phylogenetic affinities of the ciliate lineages. Mol. Phylogenet. Evol. *78*, 36–42.

[1]Canadian Institute for Advanced Research, Botany Department, University of British Columbia, 3529-6270 University Boulevard, Vancouver, BC, V6T 1Z4, Canada. [2]Biology Department, University of Pisa, via Alessandro Volta 4/6, Pisa, 56126, Italy. [3]Department of Integrative Biology, University of Guelph, Guelph, ON, N1G 2W1, Canada. [4]Department of Zoology, University of British Columbia, 6270 University Boulevard, Vancouver, BC, V6T 1Z4, Canada.
*E-mail: pkeeling@mail.ubc.ca