RESEARCH ARTICLE

SINGLE-CELL GENOMICS

Comprehensive single-cell transcriptional profiling of a multicellular organism

Junyue Cao,[1,2]* Jonathan S. Packer,[1]* Vijay Ramani,[1]† Darren A. Cusanovich,[1]† Chau Huynh,[1] Riza Daza,[1] Xiaojie Qiu,[1,2] Choli Lee,[1] Scott N. Furlan,[3,4,5] Frank J. Steemers,[6] Andrew Adey,[7,8] Robert H. Waterston,[1]‡ Cole Trapnell,[1]‡ Jay Shendure[1,9]‡

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.

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 numbers of cells processed (*18*, *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 chromosome 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., 762 somatic cells per L2 larva), our knowledge of the molecular state of each cell and cell type has remained fragmentary. 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 5 (Tn*5*) transposase, lysis, and polymerase chain reaction (PCR) amplification are performed. The PCR primers target the barcoded polythymidine primer on one end and the Tn*5* 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 Tn*5* 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 262 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 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

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [2]Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA. [3]Ben Towne Center for Childhood Cancer Research, Seattle Children's Research Institute, Seattle, WA, USA. [4]Department of Pediatrics, University of Washington, Seattle, WA, USA. [5]Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [6]Advanced Research Group, Illumina, San Diego, CA, USA. [7]Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR, USA. [8]Knight Cardiovascular Institute, Portland, OR, USA. [9]Howard Hughes Medical Institute, Seattle, WA, USA.
*These authors contributed equally to this work. †These authors contributed equally to this work. ‡Corresponding author. Email: watersto@uw.edu (R.H.W.); coletrap@uw.edu (C.T.); shendure@uw.edu (J.S.)

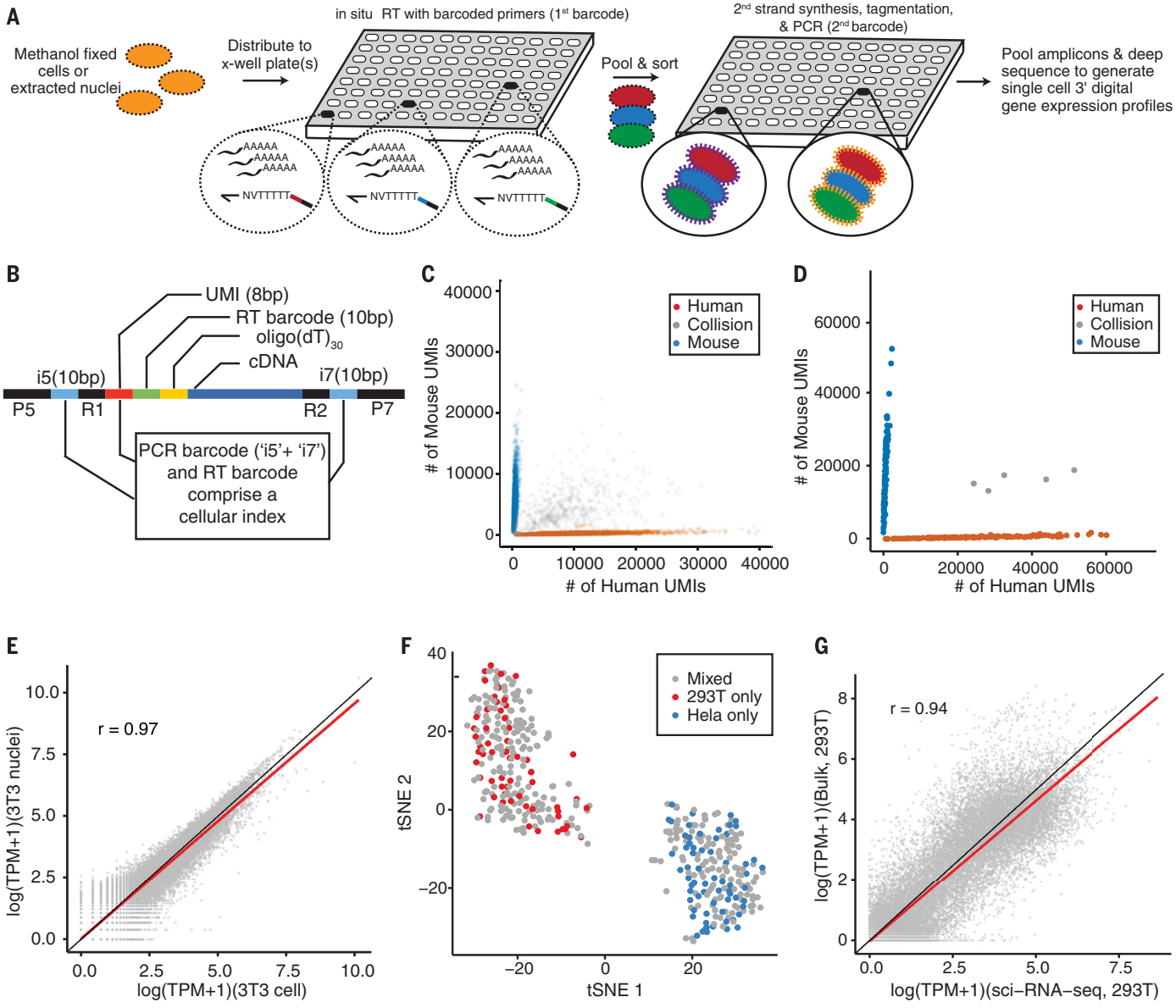

Fig. 1. sci-RNA-seq enables multiplex single-cell transcriptome profiling. (**A**) Schematic of the sci-RNA-seq workflow. AAAAA, polyadenosine tail; NVTTTTT, polythymidine primer. (**B**) Schematic of sci-RNA-seq library amplicons for Illumina sequencing. bp, base pairs; R, annealing sites for Illumina sequencing primers; P, Illumina P5 or P7 adaptor sequence. (**C**) Scatter plot of unique molecular identifier (UMI) counts from human and mouse cells, determined by 384 × 384 sci-RNA-seq. Blue, inferred mouse cells (n = 5953). Red, inferred human cells (n = 3967). Gray, collisions (n = 884). (**D**) Scatter plot of UMI counts from human and mouse cells, determined by 96 × 96 sci-RNA-seq with an optimized protocol. Blue, inferred mouse cells (n = 129). Red, inferred human cells (n = 160). Gray, collisions (n = 5). In (C) and (D), only cells originating from wells containing mixed human and mouse cells are shown. (**E**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of NIH/3T3 cells (n = 238) and nuclei (n = 124). (**F**) t-SNE plot of cells originating in wells containing HEK293T (red; n = 60), HeLa S3 (blue; n = 69), or a mixture (gray; n = 321). (**G**) Correlation between gene expression measurements from aggregated sci-RNA-seq data and bulk RNA-seq data obtained using a related protocol (*29*). In (E) and (G), the red line is the linear regression, and the black line is $y = x$.

(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

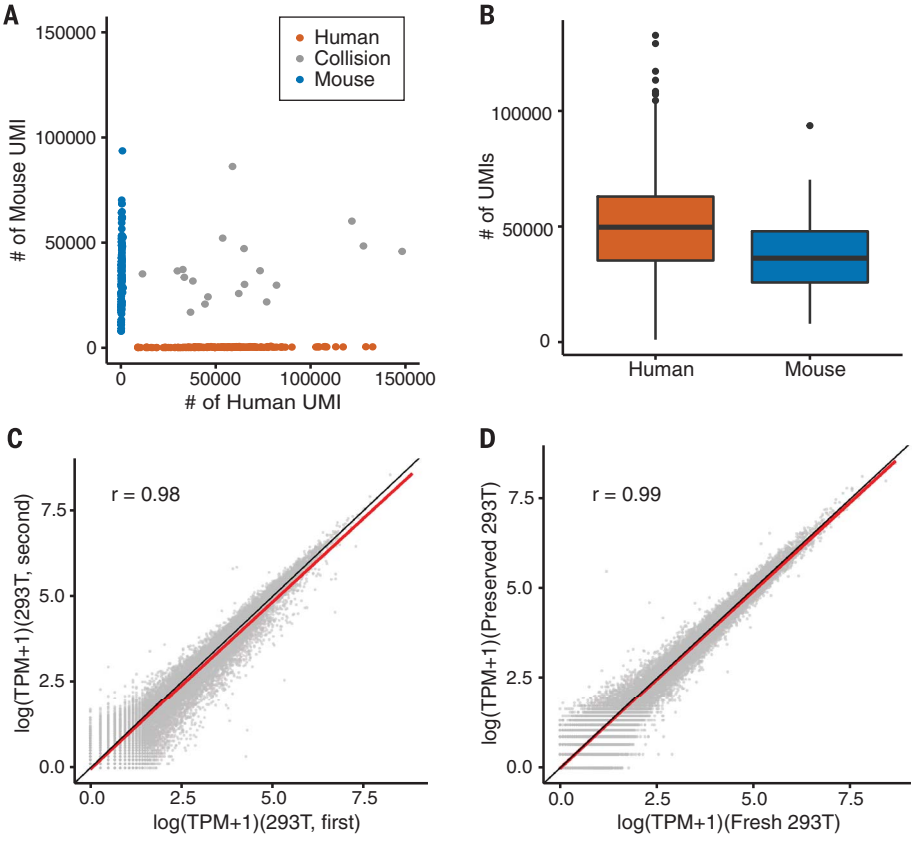

Fig. 2. sci-RNA-seq shows robust gene expression measurements. (**A**) Scatter plot of UMI counts from human and mouse cells, determined by a 16 × 84 sci-RNA-seq experiment on mixed HEK293T and NIH/3T3 cells (table S1). Blue, inferred mouse cells (n = 109). Red, inferred human cells (n = 168). Gray, collisions (n = 19). (**B**) Box plots showing the number of UMIs detected per cell (thick horizontal lines, medians; upper and lower box edges, first and third quartiles, respectively; whiskers, 1.5 times the interquartile range; circles, outliers). (**C**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles from two experiments 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 efficiency per molecule. Consistent with this, optimizing the number of cells per RT reaction increased UMI counts per cell (*28*).

Estimates of gene expression from the aggregated transcriptomes of nuclei and cells were well correlated [Pearson correlation coefficient (r) = 0.96 for HEK293T and 0.97 for NIH/3T3; Fig. 1E and fig. S2D]. From cells, 81% of reads mapped to the expected strand of genic regions (47% exonic and 34% intronic), and 19% mapped to intergenic regions or the unexpected strand of genic regions. From nuclei, 84% of reads mapped to the expected strand of genic regions (35% exonic and 49% intronic), and 16% mapped to intergenic regions or the unexpected strand of genic regions, similar to results from previous studies (*14*). Whereas exonic reads showed an expected enrichment at the 3′ ends of gene bodies, intronic reads did not, and they may be the result of polythymidine priming from polyadenosine tracts in heterogeneous nuclear RNA (fig. S3).

Transcriptomes originating in the 48 wells containing pure or an intraspecies mixture of HEK293T and HeLa S3 cells were readily separated into two clusters by *t*-distributed stochastic neighbor embedding (t-SNE) (Fig. 1F and fig. S4). Estimates of gene expression from the aggregated transcriptomes of all identified HEK293T cells versus those from a related bulk RNA-seq workflow without methanol fixation [Tn*5*-RNA-seq (*29*)] were well correlated (r = 0.94; Fig. 1G).

Robustness of sci-RNA-seq

After optimizing the number of cells per RT reaction, we fixed a mixture of HEK293T and NIH/3T3 cells and performed 16 × 84–well sci-RNA-seq (table S1) (*28*). We recovered 168 human cells and 109 mouse cells with 19 collisions (Fig. 2A). At ~240,000 reads per cell (73% duplication rate), we observed an average of 49,043 UMIs (7563 genes) per human cell and 36,737 UMIs (6263 genes) per mouse cell (Fig. 2B and fig. S5A), with 0.9 and 1.2% 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 (4503 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 routine profiling of ~10^4 single cells per experiment. We tested an additional level of indexing during Tn*5* transposition of double-stranded cDNA (*22*). We performed 16 × 6 × 16–well sci-RNA-seq on mixed HEK293T and NIH/3T3 cells after methanol fixation. After RT with 16 barcodes and second-strand synthesis, cells were pooled and distributed to six wells for tagmentation with indexed Tn*5* (six barcodes), then pooled again and sorted to 16 wells for PCR with indexed primers. 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 average, 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 Tn*5*. This limitation notwithstanding, three-level combinatorial indexing has the potential to enable routine profiling of >10^6 single cells per experiment [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 optimized the protocol. We pooled ~150,000 larvae synchronized at the L2 stage and dissociated them into single-cell suspensions. We then performed in situ RT across six 96-well plates (576 first-round barcodes), each well containing ~1000 *C. elegans* 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 barcoding (960 second-round barcodes), gating on DNA content to distinguish between *C. elegans*

and HEK293T cells. This sorting resulted in 96% of wells harboring only *C. elegans* cells (140 each) and 4% of wells harboring a mix of *C. elegans* and HEK293T cells (140 *C. elegans* and 10 HEK293T each).

This experiment yielded 42,035 *C. elegans* single-cell transcriptomes (UMI counts per cell for protein-coding genes ≥ 100). Ninety-four percent of reads mapped to the expected strand of genic regions (92% exonic and 2% intronic). At a sequencing depth of ~20,000 reads per cell and a duplication rate of 80%, we identified a median of 575 UMIs mapping to protein-coding genes 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 37,734 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 somatic cells 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 7325 cells (UMI counts per cell for protein-coding genes ≥ 200), of which 6335 were somatic and 511 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

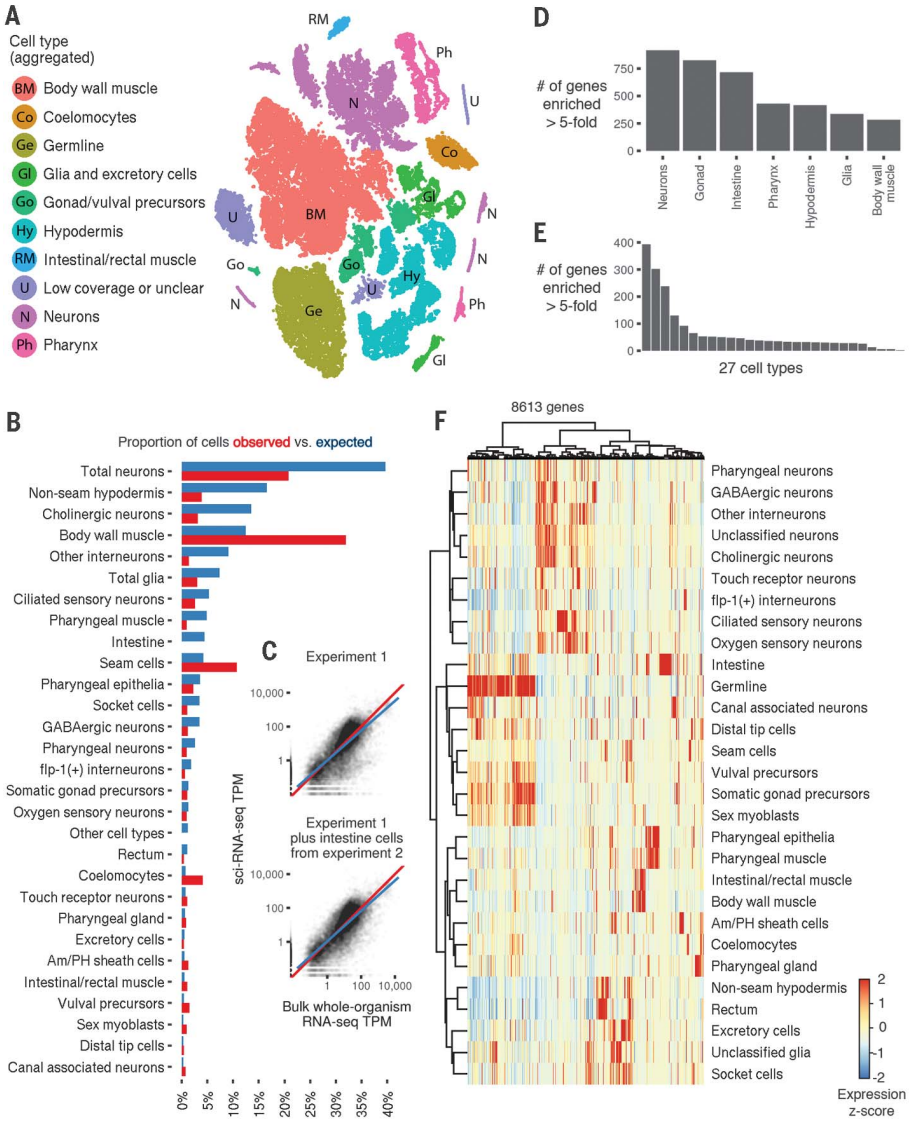

Fig. 3. A single sci-RNA-seq experiment highlights the single-cell transcriptomes of the *C. elegans* larva. (**A**) t-SNE visualization of the high-level cell types identified. (**B**) Bar graph showing the percentage of somatic cells profiled in the first sci-RNA-seq *C. elegans* experiment that could be identified as belonging to each cell type (red), compared with the percentage of cells from that type expected in an L2 *C. elegans* individual (blue). (**C**) Scatter plots showing the log-scaled transcripts per million (TPM) values of genes in the aggregation of all sci-RNA-seq reads (x axis) or in bulk RNA-seq (y axis; geometric mean of three experiments). Red line, $y = x$; blue line, linear regression. The top plot includes only the first sci-RNA-seq experiment. The bottom plot also includes intestine cells from the second sci-RNA-seq experiment. (**D**) Number of genes that are at least five times as highly expressed in a specific tissue as in the second-highest-expressing tissue, excluding genes for which the differential expression between the first- and second-highest expressing tissues is not significant ($q > 0.05$). (**E**) Same as (D), except comparing cell types instead of tissues. (**F**) Heat map showing the relative expression of genes in consensus transcriptomes for each cell type, estimated by sci-RNA-seq. Genes are included if they have a size factor–normalized mean expression of >0.05 in at least one cell type (8613 genes in total). The raw expression data (UMI count matrix) is log-transformed, column-centered, and scaled (using the R function scale), and the resulting values are clamped to the interval (−2, 2). GABA, γ-aminobutyric acid.

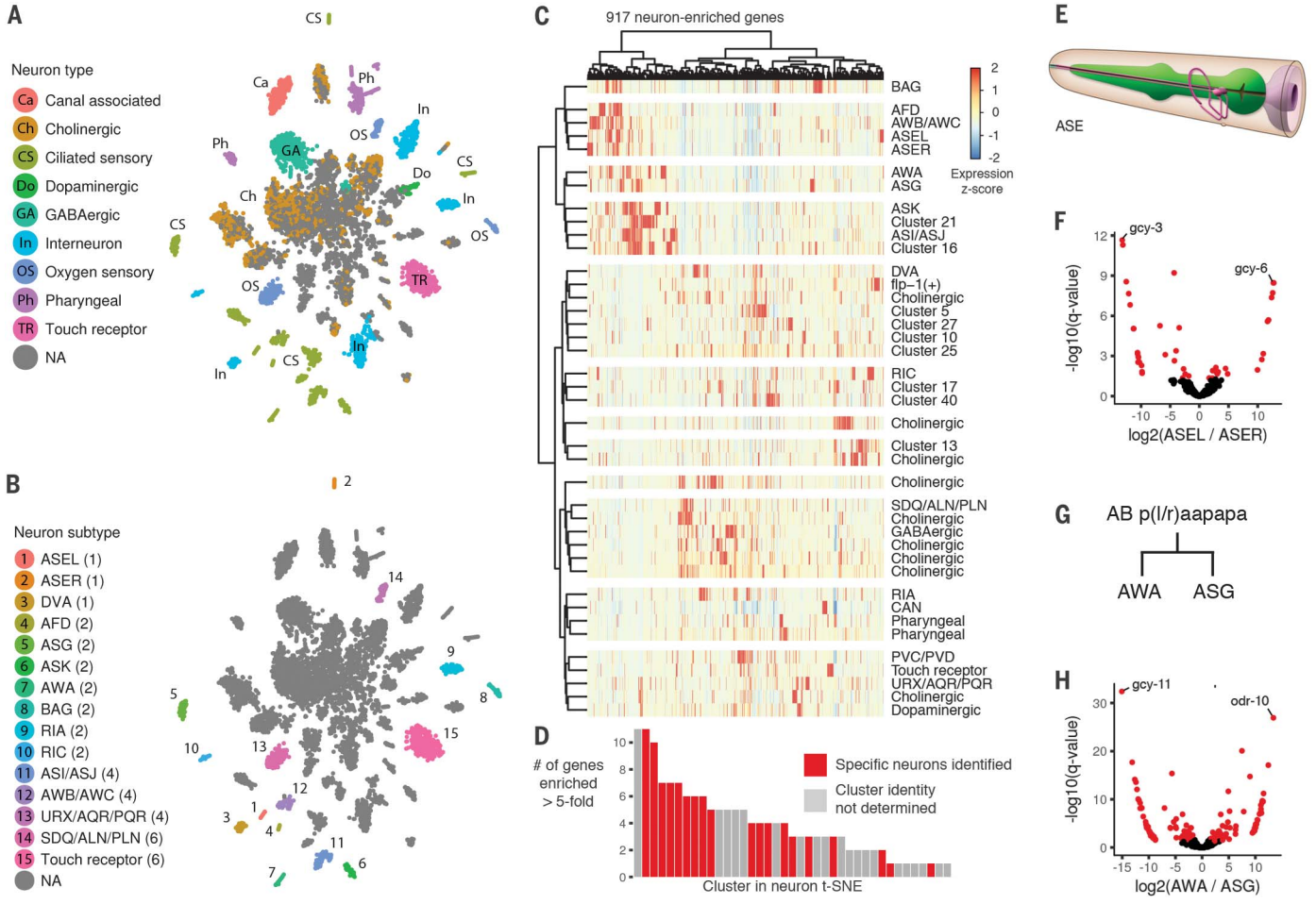

Fig. 4. sci-RNA-seq reveals the transcriptomes of fine-grained anatomical classes of *C. elegans* neurons. (**A**) t-SNE visualization of high-level neuronal subtypes. Cells identified as neurons from the t-SNE clustering shown in Fig. 3A were reclustered with t-SNE. NA, not assigned. (**B**) Clusters 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 (q 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 ($q < 0.05$) 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 plaapapa; the right AWA and ASG neurons arise from AB praapapa. (**H**) Volcano plot showing differentially expressed genes between the AWA and ASG neurons.

C. elegans experiment in subsequent analyses, with all other cell types being represented by the first experiment only.

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 GExplore (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

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 reclustering 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 AFD, ASG, ASK, AWA, BAG, CAN, RIA, and RIC neuron pairs), and three clusters corresponded to exactly one neuron [the left 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 (out of 40 total) in which they were expressed at ≥10% of the level of the highest-expressed cluster. Furthermore, 155 of these genes were expressed predominantly in a single neuronal cluster (at least a fivefold difference between the highest- and second-highest-expressing neuronal cluster) (Fig. 4D and table S8).

Expressions of marker genes, such as *gcy-3* and *gcy-6*, were key in identifying two neuronal clusters as the ASEL and ASER gustatory neurons, respectively (Fig. 4E). These neurons have asymmetry in gene expression (*38*), and we observed 44 genes to be differentially expressed (fold difference > 3, false discovery rate < 5%; Fig. 4F and table S12). mRNA from these neurons has previously been profiled with coimmunoprecipitation of RNA and a transgenic polyadenylate-binding protein expressed specifically in ASEL or ASER, followed by microarray analysis (*39*). The differentially expressed genes that we observed are consistent with this previous study (fig. S11), highlighting the ability of sci-RNA-seq to facilitate the analysis of cell types with frequencies as rare as a single cell per individual.

Two neuronal clusters correspond to sister cells, the AWA and ASG neurons (Fig. 4G), which arise from the same parental cell in the last round of *C. elegans* embryonic cell divisions. Their differentiation has previously been used as a model for the study of the regulation of cell fate decisions (*40*). In our data, 136 genes were differentially expressed between these two cell types (fold difference > 3, false discovery rate < 5%; Fig. 4H and table S13). The divergent transcriptomes of the AWA and ASG neurons, along with those of the ASEL 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—profiled 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

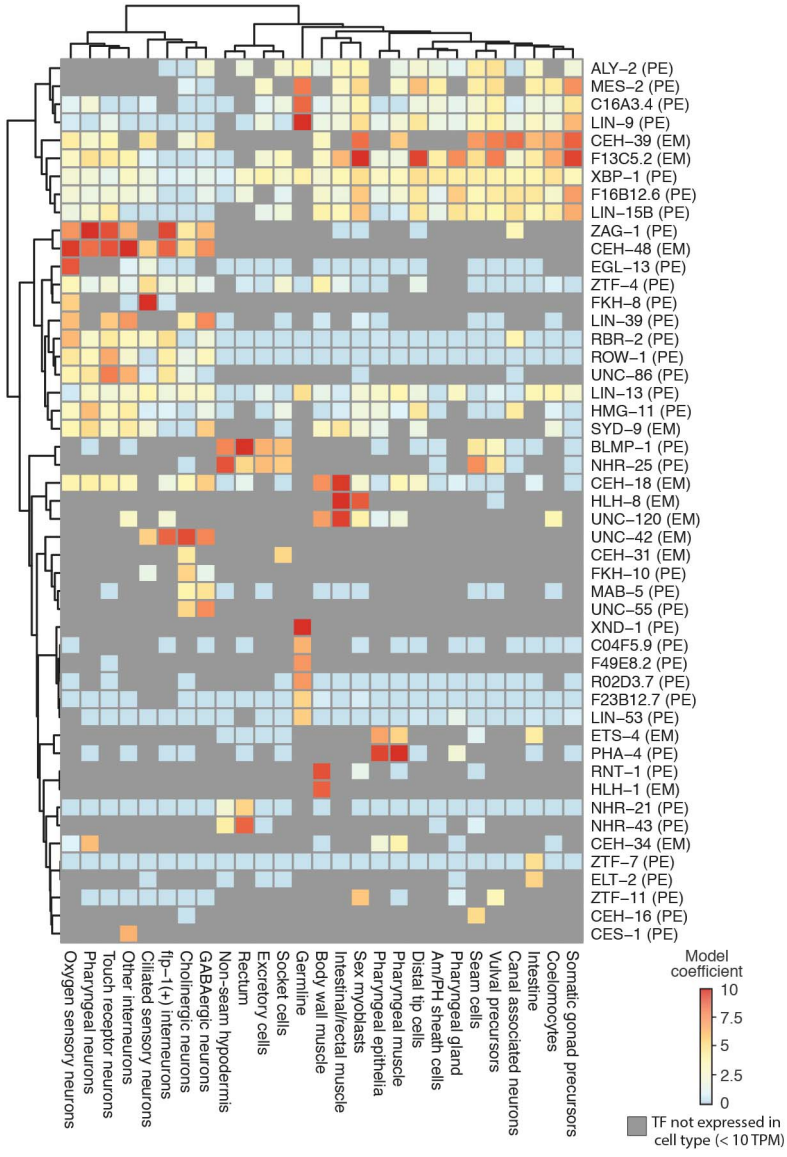

Fig. 5. Cell type–specific expression profiles from sci-RNA-seq enable the deconvolution of whole-animal transcription factor ChIP-seq data. For each of 27 cell types, a regularized regression model was fit to predict log-transformed gene expression levels in that cell type on the basis of ChIP-seq peaks in gene promoters (*28*). The ChIP-seq data were generated by the modENCODE (*41*) and modERN (*42*) consortia, profiling transcription factor binding in whole *C. elegans* animals. "EM" next to a TF label indicates that the ChIP-seq data for the TF are from an embryonic stage; "PE" indicates that the data are from a postembryonic stage. Colors in the heat map show the extent to which having a ChIP-seq peak for a given TF in a gene promoter correlates with increased expression in a given cell type. Peaks in "HOT" regions (*28*) are excluded. Gray cells in the heat map correspond to cases where a TF is not expressed in a cell type (<10 TPM), in which case ChIP-seq data for that TF are not considered by the regression model.

were detectably expressed within it [>10 transcripts per million (TPM)], increasing the proportion of TF–cell type associations that are likely to reflect causal gene regulation. Our regression model selected numerous regulators that are critical for the development or proper function of specific cell types, including *hlh-1* and *unc-120* in body wall muscle (*43*), *pha-4* in pharyngeal cell types (*44*), *hlh-8* (encoding CeTwist) in sex myoblasts (*45*), *blmp-1* and *nhr-25* in the hypodermis (*46*, *47*), *elt-2* in the intestine (*48*), and *xnd-1* in the germ line (*49*, *50*).

The regression identified several previously unknown regulators of cell type–specific expression. For example, *fkh-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 germ line and associated with germline gene expression (fig. S12). The gene encoding F49E8.2 is an ortholog of the human gene *E2F-associated phosphoprotein* (*EAPP*) (*52*), and F49E8.2 ChIP-seq peaks colocalize with germ line-specific EFL-1 peaks [ortholog of E2F; data from (*53*)] more often than could be expected as a result of chance (χ^2 test, $P = 2.8 \times 10^{-21}$; fig. S13, A and B), suggesting that these proteins may physically interact. The hypodermis-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 socket cells as in seam cells (560 versus 124 TPM) and 8.7 times as much as in the nonseam hypodermis (560 versus 64 TPM), suggesting a role in glial development.

Discussion

Our method for single-cell RNA-seq combinatorial indexing of cells or nuclei, sci-RNA-seq, can be applied to profile the transcriptomes of tens of thousands of single cells per experiment through a library construction completed by a single person in 2 days, at a cost of $0.03 to $0.20 per cell. sci-RNA-seq is compatible with cell fixation, which can minimize perturbations to cell state or RNA integrity before or during processing and facilitates the concurrent processing of multiple samples within a single experiment, potentially reducing batch effects relative to platforms requiring serial processing, an area of concern for the single-cell RNA-seq field (*54*). Given that the second barcode is introduced after flow sorting, it is also possible to associate wells on the PCR plate with FACS-defined subpopulations. sci-RNA-seq is also compatible with nuclei, which may be important for tissues for which unbiased cell disaggregation protocols are not well established (possibly most tissues). Lastly, sci-RNA-seq is scalable. We demonstrated indexing up to 576 × 960, which enabled the generation of ~4×10^4 single-cell transcriptomes in one experiment. However, processing of more cells with sublinear cost scaling is possible by using more barcoded RT and PCR primers (e.g., 1536 × 1536 combinatorial indexing) and/or introducing additional rounds of indexing. With 384 × 384 × 384 combinatorial indexing, one could hypothetically profile the transcriptomes of more than 10 million cells per experiment.

With sci-RNA-seq, we generated a catalog of single-cell transcriptomes with >50-fold "shotgun" cellular coverage of the L2 *C. elegans* soma. We detected 18 non-neuronal cell types and many neuronal cell types, which we grouped into either 10 broad classes or 40 fine-grained clusters from an unsupervised analysis, highlighting the potential of an organism's gene regulatory programs to be enacted at a fine-grained level. We anticipate that these data will be a rich resource for nematode biology—a starting point for an atlas that leverages Sulston's lineage map to define the molecular state of every cell throughout the life cycle of *C. elegans*. In addition, as illustrated by our experience with intestinal cells, the greater knowledge of "ground truth" in *C. elegans* may further the refinement of experimental and computational methods for recovering and distinguishing cell types and states. To this end, we have created a website to facilitate the further annotation of these data by the community (http://atlas.gs.washington.edu). Gene-by-cell matrices and vignettes for how to work with the data are also hosted at this site.

sci-RNA-seq expands the repertoire of single-cell molecular phenotypes that can be resolved by combinatorial indexing (*22*–*25*). Provided that multiple aspects of cellular biology can be concurrently barcoded, combinatorial indexing may also facilitate the scalable generation of "joint" single-cell molecular profiles (e.g., 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 (*55*) of single cells in other organisms will begin to give shape to global views of their developmental biology.

REFERENCES AND NOTES

1. C. Trapnell, *Genome Res.* **25**, 1491–1498 (2015).
2. D. Ramsköld *et al.*, *Nat. Biotechnol.* **30**, 777–782 (2012).
3. A. K. Shalek *et al.*, *Nature* **498**, 236–240 (2013).
4. Q. F. Wills *et al.*, *Nat. Biotechnol.* **31**, 748–752 (2013).
5. G. X. Y. Zheng *et al.*, *Nat. Commun.* **8**, 14049 (2017).
6. A. A. Pollen *et al.*, *Nat. Biotechnol.* **32**, 1053–1058 (2014).
7. A. Zeisel *et al.*, *Science* **347**, 1138–1142 (2015).
8. B. B. Lake *et al.*, *Science* **352**, 1586–1590 (2016).
9. I. Tirosh *et al.*, *Science* **352**, 189–196 (2016).
10. W. Zeng *et al.*, *Nucleic Acids Res.* **44**, e158 (2016).
11. C. Trapnell *et al.*, *Nat. Biotechnol.* **32**, 381–386 (2014).
12. F. Tang *et al.*, *Nat. Methods* **6**, 377–382 (2009).
13. S. Picelli *et al.*, *Nat. Methods* **10**, 1096–1098 (2013).
14. R. V. Grindberg *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **110**, 19802–19807 (2013).
15. H. C. Fan, G. K. Fu, S. P. A. Fodor, *Science* **347**, 1258367 (2015).
16. E. Z. Macosko *et al.*, *Cell* **161**, 1202–1214 (2015).
17. A. M. Klein *et al.*, *Cell* **161**, 1187–1201 (2015).
18. A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, S. A. Teichmann, *Mol. Cell* **58**, 610–620 (2015).
19. S. Liu, C. Trapnell, *F1000Res.* **5(F1000 Faculty Rev)**, 182 (2016).
20. A. Adey *et al.*, *Genome Res.* **24**, 2041–2049 (2014).
21. S. Amini *et al.*, *Nat. Genet.* **46**, 1343–1349 (2014).
22. D. A. Cusanovich *et al.*, *Science* **348**, 910–914 (2015).
23. S. A. Vitak *et al.*, *Nat. Methods* **14**, 302–308 (2017).
24. V. Ramani *et al.*, *Nat. Methods* **14**, 263–266 (2017).
25. R. M. Mulqueen *et al.*, Scalable and efficient single-cell DNA methylation sequencing by combinatorial indexing. BioRxiv 157230 [Preprint]. 28 June 2017. https://doi.org/10.1101/157230.
26. J. E. Sulston, E. Schierenberg, J. G. White, J. N. Thomson, *Dev. Biol.* **100**, 64–119 (1983).
27. J. E. Sulston, H. R. Horvitz, *Dev. Biol.* **56**, 110–156 (1977).
28. Materials and methods are provided as supplementary materials.
29. J. Gertz *et al.*, *Genome Res.* **22**, 134–141 (2012).
30. E. M. Hedgecock, J. G. White, *Dev. Biol.* **107**, 128–133 (1985).
31. G. V. Clokey, L. A. Jacobson, *Mech. Ageing Dev.* **35**, 79–94 (1986).
32. G.-J. Hendriks, D. Gaidatzis, F. Aeschimann, H. Großhans, *Mol. Cell* **53**, 380–392 (2014).
33. G. Heimberg, R. Bhatnagar, H. El-Samad, M. Thomson, *Cell Syst.* **2**, 239–250 (2016).
34. M. E. Boeck *et al.*, *Genome Res.* **26**, 1441–1450 (2016).
35. R. Ruksana *et al.*, *Genes Cells* **10**, 261–276 (2005).
36. J. G. White, E. Southgate, J. N. Thomson, S. Brenner, *Philos. Trans. R. Soc. London B Biol. Sci.* **314**, 1–340 (1986).
37. O. Hobert, L. Glenwinkel, J. White, *Curr. Biol.* **26**, R1197–R1203 (2016).
38. O. Hobert, R. J. Johnston Jr., S. Chang, *Nat. Rev. Neurosci.* **3**, 629–640 (2002).
39. J. Takayama, S. Faumont, H. Kunitomo, S. R. Lockery, Y. Iino, *Nucleic Acids Res.* **38**, 131–142 (2010).
40. T. R. Sarafi-Reinach, T. Melkman, O. Hobert, P. Sengupta, *Development* **128**, 3269–3281 (2001).
41. C. L. Araya *et al.*, *Nature* **512**, 400–405 (2014).
42. modERN consortium, ENCODE, 2017; https://encodeproject.org/.
43. T. Fukushige, T. M. Brodigan, L. A. Schriefer, R. H. Waterston, M. Krause, *Genes Dev.* **20**, 3395–3406 (2006).
44. J. Gaudet, S. E. Mango, *Science* **295**, 821–825 (2002).
45. B. D. Harfe *et al.*, *Genes Dev.* **12**, 2623–2635 (1998).
46. M. Horn *et al.*, *Dev. Cell* **28**, 697–710 (2014).
47. C. R. Gissendanner, A. E. Sluder, *Dev. Biol.* **221**, 259–272 (2000).
48. T. Fukushige, M. G. Hawkins, J. D. McGhee, *Dev. Biol.* **198**, 286–302 (1998).
49. C. R. Wagner, L. Kuervers, D. L. Baillie, J. L. Yanowitz, *Nature* **467**, 839–843 (2010).
50. R. Mainpal, J. Nance, J. L. Yanowitz, *Development* **142**, 3571–3582 (2015).
51. I. A. Hope, A. Mounsey, P. Bauer, S. Aslam, *Gene* **304**, 43–55 (2003).
52. D. D. Shaye, I. Greenwald, *PLOS ONE* **6**, e20085 (2011).
53. M. Kudron *et al.*, *Genome Biol.* **14**, R5 (2013).
54. P.-Y. Tung *et al.*, *Sci. Rep.* **7**, 39921 (2017).
55. A. McKenna *et al.*, *Science* **353**, aaf7907 (2016).
56. Z. Altun, D. Hall, "ASEL, ASER," in *WormAtlas*, 2017; www.wormatlas.org/neurons/Individual%20Neurons/ASEframeset.html.

ACKNOWLEDGMENTS

The raw data have been deposited with the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession code GSE98561. We thank members of the Shendure, Trapnell, and Waterston laboratories for helpful discussions and feedback, particularly A. Hill, V. Agarwal, M. Gasperini, L. Starita, Y. Yin, and B. Martin; S. Zimmerman and C. Berg for helpful technique suggestions; the modERN consortium for allowing us to use their ChIP-seq data; D. Prunkard and L. Gitari in the Pathology Flow Cytometry Core Facility for their exceptional assistance in flow sorting; the T. Reh laboratory for sharing the NIH/3T3 cell line; and H. Hutter for adding our tissue-specific expression profiles to gExplore. HeLa S3 cells were used as part of this study. Henrietta Lacks, and the HeLa cell line that was established from her tumor cells in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. This work was funded by grants from the NIH (DP1HG007811 and R01HG006283 to J.S., U41HG007355 and R01GM072675 to R.H.W, and DP2 HD088158 to C.T.), the Paul G. Allen Family Foundation (to J.S.), the W. M. Keck Foundation (to C.T. and J.S.), the Dale F. Frey Award for Breakthrough Scientists (to C.T.), the Alfred P. Sloan Foundation Research Fellowship (to C.T.), and the William Gates III Endowed Chair in Biomedical Sciences (to R.H.W.). D.A.C. was supported in part by T32HL007828 from the National Heart, Lung, and Blood Institute. J.S. is an investigator of the Howard Hughes Medical Institute. F.J.S. declares competing financial interests in the form of stock ownership and paid employment by Illumina. One or more embodiments of one or more patents and patent applications filed by Illumina may encompass the methods, reagents, and data disclosed in this manuscript.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/357/6352/661/suppl/DC1
Materials and Methods
Figs. S1 to S24
Tables S1 to S14
References (*57*–*140*)

1 February 2017; resubmitted 12 May 2017
Accepted 19 July 2017
10.1126/science.aam8940

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