Comprehensive single-cell transcriptional profiling of a multicellular organism

Junyue Cao, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston, Cole Trapnell, Jay Shendure

To resolve cellular heterogeneity, we developed a combinatorial indexing strategy to profile the transcriptomes of single cells or nuclei, termed sci-RNA-seq (single-cell combinatorial indexing RNA sequencing). We applied sci-RNA-seq to profile nearly 50,000 cells from the nematode Caenorhabditis elegans at the L2 larval stage, which provided 50-fold “shotgun” cellular coverage of its somatic cell composition. From these data, we defined consensus expression profiles for 27 cell types and recovered rare neuronal cell types corresponding to as few as one or two cells in the L2 worm. We integrated these profiles with whole-animal chromatin immunoprecipitation sequencing data to deconvolve the cell type-specific effects of transcription factors. The data generated by sci-RNA-seq constitute a powerful resource for nematode biology and foreshadow similar atlases for other organisms.

I individual cells are the natural unit of form and function in biological systems. However, conventional methods for profiling the molecular content of biological samples mask cellular heterogeneity, which is likely present even in ostensibly homogeneous tissues (1). Recently, profiling the transcriptome of individual cells has emerged as a powerful strategy for resolving such heterogeneity. The expression levels of mRNA species are linked to cellular function and therefore can be used to classify cell types (2–10) and order cell states (11). Although methods for single-cell RNA sequencing (RNA-seq) have proliferated, they rely on the isolation of individual cells within physical compartments (2, 5, 8, 12–17). Consequently, preparing single-cell RNA-seq libraries with these methods can be expensive, the cost scaling linearly with the number of cells processed (13, 19).

We recently developed combinatorial indexing, a method using split-pool barcoding of nucleic acids to uniquely label a large number of single molecules or single cells. Single-molecule combinatorial indexing can be used for haplotype-resolved genome sequencing and de novo genome assembly (20, 21), whereas single-cell combinatorial indexing (“sci”) can be used to profile chromatin accessibility (sci-ATAC-seq) (22), genome sequence (sci-DNA-seq) (23), genome-wide chromosomal conformation (sci-Hi-C) (24), and DNA methylation (sci-MET) (25) in large numbers of single cells.

In this work, we developed a combinatorial indexing method to uniquely label the transcriptomes of large numbers of single cells or nuclei, termed sci-RNA-seq. We applied sci-RNA-seq to deeply profile single-cell transcriptomes in the nematode Caenorhabditis elegans at the L2 stage. C. elegans is the only multicellular organism for which all cells and cell types are defined, as is its entire developmental lineage (26, 27). However, despite its modest cell count (e.g., 782 somatic cells per L2 larva), our knowledge of the molecular state of each cell and cell type has remained fragmented. We therefore saw an opportunity to generate a powerful resource for nematode biologists, as well as for the single-cell genomics community.

Overview of sci-RNA-seq

In its current form, sci-RNA-seq relies on the following steps (Fig. 1A): (i) Cells are fixed and permeabilized with methanol (alternatively, cells are lysed and nuclei are recovered), then distributed across 96- or 384-well plates. (ii) A first molecular index is introduced to the mRNA of cells within each well, with in situ reverse transcription (RT) incorporating a barcode-bearing, well-specific polythymidine primer containing unique molecular identifiers (UMIs). (iii) All cells are pooled and redistributed by fluorescence-activated cell sorting (FACS) to 96- or 384-well plates in limiting numbers (e.g., 10 to 100 per well). Cells are gated on the basis of DAPI (4′,6-diamidino-2-phenylindole) staining to discriminate single cells from doublets during sorting. (iv) Second-strand synthesis, transposition with transposon Tn5 transposase, lysis, and polymerase chain reaction (PCR) amplification are performed. The PCR primers target the barcoded polythymidine primer on one end and the Tn5 adaptor insertion on the other end, so that resulting PCR amplicons preferentially capture the 3′ ends of transcripts. These primers introduce a second barcode that is specific to each well of the PCR plate. (v) Amplicons are pooled and subjected to massively parallel sequencing, resulting in 3′-tag digital gene expression profiles, with each read associated with two barcodes corresponding to the first and second rounds of cellular indexing (Fig. 1B). In a variant of the method described below, we introduce a third round of cellular indexing during Tn5 transposition of double-stranded cDNA.

Most cells pass through a unique combination of wells, resulting in a unique combination of barcodes for each cell that tags its transcripts. The rate of two or more cells receiving the same combination of barcodes can be tuned by adjusting how many cells are distributed to the second set of wells (22). Increasing the number of barcodes used during each round of indexing boosts the number of cells that can be profiled while reducing the effective cost per cell (Fig. S1). Additional levels of indexing can potentially offer even greater complexity and lower costs. Multiple samples (e.g., from different cell populations, tissues, individuals, time points, perturbations, or replicates) can be concurrently processed in one experiment, using different subsets of wells for each sample during the first round of indexing.

Scalability of sci-RNA-seq

We tested 282 sci-RNA-seq conditions with mammalian cells, optimizing the protocol and reaction conditions. We demonstrate scalability with 384 × 384–well sci-RNA-seq. During the first round of indexing, half of 384 wells contained pure populations of either human [human embryonic kidney 293T (HEK293T) and/or HeLa S3] or mouse (NIH/3T3) cells, and the other half contained a mixed human and mouse cells (table S1). After barcoded RT, cells were pooled and then sorted to a new 384-well plate for the second round of barcoding and deep sequencing of pooled PCR amplicons. We recovered 15,997 single-cell transcriptomes and readily assigned cells as human or mouse (Fig. 1C).

Optimization of sci-RNA-seq and application to nuclei

We performed optimized 96 × 96–well sci-RNA-seq on five cell or nucleus populations, each present in distinct subsets of wells during the first round of barcoding (table S1): HEK293T cells...
(8 wells), HeLa S3 cells (8 wells), an intraspecies mixture of HEK293T and HeLa S3 cells (32 wells), and interspecies mixtures of HEK293T and NIH/3T3 cells (24 wells) or nuclei (24 wells). We deeply sequenced the resulting library (~250,000 reads per cell, ~210,000 reads per nucleus, ~88% duplication rate), profiling 744 single-cell and 175 single-nucleus transcriptomes.

Transcriptomes in the 24 wells containing an interspecies mixture of human and mouse cells overwhelmingly mapped to the genome of one species or the other (289 of 294 cells), with only five “collisions” (which likely represent coincidental passage through the same wells by two or more cells) (Fig. 1D). Excluding collisions, we observed an average of 24,454 UMIs (5604 genes) per human cell and 17,665 UMIs (4065 genes) per mouse cell, with 1.9 and 3.3% of reads per human and mouse cell, respectively, mapping to the incorrect species.

Transcriptomes originating in the 24 wells containing an interspecies mixture of human and mouse nuclei also overwhelmingly mapped to the genome of one species or the other (172 of 175 nuclei), with only three collisions (fig. S2A). Excluding collisions, we observed an average of 32,951 UMIs (5737 genes) per human nucleus and 20,123 UMIs (4107 genes) per mouse nucleus (fig. S2, B and C), with 2.2 and 1.9% of reads per human and mouse nucleus, respectively, mapping...
respectively; whiskers, 1.5 times the interquartile range; circles, outliers). (D), the red line is the linear regression, and the black line is
measurements in aggregated sci-RNA-seq profiles of fixed-fresh and fixed-frozen cells. In (C) and
2 months apart on independently grown and fixed cells. (Image 37x715 to 558x728)

Cao et al

optimizing the number of cells per RT reaction

efficiency per molecule. Consistent with this, op-
of mRNA in cells resulting in a reduced RT ef-
faction, we fixed a mixture of HEK293T and NIH/

regions, similar to results from previous studies

gated transcriptomes of nuclei and cells were

transposition of double-stranded cDNA (22). We

Three-level, 9942 for HEK293T and 8611 for

mixture of HEK293T and NIH/3T3 cells after meth-

ers. At ~20,000 reads per cell (51% duplication

rate), we recovered 119 human and 62 mouse cells

with five collisions (fig. S6A). The aggregated

transcriptomes of three-level and two-level sci-

were performed 2 months apart on independently
grown and fixed cells. (D) Correlation between gene expression
measurements in aggregated sci-RNA-seq profiles of fixed-fresh and fixed-frozen cells. In (C) and
(D), the red line is the linear regression, and the black line is $y = x$.

to the incorrect species. The greater UMI counts
in nuclei are potentially due to the higher amounts
of mRNA in cells resulting in a reduced RT ef-

cell, respectively, mapping to the incorrect spe-

genes) per mouse cell (Fig. 2B and fig. S5A), with 0.9 and 12% of reads per human and mouse cell, respectively, mapping to the incorrect species. Although this and the previous experiment were performed 2 months apart on independently grown and fixed cells, the aggregated transcriptomes were well correlated ($r = 0.98$ for HEK293T and 0.98 for NIH/3T3 cells; Fig. 2C and fig. S5B).

We stored a portion of the methanol-fixed mixture of HEK293T and NIH/3T3 cells at −80°C for 4 days and repeated sci-RNA-seq (table S1). At ~200,000 reads per cell (73% duplication rate), we observed an average of 30,024 UMIs (5965 genes) per human cell and 21,393 UMIs (4603 genes) per mouse cell, with comparable purity (fig. S5C). The aggregated transcriptomes of the fixed-fresh and fixed-frozen cells were well correlated ($r = 0.99$ for HEK293T and 0.98 for NIH/3T3 cells; Fig. 2D and fig. S5D).

sci-RNA-seq with three levels of indexing

Two-level combinatorial indexing enables rou-
tine profiling of ~10^5 single cells per experiment. We tested an additional level of indexing during Tn5 transposition of double-stranded cDNA (22). We performed 16 × 6 × 16–well sci-RNA-seq on mixed HEK293T and NIH/3T3 cells after meth-
amol fixation. After RT with 16 barcodes and second-strand synthesis, cells were pooled and
distributed to six wells for fragmentation with indexed Tn5 (six barcodes), then pooled again
and sorted to 16 wells for PCR with indexed prim-
ers. At ~20,000 reads per cell (51% duplication
rate), we recovered 119 human and 62 mouse cells
with five collisions (fig. S6A). The aggregated
transcriptomes of three-level and two-level sci-
RNA-seq were well correlated ($r = 0.96$ for

HEK293T and 0.94 for NIH/3T3 cells; fig. S6, B and
C). Down-sampling to 15,000 reads per cell,
three-level indexing recovered fewer UMIs per cell than two-level indexing (three-level, on aver-

age, 6033 for HEK293T and 3640 for NIH/3T3
cells; two-level, 9942 for HEK293T and 8611 for

NIH/3T3 cells; fig. S6, D to G), possibly because
of lower efficiency of indexed versus unindexed
Tn5. This limitation notwithstanding, three-level
combinatorial indexing has the potential to en-
able routine profiling of >10^4 single cells per ex-
periment [fig. S6H (28)].

Single-cell RNA profiling of C. elegans

We next applied sci-RNA-seq to C. elegans. The cells in C. elegans larvae are much smaller, are
more variably sized, and have lower mRNA content
than the mammalian cell lines on which we opti-
mized the protocol. We pooled ~150,000 larvae
synchronized at the L2 stage and dissociated them
to single-cell suspensions. We then performed
in situ RT across six 96-well plates (576 first-round
barcodes), each well containing ~1000 C. elegans

and HEK293T cells, along with ~1000 human (HEK293T) cells
as internal controls. After pooling all cells, we
sorted the mixture of C. elegans and HEK293T
cells into 10 new 96-well plates for PCR bar-
coding (960 second-round barcodes), gating on
dNA content to distinguish between C. elegans


These same experiments were well correlated (逻辑 variation from batch effects. However, the stage (cillations in gene expression within each larval S8, B to F).

Operational timing to the first experiment (fig. 3.1 and 0.2% of reads per cell (mean, 1121 UMIs and 431 genes per cell) (fig. S7A). Importantly, control wells containing both C. elegans and HEK293T cells demonstrated clear separation between species (fig. S7B), with 3.1 and 0.2% of reads per C. elegans and human cell mapping to the incorrect species, respectively.

Identifying cell types

Semi-supervised clustering analysis segregated the cells into 29 distinct groups, the largest containing 13,205 (31.4% of) and the smallest only 131 (0.3% of) cells (Fig. 3A). Somatic cell types totaled 37,754 cells. We identified genes that were expressed specifically in a single cluster, and, by comparing those genes to expression patterns reported in the literature, assigned the clusters to cell types (figs. S15 to S23). Twenty-six cell types were represented in the 29 clusters: Nineteen represented exactly one literature-defined cell type, seven contained multiple distinct cell types, two contained cells of a specific cell type but had abnormally low UMI counts, and one could not be readily assigned. Neurons, which were present in seven clusters in the global analysis, were independently reclustered, initially revealing 10 major neuronal subtypes.

Intestine cells were not represented in any cluster. Intestine cells make up 2.5% of the soma, but are polyploid in C. elegans larvae (30) and autofluorescent in the DAPI channel used to measure DNA content (31). We speculated that they may have been excluded by how we gated on DNA content. We therefore performed a second, 384 × 144–well C. elegans experiment, collecting all cells, including polyploid cells, on the basis of DAPI fluorescence (96 wells) or gating to enrich for polyploid cells (48 wells). Intestine cells were present (unlike in the previous experiment) and, in wells gated for polyploidy, enriched twofold. This experiment yielded 7025 cells (UMI counts per cell for protein-coding genes ≥ 200), of which 6335 were somatic and 611 were intestine cells (fig. S8A).

Gene expression patterns in hypodermal cells suggested that the worm cells from the second C. elegans experiment were more tightly synchronized, overlapping but not identical in developmental timing to the first experiment (fig. S8, B to F). C. elegans larvae have pervasive oscillations in gene expression within each larval stage (32), making it difficult to distinguish biological variation from batch effects. However, the aggregated transcriptomes of HEK293T cells from these same experiments were well correlated ($r = 0.97$) and not readily separated by t-SNE (fig. S9).

This suggests that the variation observed is primarily due to differences in the developmental timing or preparation of the C. elegans larvae and cells, rather than technical variation in the sci-RNA-seq protocol. Regardless of its source, to minimize confounding by this variation, we only included the intestine cells from the second
The global and neuron-specific clustering analyses from the first C. elegans experiment, supplemented with intestine cells from the second experiment, allowed us to construct aggregate expression profiles for 27 cell types (tables S2 to S4; a 28th cell type, dopaminergic neurons, was excluded because of small cell numbers). These profiles are available online through GEExplore (http://genome.sfu.ca/gexplore/gexplore_search_tissues.html; fig. S14). Comparing the observed proportions of each cell type with their known frequencies in L2 larvae showed that sci-RNA-seq captured many cell types at or near expected frequencies (15 of 28 types had abundances ≥50% and 27 of 28 had abundances ≥20% of expectation; fig. 3B).

Transcriptional programs can be readily distinguished within single-cell transcriptome data sets at shallow sequencing depths (33). Therefore, our molecular profile for individual cell types in L2 worms may still be incomplete. However, we observed that half of all C. elegans protein-coding genes were expressed in at least 100 cells in the full data set, and 66% of protein-coding genes were expressed in at least 20 cells. This compares favorably with the estimates of expressed genes at the L2 stage from whole-animal RNA-seq (69%) (34). The "whole-worm" expression profile derived by aggregating all sci-RNA-seq reads correlated well with whole-animal bulk RNA-seq (34) for L2 C. elegans (Spearman correlation coefficient = 0.796 with cells from the first experiment only and 0.824 including intestine cells from the second experiment; Fig. 3C).

Furthermore, 3925 genes were significantly enriched in a single tissue (Fig. 3D and table S6), and 1939 genes were enriched for expression in a single cell type (Fig. 3E and table S7). Thus, despite the fact that sci-RNA-seq captures a minority of transcripts in each cell, our "oversampling" of the cellular composition of the organism enabled us to construct representative expression profiles for individual cell types (Fig. 3F).

Neuronal cell types

Because the transcripts of tissue or cell type clusters suggested subdivisions within groups (Fig. 3A), we examined expression in several tissues.

**Fig. 4. sci-RNA-seq reveals the transcriptomes of fine-grained anatomical classes of C. elegans neurons.** (A) t-SNE visualization of high-level neuronal subtypes. Cells identified as neurons from the t-SNE clustering shown in Fig. 3A were reclustered with t-SNE. NA, not assigned. (B) Clusters in the neuron t-SNE that can be identified as corresponding to one, two, or four specific neurons in an individual C. elegans larva. The number of neurons of each type is shown in parentheses. (C) Heat map showing the relative expression of high-neuronal-expression genes across 40 neuron clusters identified by t-SNE and density peak clustering. Genes are included if their expression in the aggregate transcriptome of all neurons in our data is more than five times that of their expression in any other tissue, excluding cases where the differential expression is not significant (q > 0.05). (D) Distribution for each neuron cluster of the number of genes in that cluster whose expression is more than five times that in the second-highest expressing neuron cluster (ψ for differential expression < 0.05). (E) Cartoon illustrating the position of the left and right ASE neurons (pink) relative to the pharynx (green). [From www.wormatlas.org (56)] (F) Volcano plot showing differentially expressed genes between the left and right ASE neurons. Points in red correspond to genes that are differentially expressed (ψ < 0.05) with more than a threefold difference between the higher- and lower-expressing neuron(s). (G) The left AWA and ASG neurons arise from the embryonic cell AB praapapa; the right AWA and ASG neurons arise from AB praapapa. (H) Volcano plot showing differentially expressed genes between the AWA and ASG neurons.
in more detail. We confirmed and extended findings that anterior and posterior body wall muscles have distinct expression patterns (fig. S10, A and B, and table S9) (35) and observed distinct expression patterns for posterior versus other intestine cells (fig. S10, C and D, and table S10) and amphid versus phasmid sheath cells (fig. S10, E and F, and table S11). But gene expression patterns were particularly diverse in neuronal cell types.

By morphological criteria, the 302 neurons of the worm are classified into 118 distinct types (36), and from the database of reporter transgene expression patterns, most of these are postulated to have unique molecular signatures (37). Our initial reorganization of neuronal cells divided them into 10 broad classes (Fig. 4A). Most classes of neurons were represented by several small but highly distinct clusters in the t-SNE plot. Further analysis of cluster-specific gene expression showed that many clusters corresponded to highly specific subsets of neurons in the L2 worm (Fig. 4B and table S7). Three clusters corresponded to sets of four neurons in an individual worm, eight clusters corresponded to a single pair of neurons (the ASE (ASEL), right ASE (ASER), and DVA neurons). Hierarchical clustering analysis showed that most of the 917 genes that were highly expressed in neurons, relative to other tissues, were expressed in only a minority of neuronal clusters (Fig. 4C). Of these 917 genes, 73% had no more than 10 neuron clusters as the ASEL and ASER gustatory neurons, follow by microarray analysis (38). The divergent transcriptomes of the AWA and ASG neurons, along with those of the ASE and ASER neurons, highlight the potential of cells that are extremely closely related in morphology and developmental lineage to feature distinct programs of gene regulation.

Integration with transcription factor binding sites

We hypothesized that correlating transcription factor (TF) binding patterns—profiling in chromatin immunoprecipitation (ChIP)-seq experiments by the modENCODE (41) and modERN (42) consortia—with gene expression profiles by cell type could give insights into the regulatory programs underlying the gene expression profiles. For each of 27 cell types, we constructed regularized regression models to predict each gene’s expression as a function of the TF ChIP peaks present in its promoter (Fig. 5). We restricted a cell type’s model to those TFs that
The expression identified several previously unknown regulators of cell type-specific expression. For example, *Rhl-8*, which is expressed specifically in ciliated sensory neurons (our data and reporter construct from [51]), was predictive of their gene expression program (fig. S12). The uncharacterized TF F49E8.2 is expressed specifically in the germline and associated with germline gene expression (fig. S12). The gene encoding F49E8.2 is an ortholog of the human gene **E2F-associated phosphophotrotein in comining indexing** (EAPP) ([52]), and F49E8.2 ChIP-seq peaks colocality with germ cell-specific EFL-I peaks [ortholog of E2F; data from [53]] more than could often be expected as a result of chance (χ² test, *P* = 2.8 × 10⁻²¹; fig. S13, A and B), suggesting that these proteins may physically interact. The hypothesism-associated TF-encoding genes *blmp-1* and *nhr-25* were also associated with gene expression in socket cells, excretory cells, and rectal cells. *nhr-25* is expressed 4.5 times as much in sex myoblasts ([54] versus 1.24 code-UTM) act. The hypodermis-associated TF-encoding genes *blmp-1* and *nhr-25* in the hydrolase associated with germ line-specific EFL-1 peaks [ortholog of E2F; data from [53]] more than could often be expected as a result of chance ([55]; table S29), indicating that multiple aspects of cellular biology can be concurrently barcoded, combinatorial indexing may also facilitate the scalable generation of **single-cell transeptome** (g.e., RNA-seq and ATAC-seq from each of many single cells). We also envision that large-scale, integrated profiling of the molecular states and lineage histories of **single cells** in other organisms will begin to give shape to global views of their development biology.

**REFERENCES AND NOTES**

20. A. Adey et al., Genome Res. 24, 2041–2049 (2014).
Comprehensive single-cell transcriptional profiling of a multicellular organism

Junyue Cao, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston, Cole Trapnell and Jay Shendure

Science 357 (6352), 661-667.
DOI: 10.1126/science.aam8940

Sequencing each cell of the nematode

Single-cell sequencing is challenging owing to the limited biological material available in an individual cell and the high cost of sequencing across multiple cells. Cao et al. developed a two-step combinatorial barcoding method to profile both single-cell and single-nucleus transcriptomes without requiring physical isolation of each cell. The authors profiled almost 50,000 single cells from an individual Caenorhabditis elegans larva and were able to identify and recover information from different, even rare, cell types.

Science, this issue p. 661

ARTICLE TOOLS

http://science.sciencemag.org/content/357/6352/661

SUPPLEMENTARY MATERIALS

http://science.sciencemag.org/content/suppl/2017/08/16/357.6352.661.DC1

REFERENCES

This article cites 133 articles, 57 of which you can access for free
http://science.sciencemag.org/content/357/6352/661#BIBL

PERMISSIONS

http://www.sciencemag.org/help/reprints-and-permissions

Use of this article is subject to the Terms of Service

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. 2017 © The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. The title Science is a registered trademark of AAAS.