



Comprehensive characterization of tissue-specific chromatin accessibility in L2 *C. elegans* nematodes

Timothy J Durham, Riza M Daza, Louis Gevirtzman, et al.

Genome Res. published online April 22, 2021

Access the most recent version at doi:[10.1101/gr.271791.120](https://doi.org/10.1101/gr.271791.120)

P<P	Published online April 22, 2021 in advance of the print journal.
Accepted Manuscript	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
Open Access	Freely available online through the <i>Genome Research</i> Open Access option.
Creative Commons License	This manuscript is Open Access. This article, published in <i>Genome Research</i> , is available under a Creative Commons License (Attribution 4.0 International license), as described at http://creativecommons.org/licenses/by/4.0/ .
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here .



To subscribe to *Genome Research* go to:
<https://genome.cshlp.org/subscriptions>

Published by Cold Spring Harbor Laboratory Press

Comprehensive characterization of tissue-specific chromatin accessibility in L2 *Caenorhabditis elegans* nematodes

Timothy J. Durham^{1,†}, Riza M. Daza¹, Louis Gevirtzman¹, Darren A. Cusanovich²,
Olubusayo Bolonduro^{1,‡}, William Stafford Noble^{1,3}, Jay Shendure^{1,4,5,6}, and Robert H.
Waterston^{1,*}

¹Department of Genome Sciences, University of Washington, Foege Building S-250, Box
355065, Seattle, WA 98195

²Department of Cellular and Molecular Medicine, University of Arizona, BSRL 423, 1230 N
Cherry Avenue, Tucson, AZ 85721

³Paul G. Allen School of Computer Science & Engineering, University of Washington, Box
352350, Seattle, WA 98195

⁴Howard Hughes Medical Institute, 4000 Jones Bridge Road, Chevy Chase, MD 20815

⁵Brotman-Baty Institute for Precision Medicine, Magnuson Health Sciences Building H-564,
Box 357657, Seattle, WA 98195

⁶Allen Discovery Center for Cell Lineage Tracing, University of Washington, Box 355065,
Seattle, WA 98195

[†]Present address: Broad Institute of MIT and Harvard, 415 Main Street, Cambridge, MA
02142

[‡]Present address: Committee on Genetics, Genomics, and System Biology, The University of
Chicago, Cummings Life Science Center, 920 East 58th Street, Chicago, IL 60637

*Corresponding author: watersto@uw.edu

April 11, 2021

Abstract

Recently developed single cell technologies allow researchers to characterize cell states at ever greater resolution and scale. *C. elegans* is a particularly tractable system for studying development, and recent single cell RNA-seq studies characterized the gene expression patterns for nearly every cell type in the embryo and at the second larval stage (L2). Gene expression patterns give insight about gene function and into the biochemical state of different cell types; recent advances in other single cell genomics technologies can now also characterize the regulatory context of the genome that gives rise to these gene expression levels at a single cell resolution. To explore the regulatory DNA of individual cell types in *C. elegans*, we collected single cell chromatin accessibility data using the sci-ATAC-seq assay in L2 larvae to match the available single cell RNA-seq data set. Using a novel implementation of the latent Dirichlet allocation algorithm, we identify 37 clusters of cells that correspond to different cell types in the worm, providing new maps of putative cell type-specific gene regulatory sites, with promise for better understanding of cellular differentiation and gene regulation.

Introduction

Critical cellular processes are dependent on fine-tuned control of gene expression levels. From properly responding to environmental stimuli, to progressing through the stages of development and differentiation, specific changes in gene expression play an important role in facilitating precise changes in cellular state. Recent advances in single cell transcriptomics have enabled the massively parallel measurement of gene expression in individual cells, giving unprecedented genome-wide insight into which genes are regulated together in the same cell, and into expression dynamics over time. Determining which genes are important in which cells and under which conditions is critical to attaining a deeper understanding of gene function. However, in order to truly understand how gene expression reflects and influences cell state, we must also understand how it is controlled. The nematode *Caenorhabditis elegans* is a particularly powerful system in which to apply single cell genomics technologies because it has limited cell numbers that nonetheless form diverse tissue and cell types; it is very amenable to genetic manipulation; and the developmental lineage of every cell is known and invariant. In the past few years, the worm has been the subject of perhaps the most comprehensive cell type-specific metazoan gene expression atlas by single cell RNA-seq (scRNA-seq) (Cao et al., 2017; Packer et al., 2019). These studies of the *C. elegans* embryo (Packer et al., 2019) and second larval stage (L2) (Cao et al., 2017) provide a survey of the full complement of genes expressed in each major cell type, and even some cells present only once in the worm (e.g. the ASEL and ASER gustatory neurons). Now, to understand how these tissue-specific expression patterns arise, we also need to have a similarly comprehensive catalog of regulatory elements to map their activity in different cell types and at different stages of the life cycle.

Several efforts have been undertaken to map regulatory DNA in the worm (Araya et al., 2014; Ho et al., 2017; Daugherty et al., 2017; Kudron et al., 2018; Jänes et al., 2018). Collectively, these studies have identified tens of thousands of chromatin accessibility regions and transcription factor binding sites, using DNase-seq (Ho et al., 2017), ATAC-seq (Daugherty et al., 2017; Jänes et al., 2018), and ChIP-seq (Araya et al., 2014; Kudron et al., 2018) to assay developmental stages throughout the worm life cycle. The results show that the activity at many regulatory sites changes over the worm's lifespan. However, the data from all of these studies is from whole worms and thus does not resolve differences in regulatory activity across cell types. The lack of cell type resolution is problematic for three main reasons. First, gene regulation is often highly cell type-specific, and even when different cell types express the same gene, they may use different enhancers or promoters to regulate that gene. In the case of two sites regulating the same gene in different cell types at the same stage, a whole worm chromatin accessibility data set would only show that both sites are accessible at the same time, and would not reveal whether the sites act in concert in the same cell type or if they affect the same gene but act in different cell types. Distinguishing between these cases is critical for understanding and modeling gene regulation. The second reason is that whole worm data may lack the sensitivity to detect regulatory events that occur in cell types that make up small fractions of the whole worm. Single cell RNA-seq data (Cao et al., 2017; Packer et al., 2019) reveal important differences in gene expression that distinguish even individual cells. Such differences are presumably driven in part by regulatory regions that are only accessible in those cells; in a whole worm assay the signal from these highly cell type-specific regions could be drowned out by the noise generated from more populous cell types. Third, the lack of cell type resolution on these regulatory DNA maps confounds our ability to draw conclusions about differential activity across development. During development, the number of cells, and with them the diversity and proportion of cell types, is constantly changing. Thus, if an accessible site is less prominent in a later larval

stage compared to an embryonic stage, this change could mean the site is more accessible in embryogenesis than in later development, or it could reflect that the site is more specialized in later stages and is accessible in a smaller fraction of the cells. Given these important limitations in the available data on *C. elegans* gene regulation, we sought to generate cell type-resolved chromatin accessibility maps.

Over the past few years, the technology to collect chromatin accessibility profiles of single cells has improved greatly. This technology relies on the assay for transposase-accessible chromatin followed by high throughput sequencing (ATAC-seq) (Buenrostro et al., 2013), which treats permeabilized nuclei with a hyperactive Tn5 transposome from prokaryotes (Adey et al., 2010) to simultaneously cut accessible sites in the genome and ligate sequencing adapters onto the fragment ends on either side of the cut site (a reaction referred to as “tagmentation”). The resulting library is then amplified and sequenced. The simplicity of the assay significantly reduces the requirements for input material compared to DNase-seq (Song and Crawford, 2010), and protocols have adapted ATAC-seq to work on single cells (Buenrostro et al., 2015; Cusanovich et al., 2015; Chen et al., 2018). These and other studies have shown in multiple systems that single cell ATAC-seq (scATAC-seq) can measure thousands of sites per cell type and can identify distinct cell populations with high sensitivity. The use of the single cell combinatorial indexing variant of single cell ATAC-seq (sci-ATAC-seq) permits the cost-effective capture of scATAC-seq data. This assay probabilistically identifies DNA fragments isolated from single cells by first sorting 2500 nuclei per well into a 96 well plate and treating the nuclei in each well with a Tn5 enzyme loaded with adapters that contain a unique barcode sequence, and then pooling and re-sorting 25 nuclei per well into new 96-well plates in which a second set of barcodes is incorporated by using well-specific primers during library amplification. After sequencing, the reads can be assigned to a particular cell based on their combination of Tn5 and PCR barcodes. The sci-ATAC-seq assay has been successfully leveraged in several previous studies, including the identification of differences in gene regulation across germ layers in *Drosophila* embryogenesis (Cusanovich et al., 2018b); the generation of an atlas of 85 different clusters of cells from 13 different mouse tissues (Cusanovich et al., 2018a); and the identification of cell types in hippocampal tissue from mice (Sinnamon et al., 2019).

Here we sought to obtain a comprehensive cell type-resolved map of the regulatory DNA in a whole metazoan, the nematode *C. elegans*. We collected single cell chromatin accessibility data from tens of thousands of nuclei (hereafter referred to as cells for simplicity) isolated from second larval stage (L2) animals to match previously published single cell RNA-seq data (Cao et al., 2017). To contend with the sparsity of single cell chromatin accessibility data and to reduce dimensionality, we implemented an improved latent Dirichlet allocation (LDA) model (Blei et al., 2003; Griffiths and Steyvers, 2004) that can scale to tens of thousands of cells by parallelizing the training process across multiple cores, a feature that was unavailable in a previously published LDA implementation for single cell analysis (González-Blas et al., 2019). By training LDA models on our single cell ATAC-seq data and analyzing the results in conjunction with the L2 sci-RNA-seq expression data (Cao et al., 2017), we were able to identify the tissue and even cell type origins of the cells. We compared our maps of chromatin accessibility with those from prior studies of accessible chromatin in whole animals, both to validate our results and to determine which sites were novel to our data. Finally, we assessed whether genes expressed broadly across tissues might nevertheless show multiple, more cell-type specific regulatory patterns, for example by having tissue-specific accessible sites at alternative 5' ends. We anticipate that these data will provide a valuable resource for studying regulatory biology in the worm and set the stage for future single cell ATAC-seq experiments on

additional life stages. In conjunction with cell type-specific gene expression data, these maps of candidate regulatory regions will help reveal the gene regulatory networks driving development in *C. elegans*.

Results

Single cell chromatin accessibility in *C. elegans* with sci-ATAC-seq

To match the sci-RNA-seq data, we grew a synchronized population of wild-type worms to the middle of the second larval (L2) stage. At this stage, about 700 of the 959 somatic cells in the adult hermaphrodite have been produced, and the vast majority are terminally differentiated, but the development of the gonad has not progressed far enough to begin producing the thousands of germline nuclei that at later developmental stages would severely bias our collection of tissue types. After harvesting the worms, we fixed and isolated the nuclei, froze them in aliquots, and used these wild-type nuclei as input to the single cell combinatorial indexing assay for transposase-accessible chromatin (sci-ATAC-seq) (Cusanovich et al., 2015, 2018b,a). We collected sci-ATAC-seq data for 30,930 cells with at least 150 unique reads per cell (median 672 reads per cell), which represents about 40x sampling of each cell in the L2 worm (Fig. S1). Note that we expect thousands of genes to be expressed per cell, and thus at a median of 672 reads per cell we are only sampling a small fraction of the accessible regulatory sites in any given cell.

The post-sequencing pipeline consists of aligning the paired-end reads to the WS235/ce11 version of the *C. elegans* genome, and identifying cut sites as the 60 bp regions centered on the ends of the DNA fragments defined by the mapped mate pairs. (See Methods for details.) Next, we identified which loci were accessible in each of our cells. There exists no unbiased annotation of cell type-resolved regulatory regions in *C. elegans*, so we called peaks directly from the sci-ATAC-seq data in an iterative fashion (Fig. 1A). The first step was to call peaks with MACS2 (Zhang et al., 2008) using all of the reads together, as with a bulk ATAC-seq data set. To detect additional cell type-specific peaks that could be obscured by the background reads coming from the most abundant cell types in this complex cell mixture, we clustered the cells based on the output of the latent Dirichlet allocation (LDA) modeling technique (Blei et al., 2003) using the bulk peaks, pooled the cut sites from the cells in each cluster, and called peaks for each cluster using MACS2 (“Primary LDA” in Fig. 1B). Finally, in order to refine our peak set we repeated the clustering, pooling, and peak calling once more (“Refinement LDA” in Fig. 1B).

Each peak calling iteration centers around modeling the peak distribution in individual cells, as represented by a cells-by-peaks matrix containing binary values indicating whether or not a cell reports a read overlapping a particular peak. This binary matrix is input into LDA, a Bayesian modeling approach that was originally developed in the field of semantic analysis of text documents, and has recently been used to model single cell genomics data (González-Blas et al., 2019; Kim et al., 2019; Dey et al., 2017). We discuss this modeling approach in detail below and in Methods. After our iterative clustering and peak-calling procedure, we post-processed the merged peaks from the refinement LDA step by detecting local maxima (summits) in the signal over each peak and splitting any peaks with multiple summits into separate contiguous segments. This step splits wide peaks to better capture accessible regions that may contain multiple binding sites. We share the pipeline output in a UCSC track hub, which can be accessed at the following URL: http://genome.ucsc.edu/cgi-bin/hgTracks?db=ce11&hubClear=http://waterston.gs.washington.edu/atacTissue/Durham_hub.txt.

Single cell peaks are concordant with regulatory regions from published bulk chromatin assays

After applying our iterative peak calling procedure, the whole worm refinement LDA (Fig. 1) identified 36,339 peak regions, and splitting the multi-summit peaks resulted in a total of 38,017 peaks. To compare our peak calls to other maps of regulatory DNA in *C. elegans*, we intersected our regions with peaks from two other data sets: bulk, whole worm ATAC-seq data from samples spanning the *C. elegans* life cycle (Jänes et al., 2018), and transcription factor (TF) binding site peaks identified from 427 whole-worm TF ChIP-seq data sets from the modERN consortium (Kudron et al., 2018) (Fig. 2).

We find good overlap with both published data sets but also some differences (Fig. 2A). First, intersecting the sci-ATAC-seq peaks with the bulk ATAC-seq data set shows 25,675 peaks of 38,017 (~66%) overlapping a bulk ATAC-seq site overall. This fraction is significantly greater than expected by chance (Fisher's exact test $p = 0$), since the sci-ATAC-seq peaks cover 20,234,260 bp in total (about 20.2% of the genome), and the bulk ATAC-seq peaks cover 6,376,655 bp (about 6.4% of the genome). About 50% (12,960) of the overlaps were with bulk ATAC-seq sites classified as enhancers (Jänes et al., 2018), about 40% (10,219) were with sites classified as promoters, and 10% (2,496) were with other kinds of sites (e.g. non-coding RNAs). We find more extensive overlap with TF ChIP-seq sites; 30,886 of the 38,017 sci-ATAC-seq peaks (~81%) overlap TF ChIP-seq peaks from modERN (Fig. 2A). The ChIP TF peaks cover 23,225,218 bp (~23.2% of the genome), and similarly to the bulk ATAC-seq overlaps the number of TF ChIP-seq peak overlaps is highly significant (Fisher's exact test $p = 0$).

Looking at the data from the opposite perspective (Fig. 2B), almost three quarters of bulk ATAC-seq sites overlap a sci-ATAC-seq peak (29,021 of 42,102, ~69%), and these overlaps are fairly evenly split between sites classified as promoters (10,684 of 13,833, ~77%) and enhancers (13,715 of 19,195, ~71%), with the remaining overlaps (4,622 of 9,074, ~51%) involving other categories of regulatory elements. In the case of the modERN TF sites, we find that the majority overlap a sci-ATAC-seq peak (23,863 of 41,542, ~57%). Most of the ChIP-seq sites that do not overlap a sci-ATAC-seq peak are "singletons" that were only observed in one of the 427 ChIP-seq data sets (13,579 of 17,679, ~77%). Since both the bulk ATAC-seq data set and the TF ChIP-seq data set contain data derived from across *C. elegans* development, we reasoned that many of the sites with no overlap in our sci-ATAC-seq peaks are specific to other developmental stages than L2. In support of this hypothesis, we find that the set of ChIP-seq sites that overlap sci-ATAC-seq peaks is enriched for those found in larval samples, especially L2, and depleted for sites observed in embryo and young adult (Fig. 2C). We find similar results for the bulk ATAC-seq sites (Fig. S2A), and the enrichment for TF ChIP-seq sites from L2 is even higher when only considering the singleton ChIP-seq sites (Fig. S2B). Of the 18,543 singleton TF ChIP-seq sites that do not overlap a sci-ATAC-seq peak, only 929 (~5%) are from data sets collected in L2.

We also checked the ChIP-seq signal over our sci-ATAC-seq peaks for four histone modifications (Jänes et al., 2018). The sci-ATAC-seq peaks are enriched for signal from the two activating histone marks H3K4me3 and H3K4me1, uncorrelated with signal from the H3K36me3 histone mark associated with actively-transcribed gene bodies, and depleted for signal from the repressive histone mark H3K27me3 (Fig. S3).

Latent Dirichlet allocation modeling reveals 37 clusters of cells.

To interpret our data at the level of tissues and cell types, we applied latent Dirichlet allocation (LDA) (Blei et al., 2003), a statistical modeling technique that is particularly well-suited to finding patterns in the sparse data generated by single cell genomics assays (Fig. 3A). LDA has been applied previously to analyze single cell chromatin accessibility (González-Blas et al., 2019), single cell chromatin conformation (Kim et al., 2019), and single cell gene expression data (Dey et al., 2017). LDA is a generative Bayesian modeling approach that was developed in the context of document classification. In the document classification task, the model is trained to identify information-rich words in a document corpus and to associate those words with latent topics that can distinguish the documents. The output consists of two matrices: one that captures the probability distribution of each topic over all words, and another that captures the probability distribution of each document over all topics. Thus, each topic is defined as some combination of words, and each document is modeled with some combination of topics based on its word content.

When applied to scATAC-seq data, cells are treated as documents, and peaks are treated as words. The model learns the peaks associated with latent “regulatory topics” that capture patterns that discriminate among regulatory states and cell types. The output consists of two matrices: one representing the distribution of peaks over topics, and another representing the distribution of topics over cells. A key advantage of LDA in this setting is that it handles sparse data quite well and leverages information from all cells at once to assign peaks to topics and all peaks at once to assign topics to cells.

We trained an LDA model with 55 topics, choosing 55 topics based on a 5-fold cross-validation hyperparameter search procedure (see Methods, Fig. S4). This model yielded a cells-by-topics matrix with 30,870 rows (one for each cell after filtering out cells with too few peaks detected) and 55 columns, and a peaks-by-topics matrix with 32,214 rows (one for each peak after filtering out outlier high- and low-coverage peaks) and 55 columns (Fig. 3A). [Note that in the text we transpose the topics-by-peaks matrix and refer to it as the peaks-by-topics matrix for consistency with the cells-by-topics matrix.] We expected that differences in chromatin accessibility among cell types would be the largest source of covariation in the data, and thus that there would be many topics that corresponded to distinct cell types.

To look for topics that might distinguish cell types, we first removed 15 topics that did not have a high probability in any subset of tightly grouped cells (Methods, Fig. S5). Next, we assigned cells with greater than 50% probability for one of the 40 remaining topics to “topic clusters”. For any topic cluster with fewer than 150 cells, we assigned unassigned cells (i.e. that had less than 50% probability for all topics) from the periphery based on their proximity to the centroid of that topic cluster. We removed three topics that still had fewer than 50 cells because they would have too little coverage for peak calling. Thus, using this procedure we assigned a total of 24,503 cells to 37 topic clusters for further analysis. We report the cells-by-topics matrix in Fig. 3B, and the number of cells per topic in Table 1. We describe how we assigned cell types to these topic clusters in the ensuing sections.

Visualizing these clusters with Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018; Becht et al., 2019) shows clear separation among groups of cells (Fig. 3C). We also note that, although we focus on the 37 topics that drive our cell clusters, the remaining topics could still contain useful information. Some topics with high probability in too few cells to be used for clustering may characterize cell types that are rare or that were less-successfully isolated. In addition, topics with more diffuse probability across cells could correspond to other phenomena, for example, to different kinds of regulatory activity (e.g. promoters or enhancers (González-Blas et al., 2019)), cells with more

complex patterns of regulatory activity, cells with noisy signal, or cell types with insufficient signal to be confidently clustered.

Topics correspond to specific tissue identities.

After clustering our cells based on 37 topics, we sought to determine whether these clusters of cells represent different cell types. As with other dimensionality reduction techniques (e.g. principal component analysis), LDA is an unsupervised algorithm with no restrictions on what qualities of the data it uses to determine the topics, and interpretation of the topics can be challenging.

One way to assess whether the topics show some tissue-specificity is by cross-referencing the sci-ATAC-seq peaks with what is known about those loci in the literature, similarly to how marker genes are identified for clusters in scRNA-seq data (Cao et al., 2017; Packer et al., 2019). In the absence of broad data sets for cell-specific regulatory elements, we began by looking for overlap of the ATAC-seq peaks for each topic with the ChIP-seq peaks from cell type-specific transcription factors. For each of the 37 topics that we used to cluster the cells, we found all peaks in the peaks-by-topics matrix with probability greater than zero for that topic and overlapped them with all available ChIP-seq peaks from sites found in 40 or fewer other ChIP-seq data sets (i.e. non-high-occupancy-target, or non-HOT, sites (Kudron et al., 2018)) for three transcription factors with known cell type-specific expression patterns: HLH-1, a master regulator for body wall muscle (Krause et al., 1990); ELT-1, a master regulator for hypodermis in embryos and seam cells in L2 larvae (Page et al., 1997); and ELT-2, a transcription factor important in intestine development (Fukushige et al., 1998). We compared the number of overlaps in each topic to the number we would expect if the overlaps were random (i.e. if topics were not cell type-specific), and expressed this comparison as a \log_2 ratio between observed and random overlap counts (Fig. 4). We find topics with specific enrichment for overlaps with peaks from each transcription factor (95% confidence intervals are provided in Fig. 4). Peaks that characterize topics 37 and 40 are most enriched for overlaps with HLH-1 sites; peaks for topics 9, 10, 25, and 46 are particularly enriched for overlaps with ELT-1 sites; and peaks for topics 13, 23, 47, and 51 are most enriched for overlaps with ELT-2 sites. This analysis suggests that at least some of the topics are representing different tissues, and that the cells enriched in topics characterized by many peaks overlapping HLH-1, ELT-1, and ELT-2 sites are likely muscle, hypodermis, and intestine cells, respectively. We show the same overlap analysis for all 283 transcription factors in Figure S6.

Encouraged by the analysis of overlaps with ChIP-seq data from cell type-specific TFs, we sought to leverage the L2 sci-RNA-seq data (Cao et al., 2017) to do a more comprehensive analysis of all 37 topics. To do this, we mapped the sci-ATAC-seq peaks to the nearest downstream exon within 1200 bp, thereby accounting for peaks at alternative promoters and within introns. Assigning regulatory regions to their nearest gene is a simple and commonly used heuristic that works well in *C. elegans* (Araya et al., 2014). In contrast to mammals and other more complex animals (Pliner et al., 2018), there is little evidence that regulation by distal sites plays a prominent role in *C. elegans* (Reinke, 2013). Furthermore, the *C. elegans* genome is compact and gene-dense, meaning that most regulatory sites are found close to genes, either in intergenic sequences or in introns (Reinke, 2013). In total, we were able to assign 19,532 peaks of our 36,339 total peak calls to 17,389 genes. The number of genes we associate with accessible regions is higher than the number of genes known to be expressed in L2 worms (Boeck et al., 2016; Cao et al., 2017). There are several technical reasons for this, as well as potential biological reasons, including sites that play repressive

roles, persistent sites from earlier developmental stages, or sites that are poised for later expression.

With these caveats, we used these peak-gene assignments with their associated expression patterns to associate genes with topics and thereby infer whether or not the topics are clustering cells by tissue type. For each topic, we identified the top 250 most topic-specific peaks (Fig. S7) along with the set of genes that were associated with these peaks from our analysis above. Based on the normalized sci-RNA-seq expression values for each gene for 27 tissue types (http://genome.sfu.ca/gexplore/gexplore_search_tissues.html), we derived a tissue expression distribution for each gene. For each topic, we computed the mean tissue expression distribution for the genes associated with the top 250 most topic-specific peaks, and then calculated the \log_2 -ratio of that to the mean tissue expression distribution of 250 randomly selected genes (see Fig. S8 for an example schematic). Our results show that these topic-specific peaks are near genes with tissue-specific expression patterns and suggest that these topics reflect specific cell types (Figs. 5, S9). In fact, the genes associated with many of the topics show evidence of enrichment for tissue subtypes, including distinct kinds of neurons and hypodermis, and even some small but clearly distinct cell populations, such as sex myoblasts. At this resolution, there appears to be no distinction between body wall muscle and intestinal/rectal muscle (topic 40 encompasses both), and it is not possible in all cases to assign a subtype to topics enriched in neuronal genes (for example, in topic 38).

Overall, we find reasonable agreement between the number of nuclei we observe in our tissue clusters and the number of nuclei we expect based on the worm anatomy (Sulston and Horvitz, 1977; Altun and Hall, 2002, 2009) (Table 1). During the L2 stage the number of nuclei in each worm increases from 674 to more than 731, and our synchronized worm culture likely sampled worms over several hours of development. One of the most dynamic tissues in this time is the gonad, as the germline begins mitotic divisions in early L2. Indeed topic 24 (enriched for germline) adds nearly triple the number of nuclei to the gonad total as the other (somatic) gonad topics combined.

After identifying cell types associated with our topics, we revisited the list of peaks that had no overlap with TF ChIP-seq sites and looked for differences between those with overlaps and those without (Fig. 6). In general, most peaks contribute to primarily one topic with appreciable contributions to a few additional topics. The peaks that overlap TF ChIP-seq sites (Fig. 6A) tend to be found in many more cells in the L2 animal than the peaks without TF ChIP-seq overlaps (Fig. 6B). Nevertheless, the peaks without ChIP-seq overlaps still show clear topic-specificity, suggesting that they are informative peaks. In particular, over half of the peaks with no ChIP-seq overlaps contribute to neuron or gonad topics. Neurons are the most diverse tissue type in the worm, with scRNA-seq capable of identifying transcriptional signatures consistent with specific neuron cells (Packer et al., 2019); their absence in the ChIP-seq data suggests that whole-worm ChIP-seq lacks the sensitivity to find the cell-type-specific regulatory sites present in only a few L2 cells. On the other hand, transcription factors for gonad tissues are not well-characterized by the current modERN ChIP-seq compendium. In support of the increased sensitivity of sci-ATAC-seq over bulk whole worm assays, we find that the top 250 most-specific peaks for each topic that we used to infer the cell types of our topic clusters (Figs. 5, S7) are depleted for overlaps with TF ChIP-seq sites and bulk ATAC-seq sites (Fig. S10). Thus, we conclude that the sci-ATAC-seq peaks that do not overlap TF ChIP-seq sites are most likely either highly cell-type-specific, or specific to transcription factors that have not yet been tested with ChIP-seq.

We also investigated the sci-ATAC-seq signal at known tissue-specific genes, and we find strong tissue-specific chromatin accessibility that is consistent with the known expression patterns of

these genes (Fig. 7). The genes *hlh-1*, *pha-4* (a master regulator of pharyngeal tissue), *elt-1*, *col-160* (a collagen gene that is expressed in non-seam hypodermis in L2), *bbs-8* (a gene encoding a receptor expressed in ciliated sensory and oxygen sensory neurons), *unc-47* (a gene expressed in GABA-ergic neurons), *elt-2*, T02B11.3 (a gene that is specifically expressed in sheath glial cells), and *glh-1* (a gene expressed specifically in the germline), all show accessibility enriched in the expected tissue types. The data also suggest patterns of differential isoform expression; for example, of the three *pha-4* isoforms, the 5' end of the long isoform is most accessible in intestine cells, while the 5' end of a medium isoform has almost no accessibility, and the shortest isoform has pharyngeal accessibility. There are also several sites downstream of the *pha-4* gene that are strongly accessible in pharynx and perhaps represent other sites that play a role in regulation of this locus.

Given the isoform-specific accessibility pattern over the *pha-4* locus, we asked whether this is a more general phenomenon and searched for other genes that show similar patterns. Many genes, both with and without multiple 5' ends, are broadly expressed and show complicated patterns of chromatin accessibility in the sci-ATAC-seq data that suggest the presence of tissue-specific regulatory sites, but that can be difficult to interpret (Fig. S11, Supplementary Table 1). To maximize interpretability, we searched for genes with multiple 5' ends separated by at least 150 bp that are predominantly expressed in just two tissue types. We found dozens of examples of genes with compelling patterns of tissue-specific chromatin accessibility that suggest tissue-specific isoform usage (Supplementary Note 1). We compared the sci-ATAC-seq signal at these loci with bulk whole transcript RNA-seq data from FACS-isolated embryonic tissues (Warner et al., 2019) and, despite the differences in stages assayed, find that in many cases the tissue-specific isoform patterns in the sci-ATAC-seq data are supported by the RNA-seq data (Fig. S12).

LDA modeling of cells from individual tissue types detects fine-grained cell types.

While the topics we identified can distinguish cells at the level of tissue type, to yield more specific cell identities we tried a more focused analysis of cells from a particular cluster/tissue. We re-ran our LDA-based clustering procedure (Figs. 1, S13) for cells of each tissue type to identify subclusters that correspond to more fine-grained cell types. We grouped the topics into eight major tissue types (coelomocyte, glia, gonad, hypodermis, intestine, muscle, neuron, and pharynx), and iteratively trained LDA models for each tissue, clustered the cells, and called peaks for each cluster. Similarly to the whole worm analysis, we provide the results as a UCSC Genome Browser track hub, viewable here: [http://genome.ucsc.edu/cgi-](http://genome.ucsc.edu/cgi-bin/hgTracks?db=ce11&hubClear=http://waterston.gs.washington.edu/atacCellType/Durham_hub.txt)

[bin/hgTracks?db=ce11&hubClear=http://waterston.gs.washington.edu/atacCellType/Durham_hub.txt](http://genome.ucsc.edu/cgi-bin/hgTracks?db=ce11&hubClear=http://waterston.gs.washington.edu/atacCellType/Durham_hub.txt)

Body wall muscle cells are quite similar to each other, despite differentiating from four different embryonic lineages. In previous single cell RNA-seq studies, the body wall muscle cells all clustered together, without much separation (Cao et al., 2017; Packer et al., 2019). Nevertheless, within the body wall muscle cluster the cells were found to group by anatomical position, setting up an anterior-posterior axis through the cluster that was identified by looking for the expression of specific marker genes. Similarly, the intestine cells showed evidence of expression differences in sci-RNA-seq by anatomical position. We looked for a similar anatomical pattern in the sci-ATAC-seq muscle and intestinal cell clusters (Figs. 8, S14, S15). We found cells with peaks associated with marker genes that are expressed throughout muscle (*hlh-1* and *myo-3*) and intestine (*end-1* and *elt-2*) distributed throughout the clusters (Fig. 8A), whereas those cells with peaks associated with marker genes with anatomically-biased expression showed distinct patterns (Fig. 8B). The subclustering thus appears to reveal still finer distinctions between cells.

Next, we performed a similar analysis on all cells from neuron-enriched clusters. While the tissue type categories based on the L2 sci-RNA-seq data (Cao et al., 2017) already break gene expression distributions down into nine different neuronal subtypes, there are many more specific neuron types revealed by anatomical analysis and by additional scRNA-seq data sets (Packer et al., 2019; Taylor et al., 2019). (Fig. 5). To take a closer look at the neuron cells, we gathered all cells from topic clusters showing neuronal enrichment and performed the same analysis that we did for body wall muscle and intestine above. The neuron LDA model yielded 36 topic clusters – only one fewer than the number of clusters found by the whole worm model, highlighting the diversity of neuronal cell types. We evaluated five marker genes that we chose for their tissue specific expression patterns in the sci-RNA-seq data: *bbs-8* is expressed in ciliated sensory neurons; *gcy-32* is expressed in oxygen sensory neurons; *unc-30* is expressed in GABA-ergic neurons; *mec-7* is expressed in touch sensitive neurons; and *ceh-24* is expressed in cholinergic neurons. We find that the cells with chromatin accessibility near these genes are associated with distinct, well-separated clusters in UMAP space (Fig. 9A), suggesting cell type identities for these clusters.

To verify the neuron subtypes identified by single marker genes and also attempt to identify additional subtypes for the other clusters of cells, we followed up by checking additional marker genes for specific neurons (Fig. 9B). The ASE neurons are a pair of ciliated chemosensory neurons that detect water-soluble attractants like potassium and sodium ions. They express *bbs-8*, along with a highly specific repertoire of guanylyl cyclase (*gcy*) genes that encode sensory receptors.

We found chromatin accessibility peaks near six such receptor genes specific to ASE neurons: *gcy-4*, *gcy-5*, *gcy-6*, *gcy-7*, *gcy-14*, and *gcy-22* (Etchberger et al., 2009). To identify the ASE neurons in the UMAP, we colored cells by the number of co-accessible peaks they have that are near the ASE-specific *gcy* genes, which identifies one of the *bbs-8* subclusters as likely to be ASE neurons. We performed a similar analysis to confirm the UMAP cluster containing the AQR, PQR, and URX oxygen-sensory neurons. Oxygen-sensory neurons express their own repertoire of *gcy* genes, including *gcy-32*, *gcy-35*, *gcy-36*, and *gcy-37* (Packer et al., 2019), and peaks near these genes are co-accessible primarily in the cells marked by *gcy-32* in Fig. 9A. We identified additional examples of neuron subtypes, including mechanosensory neurons, two types of motor neurons, and interneurons (Fig. S16). We show the neuron UMAP colored by LDA topic probabilities in Supplementary Figures S17, S18, S19, and S20. Finally, we also subclustered coelomocyte (Figs. S21, S22), glia (Figs. S23, S24), gonad (Figs. S25, S26), hypodermis (Figs. S27, S28), and pharynx (Figs. S29, S30), finding evidence for additional specific cell types in each of them. Thus, sci-ATAC-seq can make fine-grained distinctions among cell types at a resolution that approaches scRNA-seq and thereby begin to associate specific regions of accessible chromatin with specific cell types.

Discussion

We used the sci-ATAC-seq assay to assemble the first cell type-resolved map of regulatory elements in *C. elegans*. We found 38,017 peaks, which we used to assign 24,503 of our 30,930 cells to one of 37 different clusters (Fig. 3) that represent distinct, differentiated tissues in the L2 nematode (Fig. 5). Our map, derived from data collected in essentially a single experiment, recovers the majority of L2 regulatory sites detected by hundreds of individual ChIP-seq experiments (Fig. 2) and proposes an additional 7,131 novel regulatory sites not found in the ChIP-seq compendium (Fig. 6). Accessibility at these sites can distinguish among nearly all the major cell types in the worm (Figs. 5, S9) and many minor ones, including highly similar cell types, such as muscle or intestine cells at

different anatomical positions (Fig. 8), and cell types that have only a few examples per worm, especially neuron subtypes (Figs. 9, S16). For example, there are only two ASE neurons in each worm, and our analysis suggests that we can identify one or two clusters that might harbor these cells (Fig. 9B). For accessibility peaks that overlap TF ChIP-seq sites (Kudron et al., 2018), they suggest the tissue in which the TFs are active, and in turn the ChIP sites suggest which TFs are active at the sci-ATAC-seq peak (Figs. 6, S6). We found dozens of examples where non-overlapping alternative 5' exons are likely regulated by distinct tissue-specific accessibility sites, supporting the notion that a principal reason for alternative 5' ends is for tissue-specific regulation (Supplementary Note 1, Fig. S12). In addition, there are many other broadly expressed genes, both with and without multiple isoforms, that also show tissue-specific accessibility patterns (Fig. S11, Supplementary Table 1).

The cell type resolution of sci-ATAC-seq data approaches that of scRNA-seq (Cao et al., 2017; Packer et al., 2019), despite the high levels of sparsity and low dynamic range of the data in individual cells. Single cell ATAC-seq suffers generally from sparsity and a low dynamic range because there are at most two chances in a diploid cell to sample a given accessible locus, depending on whether or not both alleles are accessible. This contrasts to single cell RNA-seq, which has hundreds or even thousands of chances for measuring the mRNA of highly expressed genes. The sparsity of scATAC-seq was particularly high in our data set because the worm samples yielded sci-ATAC-seq libraries with many fewer unique fragments per cell than other organisms. In the first reported sci-ATAC-seq results, human and mouse cell lines yielded a median of 2,503 fragments per cell (Cusanovich et al., 2015), and in more recent work on fly and mouse cells the median yield was over 10,000 fragments per cell (Cusanovich et al., 2018a,b). In contrast, despite using the most up-to-date protocol (Cusanovich et al., 2018a,b), the median *C. elegans* cell yielded only about 700 fragments (Fig. S1). We hypothesize that this could be improved by optimizing the nuclear isolation and permeabilization conditions — L2 nuclei are extremely small, compact, and dense (about 2 μm in diameter), and possibly after formaldehyde fixation Tn5 has only limited access to the chromatin. As single cell technology advances and we are able to generate more complex libraries we expect that scATAC-seq will provide even higher resolution of individual cell types in the worm.

We also expect to see improved results from applying more advanced computational techniques. We provide a new LDA implementation that implements several useful features (see Methods), and also look forward to new approaches that more tightly integrate the analysis of scATAC-seq and scRNA-seq data sets. There are multiple recent approaches for projecting single cell data from different modalities into the same embedding space (Cusanovich et al., 2018a; Welch et al., 2017; Stuart et al., 2019). By jointly analyzing the sci-ATAC-seq data and sci-RNA-seq data with one of these methods it may be possible to improve the cell type resolution of our chromatin accessibility maps.

Such accessibility maps with high cell type resolution will be important for understanding gene regulation on the scale of the whole genome across the whole organism. In addition, regulatory sites are hypothesized to play a major role in common disease and evolutionary adaptation, so maps of regulatory sites will aid in interpreting the effects of genetic variation. For example, many mutations that are linked to some phenotype by approaches like GWAS do not fall in genes. The implication is that, if one of the mutations is indeed causal, it must fall in a regulatory sequence of DNA. Thus, maps of cell type-specific regulatory regions can help interpret and prioritize candidate causal variants, and will be a useful complement to genetic resources in *C. elegans*, including the *C.*

C. elegans Natural Diversity Resource (Cook et al., 2017) and the Million Mutations Project strains (Thompson et al., 2013). Better annotations of regulatory DNA can also help with understanding comparative genomics and the evolution or conservation of stretches of non-coding DNA.

Given the importance of mapping regulatory sites for understanding genome structure and function, and the power of *C. elegans* as a model organism, improving and expanding our maps of regulatory regions should be a high priority. As single cell genomics technology continues to improve, we will be able to attain better measurements across more cells and at lower cost. Co-assays that can make chromatin accessibility and RNA-seq measurements in the same cells are showing very promising results, with the library preparation cost per cell recently dropping by orders of magnitude (Cao et al., 2018; Ma et al., 2020). Leveraging such approaches, collecting accessibility data for additional developmental stages in worm will provide valuable insight into the dynamics of gene regulation over the course of development as cells differentiate. These data can be paired with new scRNA-seq data collected from throughout *C. elegans* embryogenesis (Packer et al., 2019), moving the field closer to having a truly comprehensive map of gene expression and regulation for every cell throughout development in *C. elegans*.

Methods

Nuclear isolation from whole L2 worms

We grew wild-type *Caenorhabditis elegans* worms (VC2010 strain) at 21°C on nine 150 mm plates and synchronized the population by bleaching (2% bleach, 0.5 M KOH) young adults with 8-12 embryos to isolate embryos, hatching them at room temperature in egg buffer (118 mM NaCl, 48 mM KCl, 2 mM CaCl₂, 2 mM MgCl₂, HEPES 25 mM at pH 7.3) for 12-16 hours, and re-plating the L1 hatchlings onto nine more 150 mm plates at a density of approximately 60,000 worms per plate. After two rounds of this bleach synchronization and plating, the L1 worms were allowed to grow at 21°C for 19 hours after plating to reach the middle of the L2 stage. The worms were washed off eight of the plates with M9 buffer (22 mM KH₂PO₄, 22 mM Na₂HPO₄, 85 mM NaCl, 1 mM MgSO₄ at pH 6.5) into a 50 ml conical tube. Bacteria were removed from the suspension by spinning the tube at ~3,000 × g, aspirating the supernatant, resuspending in fresh M9. The M9 wash was repeated, and the supernatant was aspirated, leaving a worm pellet in ~1 ml of M9. The worm pellet was flash frozen by using a P1000 to transfer the worms drop by drop into a mortar containing liquid nitrogen. The frozen worms were crushed into powder with a pestle such that each worm broke into 3-4 chunks, and the powder was transferred to a 50 ml falcon tube containing 8.75 ml of 1.1% formaldehyde in egg buffer supplemented with 1x protease inhibitor. Worms were rocked at room temperature for 10 min before the fixation reaction was quenched by adding 1.25 ml 1M glycine (final concentration ~125 mM) and incubated another 5 min at room temperature. The fixed worms were pelleted at 3220 × g for 5 min at 4°C, the supernatant was removed, and the pellet was resuspended in 10ml ice cold egg buffer. Fixed worms were pelleted again by spinning at 3220 × g for 5 min at 4°C. The egg buffer supernatant was aspirated, and the pellet was resuspended in ice cold 2x nuclear preparation buffer (20 mM HEPES pH 7.6, 20 mM KCl, 3 mM MgCl₂, 2 mM EGTA, 0.5 M sucrose, 0.05% Triton X-100 in egg buffer) supplemented with protease inhibitor (NPB+PI). The following steps were all performed at 4°C or on ice: The solution was transferred to a 7 ml Dounce homogenizer and the fixed worm chunks were homogenized with 20 loose pestle strokes followed by 10 tight pestle strokes. The Dounced suspension was spun for 90 seconds at ~200 × g in a swing-

arm centrifuge to loosely pellet debris, and the top 1000 μ l of supernatant (containing the nuclei) was removed to a 15 ml falcon tube on ice. 1 ml of fresh NPB+PI was added to the Dounce, the debris pellet was gently resuspended, and the Douncing and spinning were repeated three more times, resulting in the collection of 4 ml of nuclei. The suspension of nuclei was cleaned by gently passing through a 10 μ m syringe filter pre-wetted and chased with 1 ml ice cold NPB+PI into a new 15 ml falcon on ice. The nuclei were split evenly into 1.5 ml eppendorf tubes and pelleted at 2000 \times g for 10 min at 4°C. All supernatant was removed, and the pellets were each gently resuspended in 1 ml freezing solution (50 mM Tris at pH 8.0, 25% glycerol, 5 mM Mg(OAc)₂, 0.1 mM EDTA, 5 mM DTT, 1x protease inhibitor cocktail (Roche), 1:2,500 superasin (Ambion)) (Cusanovich et al., 2018b). The resuspended nuclei were transferred to 2 ml cryotubes, flash frozen in liquid nitrogen, and stored at -80°C.

Single cell ATAC-seq via single cell combinatorial indexing

The sci-ATAC-seq protocol was as described Cusanovich, et al. 2018 (Cusanovich et al., 2018b). Briefly, flash-frozen VC2010 nuclei were thawed in a 37°C water bath and put immediately on ice. The nuclei were transferred to a 1.5 ml eppendorf tube and spun at 2000 \times g for 10 min. The supernatant was aspirated, and the pellet was resuspended in 200 μ l of ATAC-OMNI (Corces et al., 2017) RSB (10 mM Tris-HCl pH 7.4, 10 mM NaCl, and 3 mM MgCl₂ in water) supplemented with 0.01% Digitonin, 0.1% IGEPAL-630, and 0.1% Tween-20, allowed to stand for 3 min on ice, and then quenched by adding 1 ml of RSB supplemented with 0.1% Tween-20. The resuspended and lysed nuclei were stained with 1x Hoechst and a BD FACS Aria II was used to distribute 2500 nuclei into each well of a 96-well v-bottom plate (Eppendorf twin.tec LoBind skirted 96 well PCR plate) prepared with 19 μ l of tagmentation reaction solution (10 μ l 2x Nextera TD buffer, 3.3 μ l 1X DPBS, 0.2 μ l 1% Digitonin, 0.2 μ l 10% Tween-20, 5.3 μ l H₂O) (Corces et al., 2017). After sorting, 1.0 μ l of 2.5 μ M uniquely-barcoded Tn5 from Illumina (Cusanovich et al., 2015) was pipetted into each well of the 96 well plate, and the transposition reaction was allowed to proceed at 55°C for 30 minutes. Next, 20 μ l of STOP reaction buffer (40 mM EDTA and 1 mM Spermidine) was added to quench the reaction, and the plate was put at 37°C for 15 min. After stopping transposition, all nuclei were pooled into a 15 ml conical tube, re-stained with 1x Hoechst, and distributed by FACS into twenty eight 96-well v-bottom plates at 25 nuclei per well. The 96-well plates contained 12 μ l per well of reverse cross-linking buffer (0.83 mg/ml Proteinase K and 0.042% SDS in Qiagen EB buffer), and were put on ice, spun down, and frozen at -20°C in batches during sorting. Later, these plates were thawed in groups of four for reversing crosslinks by incubating at 65°C for 16 hours, after which the transposed and un-crosslinked fragments were amplified using uniquely-barcoded PCR primers. We ran four wells as test reactions in qPCR and monitored the libraries for saturation of SYBR-Green signal to identify the number of cycles required for appropriate amplification (Cusanovich et al., 2018b), and then amplified the rest of the wells for either 22 cycles with Illumina NPM 2x PCR master mix or 23 cycles with NEBNext 2x PCR master mix (PCR Reaction: 12.0 μ l of nuclei in reverse crosslinking buffer; 2.5 μ l of 5 μ M Nextera v2 barcoded P7 PCR primer; 2.5 μ l of 5 μ M Nextera v2 barcoded P5 PCR primer; 1.0 μ l of 100X BSA; 25.0 μ l of 2x NEBNext PCR Master Mix (NEB cat M0541); and 7.0 μ l nuclease free H₂O) (NPM PCR protocol: 72°C for 3:00, 98°C for 0:30, repeat \times 22(98°C for 0:10, 63°C for 0:30, 72°C for 1:00), 4°C HOLD; NEBNext PCR protocol: 72°C for 5:00, 98°C for 0:30, repeat \times 23(98°C for 0:10, 63°C for 0:30, 72°C for 1:00), 4°C HOLD). After amplification, the fragments were cleaned up by pooling the contents of all wells and splitting

across four Zymo Clean and Concentrate columns (cat. D4014), eluted each in 25 μ l Qiagen EB, combined the eluates, and then further cleaned and concentrated with 1x Ampure XP magnetic beads, and finally eluted in 25 μ l. Library quality was assessed using the Agilent TapeStation D5000 kit (Screentape cat. 5067-5588, and reagents cat. 5067-5589), and molarity was quantified for fragments between 200 and 1000 bp. Last, the libraries were combined into an equimolar pool at 2 nM for sequencing. Libraries were sequenced using the manufacturer's denaturation conditions, and loaded either on a Illumina MiSeq 300 cycle v2 kit (cat. MS-102-2002) at an input concentration of 15 pM, or on an Illumina NextSeq Mid-Output 300 cycle v2.5 kit (cat. 20024905) at an input concentration of 1.8 pM, using custom sequencing primers and recipe from Illumina.

Generation of genomic DNA input control

To control for the sequence cutting bias of Tn5 (Green et al., 2012), we treated naked *C. elegans* genomic DNA with the bulk ATAC-seq protocol (Buenrostro et al., 2013) as follows. We isolated genomic DNA with phenol:chloroform extraction and ethanol precipitation. To keep the Tn5:DNA ratio similar to a bulk ATAC-seq experiment with 50,000 cells, we estimated that a typical *C. elegans* nucleus will contain 1×10^6 bp \times 2 genomes \times 660 MW/bp \times 1.67×10^{-12} pg/MW \approx 0.22 pg/nucleus, or \sim 11 ng in 50,000 nuclei. We diluted the DNA to a concentration of \sim 0.87 ng/ μ l, as measured with the Qubit High Sensitivity assay (Invitrogen), and used 11.5 μ l as input to a 25 μ l reaction with 12.5 μ l of Nextera TD buffer and 1.0 μ l of Nextera Tn5 enzyme. The reaction was incubated at 37°C for 30 min, cleaned up with a Qiagen MinElute column, and amplified using NEBNext 2x PCR Mix with primers from Buenrostro, et al. 2013 (Buenrostro et al., 2013). The libraries were cleaned up with 1:1 AMPure XP magnetic beads and sequenced on the Illumina NextSeq platform.

ATAC-seq alignment pipeline

Initial processing of the sequencing results was done as reported in Cusanovich, et al. 2018 (Cusanovich et al., 2018b), with some changes. Sequencing results were converted to FASTQ format with the Illumina bcl2fastq program (v. 2.19). First, the integrity of the barcode sequences was checked for each of the four components of the barcode (tagmentation barcodes from both sides of the cut and the P5 and P7 primer indices added during PCR amplification) by matching the sequencing results to the known barcode sequences. Any read that had three or fewer edits compared to the best-matching known barcode sequence and that had no other known barcode sequences matching with five or fewer edits were corrected and assigned to the best-matching barcode sequence. Any read-through of short templates was corrected by trimming adapter sequences from reads using Trimmomatic (Bolger et al., 2014) (v0.36) with the options ILLUMINACLIP:NexteraPE-PE:2:30:10:1:true, TRAILING:3, SLIDINGWINDOW:4:10, and MINLEN:20. Next, read sequences were aligned to the WS235/ce11 build of the *C. elegans* genome with Bowtie 2 (Langmead and Salzberg, 2012) with options -X 2000 and -3 1, properly paired reads with mapping scores greater than 10 were kept, any reads mapping to the mitochondrial DNA were filtered out, and read pairs with identical barcode sequence and identical starting and ending mapping coordinates were identified as PCR duplicates and collapsed to 1 using a custom script (Cusanovich et al., 2018b). Next, read coverage for each cell was calculated, and cell barcodes with fewer than 150 reads were removed from further analysis (Fig. S1). The reads that made it through filtering for each batch of sequencing were merged into a single BAM file using the Picard MergeSamFiles program (<http://broadinstitute.github.io/picard>). Last, the reads in the merged BAM file were

converted into cut sites by taking 60 bp intervals centered on the fragment ends, shifting those sites for reads mapping to the forward strand by +4 bp and the negative strand by -5 bp (to account for the shape of the Tn5 cut site (Buenrostro et al., 2013)), and writing the resulting coordinates to a BED file for peak calling.

We called peaks from the cut site data using MACS2 (Zhang et al., 2008) (v. 2.2.5) with options `--format=BED, -g 9e7, --nomodel, --qvalue=0.05, --SPMR, --tsize=60, --bdg, --keep-dup all, and --call-summits`. Additionally, we aligned and mapped cut sites for some bulk ATAC-seq data collected on naked *C. elegans* genomic DNA and provided this as an input control to correct our peak calls for sequence bias in Tn5 cutting. After peak calling, we merged any overlapping peaks to produce a single set of non-overlapping genome-wide peak calls. Finally, we generated a binary cells-by-peaks matrix that records which peaks were detected (i.e. were overlapped by a cut site) in each cell. This data structure was used for further cell clustering analysis with latent Dirichlet analysis (LDA), described below.

Cell clustering pipeline overview

To identify cell types from the sci-ATAC-seq data, we used an iterative clustering and peak calling approach. After investigating dimensional reduction methods including PCA and LSI, we found LDA provided the best separation between clusters of cells. The core pipeline consisted of three main steps: first, training a latent Dirichlet allocation model (LDA, see below) on the data; second, identifying LDA topics that corresponded to coherent groups of cells and using those topics to cluster the cells; and finally, calling peaks on pooled data for each cluster of cells. Calling peaks on the cell clusters increases sensitivity to detect cluster-specific peaks compared to the peak calling at the end of the alignment pipeline; cell type-specific regulatory sites may be weaker than more commonly-accessible peaks either because they are smaller or more transiently accessible, or because they may be specific to a small subset of the cells in the whole worm. Either way, the data for such sites might not rise above the background noise from other cell types in the whole worm data set, but are detectable in more homogeneous subsets of cells. We iterated this procedure twice: first, to generate a more sensitive set of peaks than the peak calls from the alignment pipeline (we call this first iteration the primary LDA), and a second time to train a new LDA model with the improved peak set, which also results in a third, refined peak set (we will refer to the second iteration as the refinement LDA). Last, MACS2 occasionally calls broad peaks that cover large regions of the genome of hundreds or even thousands of base pairs. The signal over these peaks is usually multimodal with multiple distinct “summits” that most likely represent distinct binding sites. To better capture the distinct nature of these summit regions in the refined peak set, we implemented a custom script that identifies the local summits within a MACS2 peak and splits the peak into multiple contiguous segments that each encompass a single summit region, and we report these split peaks in our UCSC Genome Browser track hubs. (Note that our peak splitting procedure is similar to, but distinct from, the MACS2 summit-calling procedure, which reports the coordinate of the base pair with the highest signal and does not actually segment the peak around the summits.) For the peak-splitting code, see the `expand_summits2.py` script in the paper’s GitHub repository: https://github.com/tdurham86/L2_sci-ATAC-seq.

Latent Dirichlet allocation implementation

Inspired by the effectiveness of LDA as implemented in cisTopic (González-Blas et al., 2019), we decided to take this approach to analyzing the sci-ATAC-seq data. Briefly, LDA is a Bayesian modeling strategy that was originally developed in the setting of document classification. It assumes that each document is characterized by one or more latent “topics”, and that these topics are characterized by subsets of the words in the document. Documents are modeled as Dirichlet probability distributions over the set of topics, and the topics are modeled as Dirichlet probability distributions over the set of words. Training proceeds by iterating over the entire vocabulary defined by the documents it is modeling and proposing a topic for every instance of every word in every document. The probability of picking a topic for a given word and document is computed based on the current probability distribution of topics for that document and the probability of words for each topic. At the end of training, the probability distributions for the topics over the documents and the words over the topics can be calculated by summing the topic proposals for all peaks, and for all documents, respectively. Our implementation uses a collapsed Gibbs sampler to speed up training by sampling the latent parameters of the model from the full conditional posterior (Griffiths and Steyvers, 2004). When applied to single cell ATAC-seq data, as in cisTopic, cells are treated as the documents and peaks are treated as the words. The LDA model then learns topics that distinguish among the cells based on which peaks tend to be accessible in similar patterns across all cells, and outputs two matrices that capture the relationship between peaks and topics, and cells and topics: the first matrix contains the counts of the number of cells for which a given peak was assigned to each topic (the peaks-by-topics matrix), and the second matrix contains how many peaks from each cell were assigned to each topic (the cells-by-topics matrix).

We began by using cisTopic itself, but found the R implementation to take almost two days to process the full data set. (Note that a new recently released version of cisTopic is significantly faster and incorporates similar ideas to our implementation.) To speed up the modeling, we implemented a parallelized version in Java that can split the training of a single model across multiple cores, reducing the run time to just a couple of hours. In the end, for the whole worm primary LDA we used 34 topics and set the alpha parameter for the Dirichlet priors to 3.0 to concentrate the probability distributions into just a few peaks/topics. We set the beta parameter to 2000.0; the higher value allows the LDA model to spread the probability across more peaks. These alpha and beta values control the weight of the symmetric, or uniform, priors used for the cell-by-topic and topic-by-peak distributions – the higher the alpha or beta value, the more weight is given to the uniform prior, which encourages the model to spread probability across topics and peaks instead of concentrating it. The Java code can be accessed on GitHub at <https://github.com/gevirl/LDA> and is released under the GNU GPLv3 license.

Filtering the cells-by-peaks matrix for LDA

To improve the efficiency and effectiveness of LDA, it can help to remove cells that have very few peaks and to remove peaks that are either found in very few cells or too many cells. Sparse cells and peaks provide little useful information and are enriched for noise, and while peaks that are found in too many cells are likely real, they are not very helpful for distinguishing among cell types. We filtered the outliers by sorting the cells by the fraction of all peaks that was detected in each cell (or peaks by the fraction of all cells in which they were detected), mean centering the data, and then identifying the change points at the extremes of the resulting curve by doing a convolution with

values from a sigmoid function to detect the areas where the slope is changing fastest. We set the filtering thresholds as four times the inter-quartile range above the mean of the convolution output (Figs. S31, S32). We applied LDA to the resulting filtered cells-by-peaks matrix.

Latent Dirichlet allocation model selection

Choosing hyperparameter values is one of the most challenging aspects of training models like LDA. In particular, using an appropriate number of topics is critical to getting good results, and picking this number requires an empirical approach. In *cisTopic* (González-Blas et al., 2019), the authors recommend training several model instances, each with a different number of topics, and choosing the number of topics that gives the best log likelihood of your input data. However, increasing the number of topics adds parameters to the model, which makes the model better able to fit the training data, even if it has already fit the true signal and begins to train on noise (i.e. it is overfitting). Since the *cisTopic* procedure uses the same data for training and evaluation, it does not test the generalizability of the model parameters, and cannot tell when the model starts to overfit. Ultimately, it will recommend using a higher number of topics than can be supported by the data. To identify a suitable number of topics that avoids overfitting, we implemented the following cross-validation procedure.

First, the cells are evenly and randomly split into five disjoint sets for five-fold cross validation. Then, for each number of topics that we would like to test, five LDA models are trained, with each model training on four of the folds and holding one out for evaluation. Once each model is done training, it estimates the likelihood of the data in the held-out test fold with a Chib-style estimator (Wallach et al., 2009). In this estimation procedure, the peak-topic probabilities learned from the training data are fixed, and then a cell-topic vector is trained for each held out cell based on the fixed peak-topic probabilities. The log likelihood of each held out cell is estimated based on sampling from the posterior of the model trained on that held out cell. We convert these log likelihoods to perplexity, which is defined as follows:

$$\text{perplexity}(w) = \exp\left(-\frac{\mathcal{L}(w|\theta)}{N}\right)$$

where w is a held out test cell, $\mathcal{L}(w|\theta)$ is the log likelihood of that test cell given the LDA model, and N is the number of peaks found in that cell. Because perplexity is inversely related to the log likelihood, smaller values are better. The best number of topics to use is the one that produces the lowest mean perplexity from the five held out sets of test data. It is important to note that LDA is a stochastic modeling technique, and training on the same data with different random seeds will yield similar but different solutions. In addition, our *sci-ATAC-seq* data are by nature noisy and complex. Thus, the hyperparameter search procedure will not always result in a clear best number of topics to pick. We found that if we trained a model with a few extra topics beyond the optimal number, LDA would largely ignore the extra topics and still put almost all of the probability in a number of topics that approximated the underlying dimensionality of the data. Given that the model appeared robust to some extra topics, we ran our models with 1.5 times the number of topics recommended by the hyperparameter search (Figs. S4, S13).

Latent Dirichlet allocation training

To train the LDA model, we used our parallelized implementation of LDA, which could generally run on the full data set in about two hours on a machine with 8 cores and 32 GB of memory. Here is an example command line from the whole-worm refinement LDA analysis (see the GitHub repository for full documentation and usage information): `java -Xms32G -cp LatentDirichletAllocation.jar org.rhwlab.lda.cache.matrix.LDA_CommandLine -lda -a 3.0 -b 2000.0 scatac_data.bow -li 4000 -o ./out/dir -s 1 -t 55 -th 8 -tn 5 -ch 0 -rid 0000 -pe -d topic -st mode -sk 40 -v 1 -id ./out/dir/0000_topics55_alpha3.000_beta2000.000 -pr 1.0`

The model has four main outputs: the docTopic matrix, which contains the raw counts of how many peaks were assigned to each topic in each cell; the theta matrix, which contains the probability distribution across topics for each cell and takes into account the full LDA probability, including the prior; the wordTopic matrix, which contains the raw counts of how many times each peak was assigned to each topic across all cells; and the phi matrix, which contains the probability distribution across peaks for each topic, and takes into account the prior. We used the theta matrix to cluster the cells, and the wordTopic matrix for identifying the cell type for each cluster.

Cell clustering by topics

Next, we sought to cluster the cells based on the LDA modeling results. We reasoned that differences among cell types would be the dominant source of informative variation in our sci-ATAC-seq data, and that, for this reason, many topics should mostly correspond to distinct cell types. To identify which topics were most likely to distinguish among cell types, we looked for topics that had high probability in subsets of cells that were close together in LDA topic-space and used these topics to define cell clusters as follows. To identify topics that corresponded to groups of cells in LDA topic-space, for each topic, we ranked the cells by their probability for that topic in the theta matrix, computed the centroid of the topic as the mean topic probability vector for the top 50 cells, and then scored the topic by calculating the average similarity of those top 50 cells to the centroid by averaging the dot products of the topic probability vectors and the mean topic probability vector. Then, we ranked the topics by this centroid similarity and identified a threshold of 0.2 to separate the candidate cell type topics from the others (Fig. S5).

Next, we assigned cells to clusters defined by these topics. Any cell with greater than 50% probability in one of the cell type topics was automatically assigned to that topic cluster. Some topic clusters had many high probability cells, while some had few. For any clusters with fewer than 150 cells, we attempted to add nearby unassigned cells based on their distance from the cluster's centroid in a 10-dimensional UMAP space (McInnes et al., 2018). We used an iterative procedure for each small topic cluster as follows: First, we computed the cluster centroid in UMAP space by averaging 200 samples from a Gaussian kernel density estimate of the shape of the cluster. Next, we used a KDTree to identify a set of nearest neighbor cells to the centroid that was 25% larger than the cluster size. We detected whether any of these nearest neighbors were distance outliers (and thus more likely from a different cluster) by ranking the neighbors by distance from the centroid and convoluting the resulting distances with a step function to detect regions where the slope of the ranked distances increases. We added to the topic cluster the unassigned cell closest to the centroid as long as its value from the convolution was not greater than 1.5 times the interquartile range of all convolution values. Then we iterated this procedure, growing the small topic clusters one unassigned cell at a time. After this procedure, any topic clusters that still had fewer than 50

associated cells were removed from consideration as clusters and their cells were not assigned to any cluster.

To visualize the resulting topic clusters, we used UMAP (McInnes et al., 2018) to reduce the dimensions of the theta matrix to 2 (Fig. 3C). We first row-normalized the theta matrix with the L2-norm and then used the Python implementation of UMAP (`umap-learn`, v. 0.3.8) with default parameters. Finally, we plotted the cells as a scatter plot based on their coordinates in 2D UMAP space, and colored the cells by their topic cluster assignments.

Calling peaks for each topic cluster

The final step in the clustering pipeline is to call peaks for each cluster. For each topic cluster, we pooled the cut sites from the cells in that cluster, and used the pooled data as input to MACS2 (Zhang et al., 2008) (v. 2.2.5) with the same settings as in the alignment pipeline, including providing the bulk ATAC-seq data from naked genomic DNA as an input control. After calling peaks for each cluster, we merged the peak calls from all clusters using `bedtools merge` to make a master list of peak regions, and then used this master peak list to create a new cells-by-peaks matrix. We used our new cells-by-peaks matrix as input to a second round of LDA and cell clustering to refine our clusters and peak calls.

Overlapping peaks with other data sets

Supplementary data for Figure 2 from Jänes, et al. 2018 was downloaded from the eLife website (filename: `janes2018_fig2_data1_v2.txt`). This file was parsed using Unix tools and BEDTools (v. 2.25.0) (Quinlan and Hall, 2010) into three files: a file containing all of the peaks in BED format with overlapping sites merged (i.e. using `bedtools merge` with default parameters), a file containing promoter-annotated peaks (those annotated as “`coding_promoter`”, “`unassigned_promoter`”, or “`pseudogene_promoter`”) with overlapping peaks merged, and a file containing enhancer-associated peaks (those annotated as “`putative_enhancer`”) with overlapping peaks merged. These files were overlapped with the sci-ATAC-seq peaks, and the sci-ATAC-seq peaks were overlapped with these files, using `bedtools intersect`. The significance of the extent of the overlapping peaks was calculated using the Fisher’s exact test implementation in the `bedtools fisher` command.

Peak loci from the modERN project were downloaded from the EPIC website (<http://epic.gs.washington.edu/modERN/>) for reference WS245/ce11 using the “Download Aggregated Peaks” and “Download Clustered Peaks” buttons on the “Worm By Life Stage” tab of the user interface. Peaks were parsed into different files based on developmental stage, and any overlapping peak regions were merged in the final files. As above, these files were overlapped with the sci-ATAC-seq peaks, and the sci-ATAC-seq peaks were overlapped with these files using `bedtools intersect`. The significance of the extent of the overlapping peaks was calculated using the Fisher’s exact test implementation in the `bedtools fisher` command.

To test for enrichment of overlaps from bulk ATAC-seq sites or TF ChIP-seq sites from different developmental stages, we split the sites from each compendium into separate bed files by developmental stage, from embryo to young adult. For the TF ChIP-seq data, this split was accomplished using the developmental stage annotation provided in the downloaded peaks. For the bulk ATAC-seq peaks, we first z-score normalized the signal heights for each peak across life stages provided in the Figure 1 supplementary data from (Jänes et al., 2018), and then assigned each peak to any developmental stage with a z-score ≥ 1.5 for that peak and wrote the assigned peaks to the

corresponding bed files. Next, we counted the number of overlaps between the developmental stage beds and our sci-ATAC-seq peaks, then computed the \log_2 ratio of the number of overlaps observed versus the number of overlaps expected from 100 randomly-drawn samples of sites (Figs. 2C, S2).

We also downloaded ChIP-seq data for the L2 developmental stage from (Jänes et al., 2018) (GEO Accession: GSE114440) and analyzed the ChIP-seq signal over sci-ATAC-seq peaks as follows. First, we used the UCSC liftOver tool to map our peak coordinates to ce10 to match the reference of the ChIP-seq data, then generated a set of randomly shuffled peak locations using the bedtools shuffle command with options -chrom and -noOverlapping. Next, for both the true and shuffled peak sets, we computed the \log_2 ratio of the mean signal over each peak to the mean signal over the entire chromosome for each ChIP-seq data set, and then we compared the resulting distributions using split violin plots (Fig. S3).

To gain further insight into our sci-ATAC-seq peaks that did not have overlaps with either TF ChIP-seq sites or with bulk ATAC-seq sites, we reasoned that if our non-overlapping peaks are the result of the increased sensitivity of sci-ATAC-seq for tissue-specific accessibility, then our lists of the top 250 most topic-specific peaks (Fig. S7) should be depleted for overlaps with TF ChIP-seq sites and bulk ATAC-seq sites. For each topic, we calculated the fraction of the 250 most topic-specific peaks with an overlap and compared that fraction to the overall fraction of sci-ATAC-seq peaks with overlaps in the TF ChIP-seq sites and bulk ATAC-seq sites using the \log_2 ratio (Fig. S10). We find that, for most topics, the top topic-specific peaks are depleted for overlaps compared to the full list of peaks. The neuron-specific topics are particularly depleted for overlaps with bulk ATAC-seq sites, suggesting that whole worm bulk ATAC-seq has reduced sensitivity for detecting neuron-specific peaks (Fig. S10B). In addition, other important technical differences in the data collection methods could also contribute to differences in the regulatory sites detected. For example, in the case of the ChIP-seq data, the recovered sites are biased for the particular set of TFs that are assayed; in contrast, sci-ATAC-seq is relatively unbiased, but is subject to the Tn5 sequence bias (Green et al., 2012).

Cell-by-topic and peak-by-topic heatmaps

To generate the heatmaps as in Fig. 3B and Fig. 6, we selected only the columns corresponding to the topics used for clustering the cells. Next, we row-normalized the counts by dividing each row by its sum. Then we hierarchically clustered the rows and columns in Python (v. 3.6.10) with the SciPy module (v. 1.4.1). We used the `scipy.spatial.distance.pdist` function with the cosine metric to compute the pairwise distance matrix, then `scipy.cluster.hierarchy.linkage` with method average to generate clusters, and finally `scipy.cluster.hierarchy.dendrogram` to order the rows and columns after clustering. The clusters were classified by tissue type based on the dominant tissue type for each topic in Fig. 5. The cell/peak coverage information was calculated as the \log_2 -transformed sum of peaks found per cell for the cell-by-topic matrix or the \log_2 -transformed sum of cells exhibiting a peak for the peak-by-topic matrix.

Identifying tissue-specific topics

We identified topic tissue-specificity in two ways: by overlapping peaks from each topic with peaks from ChIP-seq of cell type-specific factors (Fig. 4), and by assigning peaks to the nearest gene and assessing which tissues those genes tend to be expressed in (Fig. 5).

To compare the overlap of topic-associated sci-ATAC-seq peaks with TF ChIP-seq peaks, we created separate BED files for peaks from ChIP-seq experiments for HLH-1, ELT-1, or ELT-2, removed any peaks that had overlaps in 40 or more other ChIP experiments (i.e. high occupancy target, or HOT, sites), and merged any remaining overlapping peaks with `bedtools merge`. We intersected the peaks called from each topic cluster with the peaks from each transcription factor and recorded the number of overlapping peaks for each topic. To understand whether the observed overlaps per topic were surprising, we generated a null distribution by sampling a number of sci-ATAC-seq peaks equal to the number of observed overlaps, with each peak being drawn from a particular topic with a probability based on the total number of peaks called for that topic cluster. We then took the \log_2 ratio of the topic distribution of observed overlaps to the topic distribution of each sample of randomly-drawn peaks. We plot the mean \log_2 ratio, and use the samples to compute a 95% confidence interval around each bar (Fig. 4). We expanded our analysis and repeated this same procedure for all TF ChIP-seq data sets, and present the results as a heatmap in Fig. S6.

We also used previously published sci-RNA-seq data (Cao et al., 2017) to comprehensively identify cell types for our topic clusters by assessing the tissue expression patterns of genes near our topic-specific peaks. For each topic, we ranked the peaks by their topic-specificity, which we define for a given peak and topic as the fraction of cells with evidence for that peak that are members of the topic's cluster. We wrote a BED file for each topic cluster that contained the coordinates of either the top 250 peaks by topic-specificity or all peaks with topic-specificity greater than 0.5, whichever was greater (Fig. S7). Next, we associated these topic-specific peaks with their nearest downstream gene (within 1200 bp) by using `bedtools closest` with options `-D b`, `-io`, and `-id`). Similar to the cell type-specific TF ChIP-seq analysis above, we then asked whether the expression distribution of these top genes across tissues is enriched on average for particular tissues. Accordingly, we drew 100 samples of 250 genes from the null distribution of all genes with sci-RNA-seq data, and we compared the expression distribution of these gene sets with our sets of topic-specific genes by computing the \log_2 ratio of the mean topic-specific tissue expression distribution to the mean tissue expression distribution of each random sample. We again reported the mean \log_2 ratio and computed the 95% confidence interval around the mean. See Figure S8 for an example illustrating this procedure, and Figure S9 for the tissue expression distributions associated with all 37 topic clusters. Note that we tried using different numbers of topic-specific peaks, from the top 50 to the top 500, and found that the cell type identities ultimately assigned to the topics were rather insensitive to this choice. We decided to use 250 as a balance between having input from as many potential cell type-specific genes as possible without including too many non-topic-specific peaks (Fig. S7).

Identifying tissue subtypes using marker genes

To identify fine-grained cell types, we conducted a sub-clustering analysis for each of the following major worm tissues that we identified: coelomocyte (Figs. S21, S22), glia (Figs. S23, S24), gonad (Figs. S25, S26), hypodermis (Figs. S27, S28), intestine (Figs. 8, S15), muscle (Figs. 8, S14), neuron (Figs. 9, S17, S18, S19, S20, S16), and pharynx (Figs. S29, S30). There were too few cells identified as sex myoblast to conduct a sub-clustering on that tissue. First, we pooled the data for the cells in the topic clusters corresponding to each tissue, and also merged the peaks called from these topic clusters, to create a data set per tissue (Table 1).

Next, we ran the same iterative LDA-modeling and cell-clustering procedure as detailed above for the whole worm. We identified subclusters by coloring the UMAP scatter plots by the cells with

peaks near cell type-specific marker genes (see Figs. 8, 9). We colored a cell for a marker gene if it showed evidence for any peak that either overlapped the gene body or overlapped the region from 1200 bp upstream (5') of the gene to 100 bp downstream (3') of the gene. The marker genes we plot were used to identify fine-grained cell types in embryonic and L2 *C. elegans* scRNA-seq data (Packer et al., 2019).

Comparing peak versus gene tissue specificity

To summarize the tissue specificity of peaks and genes, for each gene we calculated the entropy of the distribution of expression levels across tissues in the scRNA-seq, and for each sci-ATAC-seq peak we calculated the entropy of the distribution of number of other overlapping peaks from each tissue. The entropy of a discrete distribution is a measure of how concentrated the probability of that distribution is in some subset of the elements of the distribution; for any domain, the uniform distribution has the highest possible entropy and a distribution with 100% of the probability at one point in the domain has the lowest possible entropy. Thus, low entropy of a peak or gene indicates high tissue specificity and high entropy indicates low tissue-specificity. To compare gene expression patterns with their corresponding accessibility patterns, we assigned all peaks to their nearest downstream exon within 1200 bp (allowing overlaps) and then for each gene we averaged the entropy values of all peaks linked to the exons of that gene. Plotting these values against each other for 13,111 genes suggest that most genes have a lower mean peak entropy than expression entropy, indicating a more tissue-specific pattern of regulation than their pattern of expression (Fig. S11). Some genes with high expression entropy and low accessibility entropy simply do not have much chromatin accessibility signal, while other examples show extensive chromatin accessibility that exhibits multiple tissue-specific patterns that independently have low entropy and thus lead to a low mean peak tissue entropy for that gene. To distinguish between these two cases, we calculated a peak diversity score for each gene, which is defined as the number of distinct overlap patterns among the peaks from different tissues assigned to that gene. For example, consider a hypothetical gene that has a neuron peak, an intestine peak, and a pharynx peak. Assume that the neuron peak does not overlap any other peaks, and the intestine and pharynx peaks do overlap. That gene would have a peak diversity score of two: there is one pattern of neuron-specific accessibility, and another of intestine plus pharyngeal accessibility. We extracted the set of 2,038 genes with peak tissue entropy less than 1.0 and expression entropy greater than 1.5, which are enriched for the genes with the broadest expression but most tissue-specific gene regulation, and provide them as Supplementary Table 1.

Generating UCSC Genome Browser tracks for the topic clusters

To generate signal files in bigWig format for display in the UCSC Genome Browser, we used the MACS2 output from our peak calling steps. The data from the input control bedGraph file was subtracted from the data in the treatment bedGraph using `macs2 bdgcmp -m subtract`, and the resulting bedGraph file was converted to bigWig format using the `bedGraphToBigWig` utility from UCSC (Kent et al., 2010). These bigWig tracks are displayed along with our peak calls in two track hubs on the UCSC Genome Browser:

- sci-ATAC-seq Main Track Hub URL: http://genome.ucsc.edu/cgi-bin/hgTracks?db=ce11&hubClear=http://waterston.gs.washington.edu/atacTissue/Durham_hub.txt

- sci-ATAC-seq Subclustering Track Hub URL: http://genome.ucsc.edu/cgi-bin/hgTracks?db=ce11&hubClear=http://waterston.gs.washington.edu/atacCellType/Durham_hub.txt

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo>) under accession number GSE157017. The parallel LDA code is available through GitHub at <https://github.com/gevirl/LDA> and as Supplemental Code, and is released under the GNU GPLv3 license. The data pipeline, scripts for performing the LDA clustering analysis, and Jupyter notebooks for generating figures are available at the following GitHub repository: https://github.com/tdurham86/L2_sci-ATAC-seq and as Supplemental Code, and is released under the MIT license.

Acknowledgements

We want to thank Anh Leith for sharing her expertise in FACS and helping us to collect samples and run the sci-ATAC-seq protocol. Chau Huynh for guidance when developing the worm nuclei isolation protocol and when optimizing our early experiments on bulk ATAC-seq. We would like to thank Illumina for providing the indexed Tn5 and also PCR reagents for the preparation of some of our sci-ATAC-seq libraries. This work was funded by NIH grants U41HG007355, R01GM072675, and R01HG010478 to RHW. JS is a Howard Hughes Medical Institute investigator.

Author contributions

TJD and RHW conceived of the experiment. TJD grew the worms and isolated the nuclei. OB helped to optimize ATAC-seq conditions for worm experiments and collected the ATAC-seq data from naked *C. elegans* genomic DNA for use as an input control when calling peaks. TJD and RMD performed the sci-ATAC-seq experiments. DAC helped perform the pilot sci-ATAC-seq experiment, and shared his code for the initial processing and quality control of the sequencing data. JS provided important reagents and lab equipment for the sci-ATAC-seq protocol. TJD analyzed the data and wrote the paper with input from RHW and WSN. LG wrote the parallelized LDA Java program. RHW, LG, and WSN contributed to discussions of the data analysis.

Disclosure Declaration

The authors declare no competing interests.

References

Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X., *et al.*, 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*, **11**(12):R119.

- Altun, Z. F. and Hall, D. H., 2002. WormAtlas Hermaphrodite Handbook - Epithelial System Hypodermis. *WormAtlas*, .
- Altun, Z. F. and Hall, D. H., 2009. WormAtlas Hermaphrodite Handbook - Alimentary system Intestine. *WormAtlas*, .
- Araya, C. L., Kawli, T., Kundaje, A., Jiang, L., Wu, B., Vafeados, D., Terrell, R., Weissdepp, P., Gevirtzman, L., Mace, D., *et al.*, 2014. Regulatory analysis of the *C. elegans* genome with spatiotemporal resolution. *Nature*, **512**(7515):400–405.
- Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., and Newell, E. W., 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, **37**(1):38–44.
- Blei, D. M., Ng, A. Y., and Jordan, M. I., 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, **3**:993–1022.
- Boeck, M. E., Huynh, C., Gevirtzman, L., Thompson, O. A., Wang, G., Kasper, D. M., Reinke, V., Hillier, L. W., and Waterston, R. H., 2016. The time-resolved transcriptome of *C. elegans*. *Genome Research*, **26**(10):1441–1450.
- Bolger, A. M., Lohse, M., and Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**(15):2114–2120.
- Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNAbinding proteins and nucleosome position. *Nature Methods*, **10**(12):1213–1218.
- Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J., 2015. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, **523**(7561):486–490.
- Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., *et al.*, 2018. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, **361**(6409):1380–1385.
- Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S. N., Steemers, F. J., *et al.*, 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, **357**(6352):661–667.
- Chen, X., Miragaia, R. J., Natarajan, K. N., and Teichmann, S. A., 2018. A rapid and robust method for single cell chromatin accessibility profiling. *Nature Communications*, **9**(1):5345.
- Cook, D. E., Zdraljevic, S., Roberts, J. P., and Andersen, E. C., 2017. CeNDR, the *Caenorhabditis elegans* natural diversity resource. *Nucleic Acids Research*, **45**(Database issue):D650–D657.
- Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B., *et al.*, 2017. An improved ATACseq protocol reduces background and enables interrogation of frozen tissues. *Nature Methods*, **14**(10):959–962.

- Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., Steemers, F. J., Trapnell, C., and Shendure, J., 2015. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, **348**(6237):910–914.
- Cusanovich, D. A., Hill, A. J., Aghamirzaie, D., Daza, R. M., Pliner, H. A., Berletch, J. B., Filippova, G. N., Huang, X., Christiansen, L., DeWitt, W. S., *et al.*, 2018a. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*, **174**(5):1309–1324.e18.
- Cusanovich, D. A., Reddington, J. P., Garfield, D. A., Daza, R. M., Aghamirzaie, D., MarcoFerrerres, R., Pliner, H. A., Christiansen, L., Qiu, X., Steemers, F. J., *et al.*, 2018b. The *cis*regulatory dynamics of embryonic development at single-cell resolution. *Nature*, .
- Daugherty, A. C., Yeo, R. W., Buenrostro, J. D., Greenleaf, W. J., Kundaje, A., and Brunet, A., 2017. Chromatin accessibility dynamics reveal novel functional enhancers in *C. elegans*. *Genome Research*, .
- Dey, K. K., Hsiao, C. J., and Stephens, M., 2017. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLOS Genetics*, **13**(3):e1006599. Publisher: Public Library of Science.
- Etchberger, J. F., Flowers, E. B., Poole, R. J., Bashllari, E., and Hobert, O., 2009. Cis-regulatory mechanisms of left/right asymmetric neuron-subtype specification in *C. elegans*. *Development*, **136**(1):147–160. Publisher: The Company of Biologists Ltd Section: Research Article.
- Fukushige, T., Hawkins, M. G., and McGhee, J. D., 1998. The GATA-factor *elt-2* is essential for formation of the *Caenorhabditis elegans* intestine. *Developmental Biology*, **198**(2):286–302.
- González-Blas, C. B., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., and Aerts, S., 2019. *cisTopic*: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods*, :1.
- Green, B., Bouchier, C., Fairhead, C., Craig, N. L., and Cormack, B. P., 2012. Insertion site preference of *Mu*, *Tn5*, and *Tn7* transposons. *Mobile DNA*, **3**:3.
- Griffiths, T. L. and Steyvers, M., 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, **101**(suppl 1):5228–5235.
- Ho, M. C. W., Quintero-Cadena, P., and Sternberg, P. W., 2017. Genome-wide discovery of active regulatory elements and transcription factor footprints in *Caenorhabditis elegans* using DNase-seq. *Genome Research*, **27**(12):2108–2119.
- Jänes, J., Dong, Y., Schoof, M., Serizay, J., Appert, A., Cerrato, C., Woodbury, C., Chen, R., Gemma, C., Huang, N., *et al.*, 2018. Chromatin accessibility dynamics across *C. elegans* development and ageing. *eLife*, **7**:e37344.
- Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D., 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**(17):2204–2207.
- Kim, H.-J., Gu'rkun Yardımcı, G., Bonora, G., Ramani, V., Liu, J., Qiu, R., Lee, C., Hesson, J., Ware, C. B., Shendure, J., *et al.*, 2019. Capturing cell type-specific chromatin structural patterns by applying topic modeling to single-cell Hi-C data. preprint, Genomics.

- Kimble, J., 2005. Germline proliferation and its control. *WormBook*, .
- Krause, M., Fire, A., Harrison, S. W., Priess, J., and Weintraub, H., 1990. CeMyoD accumulation defines the body wall muscle cell fate during *C. elegans* embryogenesis. *Cell*, **63**(5):907–919.
- Kudron, M. M., Victorsen, A., Gevirtzman, L., Hillier, L. W., Fisher, W. W., Vafeados, D., Kirkey, M., Hammonds, A. S., Gersch, J., Ammouri, H., *et al.*, 2018. The ModERN Resource: GenomeWide Binding Profiles for Hundreds of *Drosophila* and *Caenorhabditis elegans* Transcription Factors. *Genetics*, **208**(3):937–949.
- Langmead, B. and Salzberg, S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**(4):357–359.
- Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., Tay, T., *et al.*, 2020. Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, **183**(4):1103–1116.e20.
- McInnes, L., Healy, J., and Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*, . arXiv: 1802.03426.
- Packer, J. S., Zhu, Q., Huynh, C., Sivaramakrishnan, P., Preston, E., Dueck, H., Stefanik, D., Tan, K., Trapnell, C., Kim, J., *et al.*, 2019. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science*, :eaax1971.
- Page, B. D., Zhang, W., Steward, K., Blumenthal, T., and Priess, J. R., 1997. ELT-1, a GATA-like transcription factor, is required for epidermal cell fates in *Caenorhabditis elegans* embryos. *Genes & Development*, **11**(13):1651–1661.
- Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., Srivatsan, S., Qiu, X., Jackson, D., Minkina, A., *et al.*, 2018. Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Molecular Cell*, **71**(5):858– 871.e8.
- Quinlan, A. R. and Hall, I. M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6):841–842.
- Reinke, V., 2013. Transcriptional regulation of gene expression in *C. elegans*. *WormBook*, :1–31.
- Sinnamon, J. R., Torkenczy, K. A., Linhoff, M. W., Vitak, S. A., Mulqueen, R. M., Pliner, H. A., Trapnell, C., Steemers, F. J., Mandel, G., and Adey, A. C., *et al.*, 2019. The accessible chromatin landscape of the murine hippocampus at single-cell resolution. *Genome Research*, :gr.243725.118.
- Song, L. and Crawford, G. E., 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harbor protocols*, **2010**(2):pdb.prot5384.
- Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R., *et al.*, 2019. Comprehensive Integration of Single-Cell Data. *Cell*, **177**(7):1888–1902.e21. Publisher: Elsevier.

- Sulston, J. and Horvitz, H., 1977. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Developmental Biology*, **56**(1):110–156.
- Taylor, S. R., Santpere, G., Reilly, M., Glenwinkel, L., Poff, A., McWhirter, R., Xu, C., Weinreb, A., Basavaraju, M., Cook, S. J., *et al.*, 2019. Expression profiling of the mature *C. elegans* nervous system by single-cell RNA-Sequencing. *bioRxiv*, :737577. Publisher: Cold Spring Harbor Laboratory Section: New Results.
- Thompson, O., Edgley, M., Strasbourger, P., Flibotte, S., Ewing, B., Adair, R., Au, V., Chaudhry, I., Fernando, L., Hutter, H., *et al.*, 2013. The million mutation project: A new approach to genetics in *Caenorhabditis elegans*. *Genome Research*, **23**(10):1749–1762.
- Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D., 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, Montreal, Quebec, Canada. ACM Press.
- Warner, A. D., Gevirtzman, L., Hillier, L. W., Ewing, B., and Waterston, R. H., 2019. The *C. elegans* embryonic transcriptome with tissue, time, and alternative splicing resolution. *Genome Research*, . Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.
- Welch, J. D., Hartemink, A. J., and Prins, J. F., 2017. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome Biology*, **18**:138.
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., *et al.*, 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, **9**(9):R137.

Figure Legends and Tables

Figure 1: An iterative peak-calling procedure yields more peaks from the complex mix of worm cell types. **a.** The core peak-calling procedure is to model the data using latent Dirichlet allocation, cluster the cells, and call peaks based on the clusters. **b.** A flow chart represents the overall peak calling strategy. First, bulk peaks are called, followed by two iterations of clustering and peak calling based on an LDA model. Then, we group the cells by tissue and repeat the two steps of clustering and peak calling. The number of cells included and the number of peaks called at each step is given in the inset table.

Figure 2: The peaks called from sci-ATAC-seq data exhibit substantial overlap with existing chromatin data collected from whole worms. **a.** The majority of sci-ATAC-seq peaks overlap sites called in the TF ChIP-seq peaks from modERN or the bulk ATAC-seq peaks from Jänes, et al. 2018. **b.** Peaks from the other data sets also show substantial overlap with sci-ATAC-seq peaks. Most of the ChIP-seq TF peaks that do not overlap a sci-ATAC-seq peak are singleton peaks that are only found in a single experiment. **c.** Breaking out the ChIP-seq peak overlaps by the developmental stage of the worms assayed and comparing the distribution across stages of the peaks with overlaps compared to the stage distribution for randomly selected ChIP-seq peaks shows an enrichment for peaks found in larval stage L2. Error bars indicate the 95% confidence interval.

Figure 3: Latent Dirichlet allocation modeling yields 37 major cell clusters that are characterized mostly by a single topic each. **a.** LDA modeling learns latent topics that explain the data and return two matrices, here designated P and C . Matrix P , referred to in the text as the peaks-by-topics matrix, captures the probability distribution of each topic over all peaks, while matrix C , referred to in the text as the cells-by-topics matrix, captures the probability distribution of each cell over all topics. **b.** Heatmap showing the normalized C matrix values for the 37 topics associated with clusters; this plot highlights that most cells have probability concentrated in one or a few topics. Cell types determined for the topics based on analysis of the P matrix are annotated on the left, and the number of peaks per cell is shown to the right. **c.** UMAP embedding of the C matrix colored to indicate the 37 cell clusters. Any cells that are not assigned to a cluster are plotted as small gray dots and are mostly found on the periphery of the clusters.

Figure 4: Overlapping peaks important for each topic with ChIP-seq peaks collected from cell type-specific TFs suggests at least some topics represent tissue types. Peaks associated with each topic were overlapped with ChIP-seq peaks for three cell type-specific transcription factors: HLH-1, which is specific for muscle (top plot); ELT-1, which is specific for seam cells (middle); and ELT-2, which is specific for intestine (bottom). Topic distributions for peaks with ChIP-seq site overlaps were compared to the topic distribution for randomly sampled peaks, and the results are plotted here as the \log_2 ratio of the overlap topic distribution to the random topic distribution. Error bars represent the 95% confidence interval for 100 random samples.

Figure 5: Topic-specific peaks tend to be near tissue-specific genes. Peaks associated with each topic were mapped to the nearest downstream gene, and the tissue expression distribution of the genes near the top 250 most-specific peaks for each topic was compared

with the tissue expression distribution of 250 randomly-selected genes. Here we plot the results as the \log_2 ratio of the topic-associated tissue expression distribution to that of randomly selected genes. Error bars represent the 95% confidence interval after comparing to the tissue expression distribution of 100 random samples. LDA topic numbers are shown in the bottom right corner of each plot. Topics with similar tissue specificity patterns are grouped together, and the tissue type names and colors are as in Cao, et al. 2017. Tissue assignments were made by eye based on the tissue with maximal fold-change, and are written in the bottom left corner of each plot. If no single tissue was clearly the maximum, then a more general tissue annotation was chosen (e.g. “Neurons” for topic 38). These annotations may need to be revisited with new data. Note that for concise visualization in this figure we display just 20 of our 37 topics, but we report a version of this figure with all 37 topics in Fig. S9. All plots have the same y-axis range, from a \log_2 ratio of -4.5 to 4.0.

Figure 6: Novel sites of accessible chromatin with no overlapping modERN ChIP-seq peaks exhibit topic specificity. We compare the normalized peak-by-topic matrix values between the peaks that overlap a ChIP-seq site (**a.**) and those that do not (**b.**). The non-overlapping peaks are enriched for topics associated with gonad (especially germline/topic 24) and topics associated with neurons. The non-overlapping peaks also tend to be observed in fewer cells.

Figure 7: Known tissue-specific genes show topic-specific chromatin accessibility. UCSC Genome Browser multi-locus view of the regions surrounding nine known tissue-specific genes (top and bottom rows), and the tissue expression patterns from sci-RNA-seq (middle row). In each genome browser view, the top track shows the locations of sci-ATAC-seq peaks colored by tissue type, the second track shows the stacked sci-ATAC-seq signal from each tissue, the third track shows consensus peak regions around local maxima in the signal track, and the fourth track shows the gene models. The gene expression bar plots show expression values for 27 tissues in TPM units, with the same coloring and order as the legend in Fig. 5.

Figure 8: Subclustering of muscle and intestinal cells separates them by position along the anterior-posterior body axis. **a.** Peaks near genes that should be expressed throughout a tissue, like *hlh-1* and *myo-3* in body wall muscle or *end-1* and *elt-2* in intestine, show accessibility in cells throughout the UMAP. **b.** In both muscle and intestine data, we can detect subclusters of cells that show peaks near genes that mark the anterior or posterior regions of these tissues based on literature and microscopy data (Packer et al., 2019).

Figure 9: Subclustering of neurons reveals finer structure that distinguishes different types of neurons. **a.** Cells with reads in peaks near genes with expression patterns specific to neuron subtypes cluster together (*bbs-8*: ciliated sensory neurons, *gcy-32*: oxygen sensory neurons, *unc-30*: GABA-ergic neurons, *mec-7*: touch receptor neurons, *ceh-24*: cholinergic neurons). **b.** Cells in the UMAP plot are colored by the number of marker genes with nearby co-accessible peaks. Here, we show marker genes for the ASE neurons, a specific pair of ciliated sensory neurons, which are identified in one of the *bbs-8* clusters from **a.** (marked by the left-facing arrow); and marker genes shared by the oxygen sensory neurons AQR, PQR, and URX, which further support the cluster marked with *gcy-32* in **a.** (marked by the right-facing arrow).

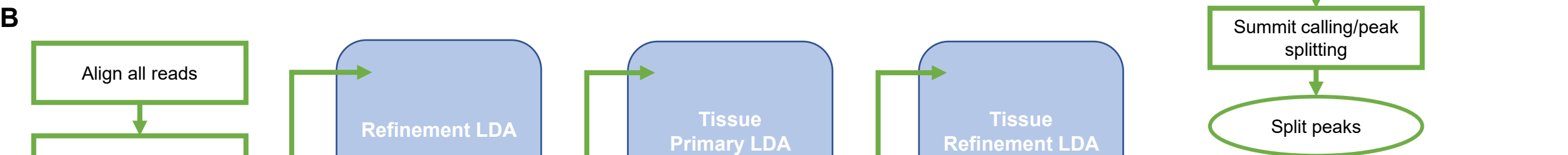
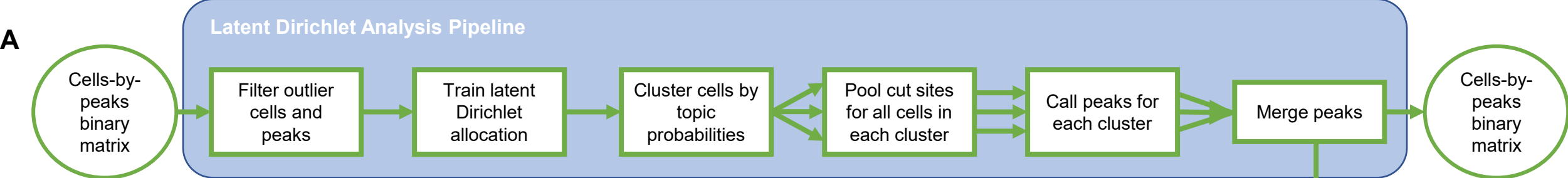
Tissue assignment	Topic number	Number of nuclei	Percent of nuclei	L2 nuclei count	L2 nuclei percent
Coelomocyte	4	286	1.2 %		
	Total:	286	1.2 %	6	0.8 %
Glia	12	321	1.3 %		
	21	663	2.7 %		
	27	421	1.7 %		
	31	166	0.7 %		
	Total:	1571	6.4 %	64	8.8 %
Gonad	7	360	1.5 %		
	24	1653	6.7 %		
	36	171	0.7 %		
	48	114	0.5 %		
	Total:	2298	9.4 %	12*	1.6 %
Hypodermis	1	2588	10.6 %		
	9	146	0.6 %		
	10	416	1.7 %		
	17	681	2.8 %		
	25	1631	6.7 %		
	30	80	0.3 %		
	41	150	0.6 %		
	43	468	1.9 %		
	46	216	0.9 %		
	Total:	6376	26.1 %	143†	19.6 %
Intestine	13	155	0.6 %		
	23	286	1.2 %		
	47	106	0.4 %		
	51	4422	18.0 %		
	Total:	4969	20.2 %	42‡	5.7 %
Muscle	40	2328	9.5 %		
	Total:	2328	9.5 %	100	13.7 %
Neuron	0	615	2.5 %		
	6	405	1.7 %		
	14	696	2.8 %		
	15	712	2.9 %		
	16	406	1.7 %		
	18	141	0.6 %		
	19	218	0.9 %		
	32	1095	4.5 %		
	33	247	1.0 %		
	38	981	4.0 %		
	45	399	1.6 %		
Total:	5915	24.2 %	282	38.6 %	
Pharynx	35	238	1.0 %		
	53	378	1.5 %		
	Total:	616	2.5 %	80	10.9 %
Sex myoblast	37	144	0.6 %		
	Total:	144	0.6 %	2	0.3 %

Table 1: Results of the iterative LDA procedure, with the number of cells assigned to each topic cluster, the tissue assignment, and the percentage of all 24,503 cells that were assigned to a topic cluster along with the predicted fraction of nuclei based on the known L2 cell composition.

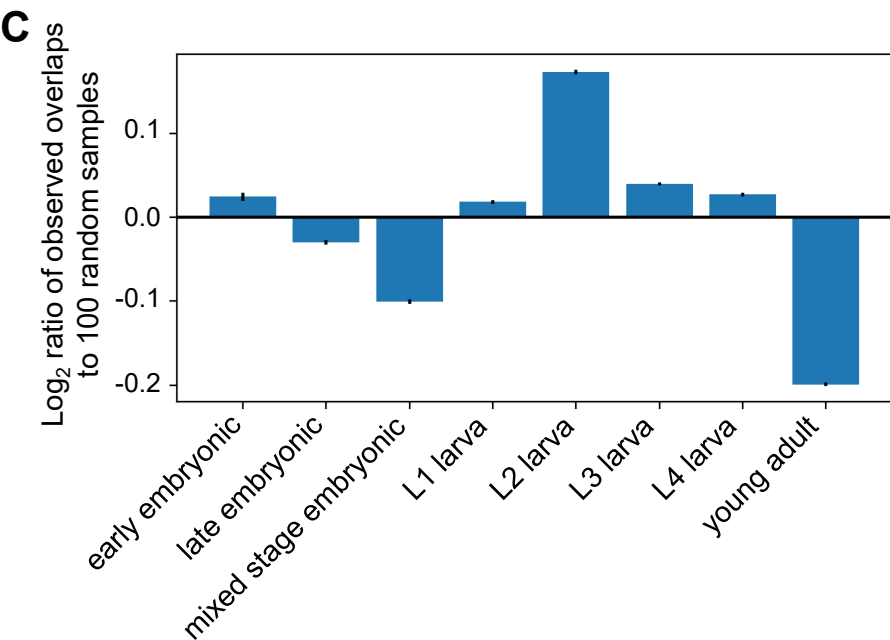
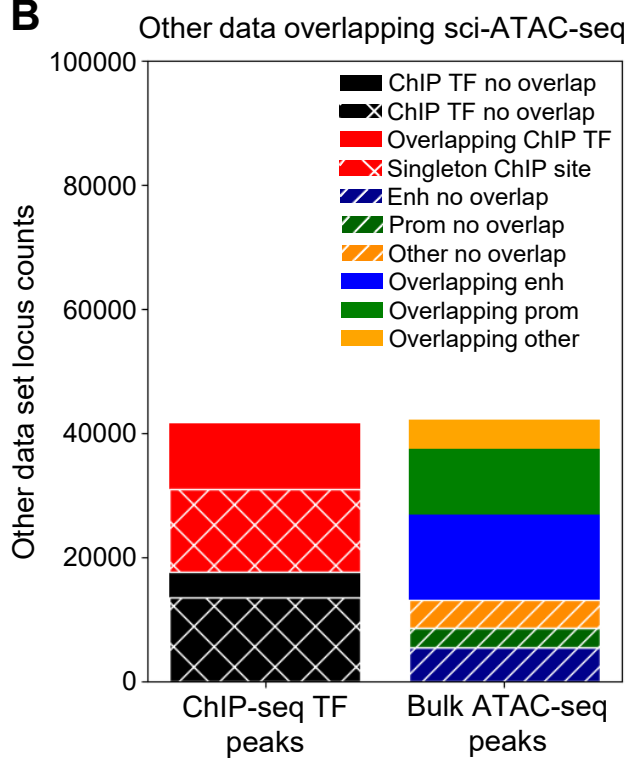
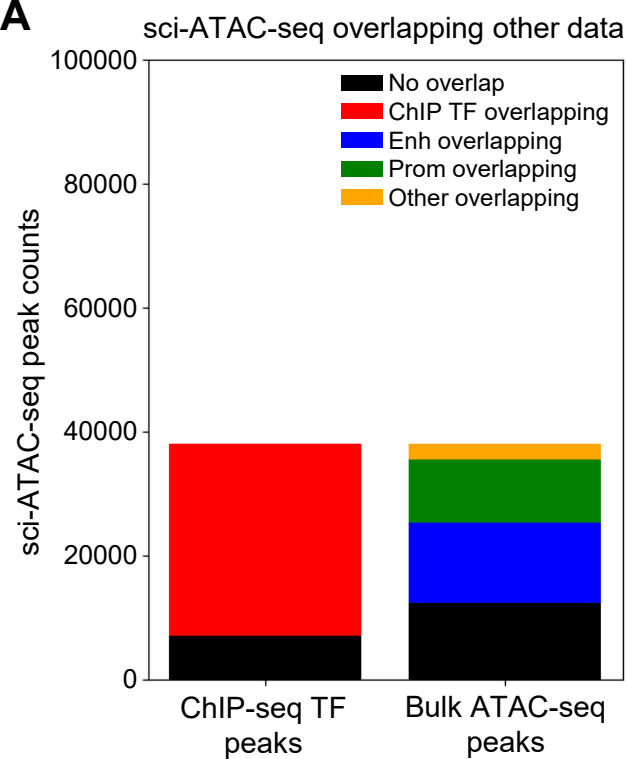
*: There are twelve somatic gonad cells at L2, plus a rapidly growing number of mitotically dividing germ cells (see text).

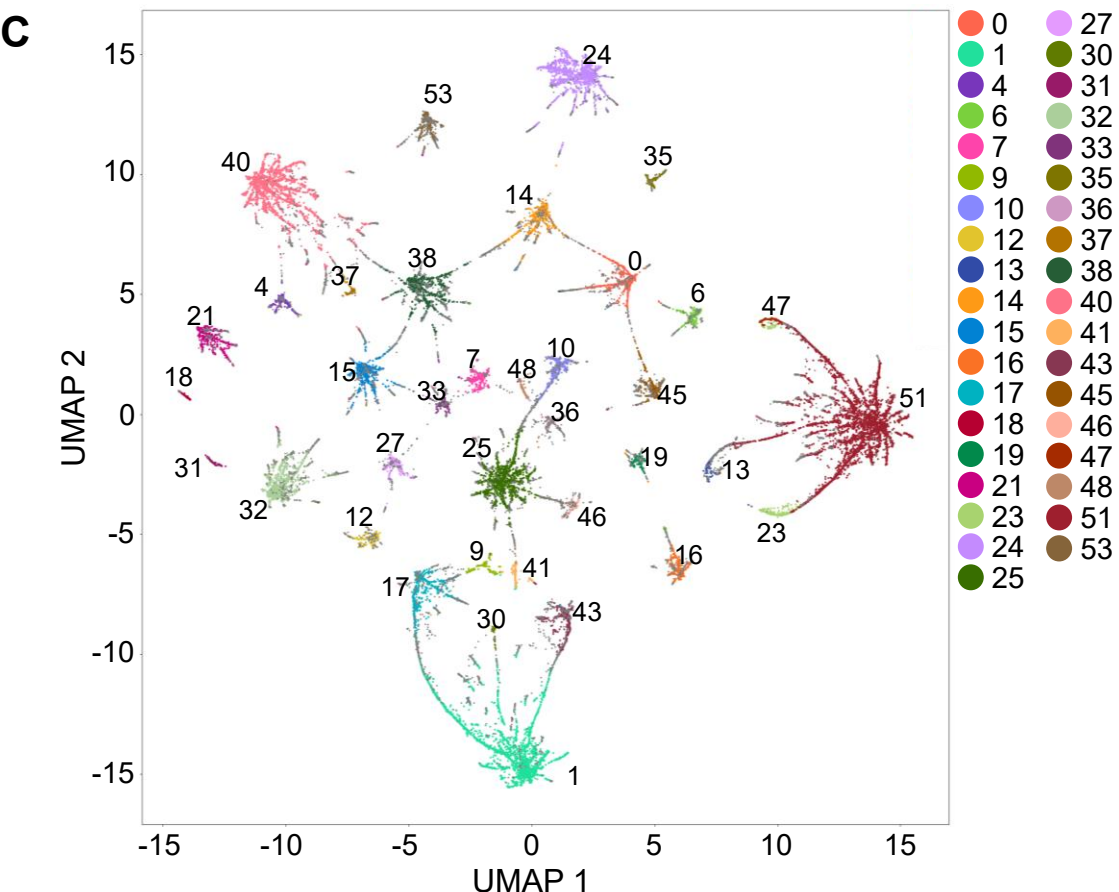
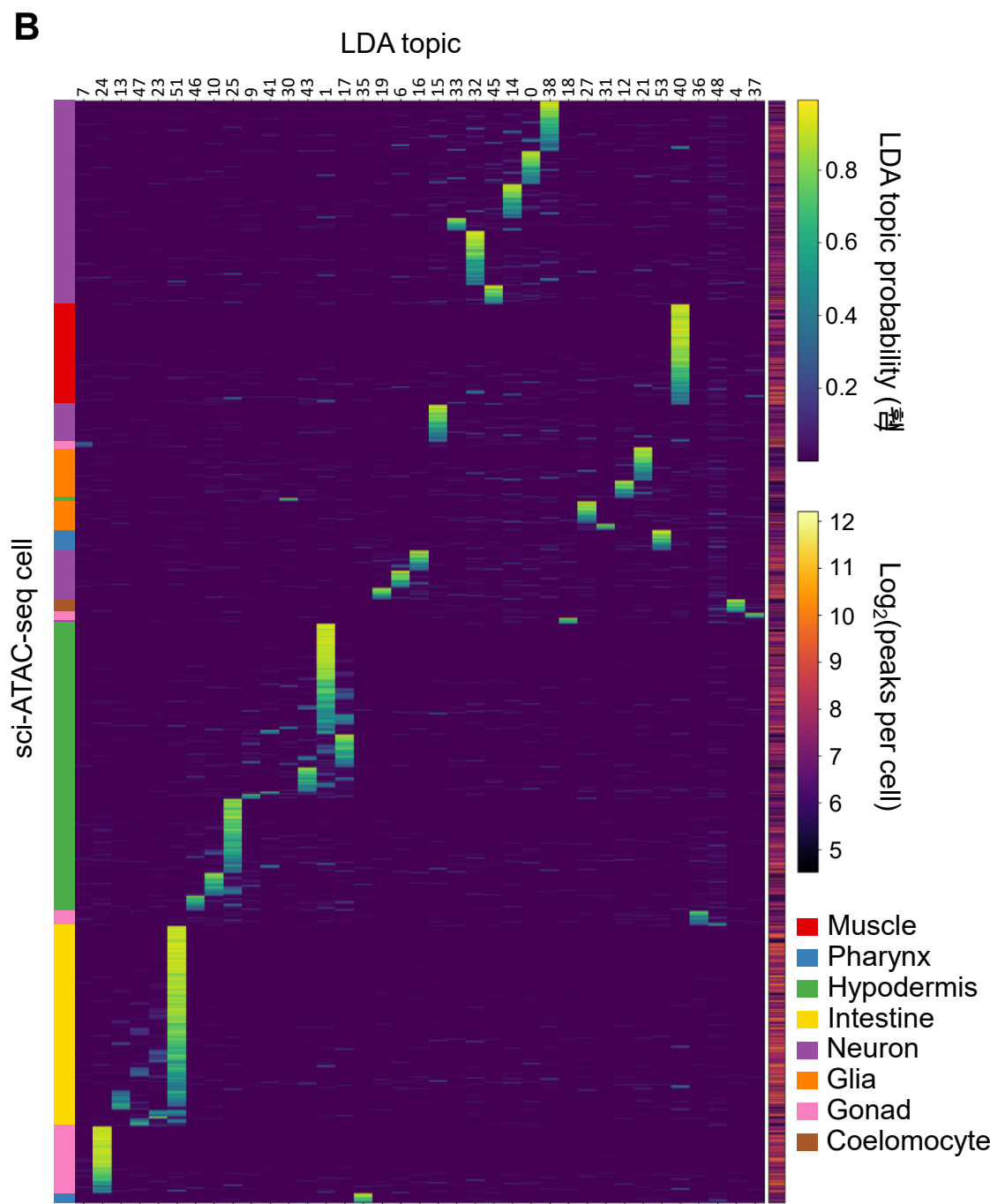
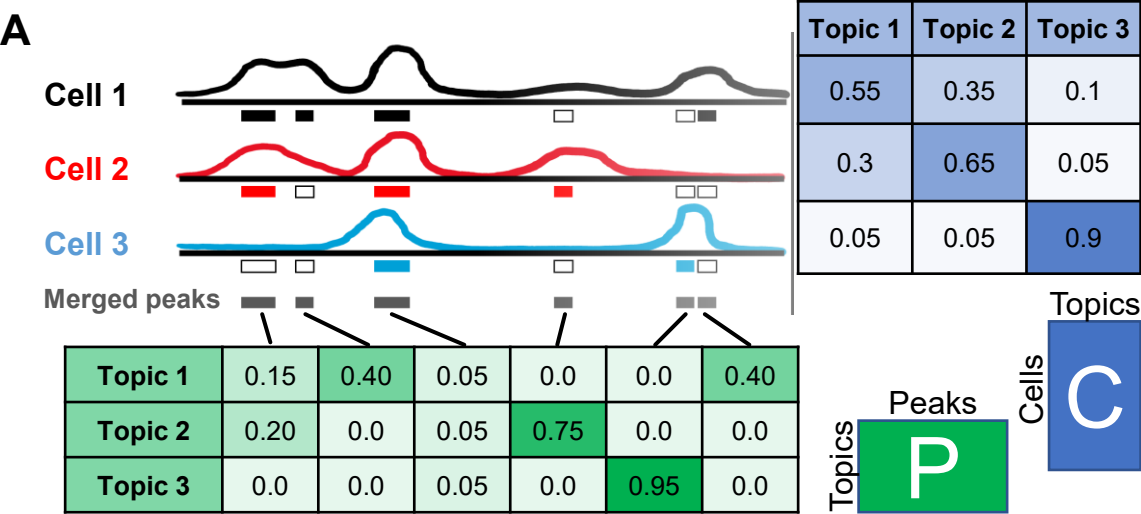
†: The hypodermis is composed of 13 hyp cells, nine of which are syncytial cells, and 16 seam cells. The number reported in the table is the number of nuclei in the hyp syncytium and seam cells at L2.

‡: The 20 intestinal cells at L2 are polyploid, and some are multinucleate. Thus, these 20 cells have 34 nuclei with an 8N DNA content. Consistent with the latter, the identified intestine cells have a much higher read coverage on average, facilitating their identification (Fig. 3B). The intestine count reported in the table also includes an additional 8 diploid nuclei found in the intestinal valve cells.

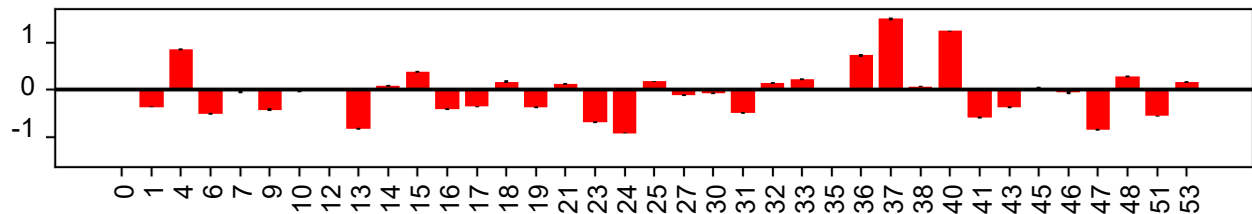


Category	Post merge peaks		Primary LDA input		Primary LDA output		Refinement LDA input		Refinement LDA output	
	Cells	Peaks	Cells	Peaks	Cells	Peaks	Cells	Peaks	Cells	Peaks
All	30930	15044	30763	13734	30763	34339	30870	32214	30870	36339
Coelomocyte	286	6116	285	5699	285	6547	285	6088	285	5825
Glia	1571	9321	1511	8676	1511	10216	1504	8432	1504	10495
Gonad	2298	14390	2211	13556	2211	13313	2191	12288	2191	13982
Hypodermis	6376	14222	6316	12864	6316	13329	6267	11875	6267	14298
Intestine	4969	11989	4850	10666	4850	12555	4867	11427	4867	12792
Muscle	2328	8472	2249	7838	2249	15779	2249	14897	2249	20418
Neuron	5915	19094	5813	17344	5813	29521	5745	27358	5745	29779
Pharynx	616	8693	616	8028	616	10946	614	9926	614	9788
Sex myoblast	144	3758	143	3301	--	--	--	--	--	--

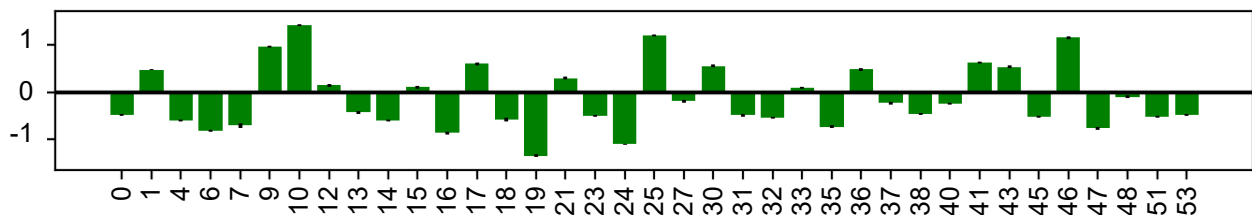




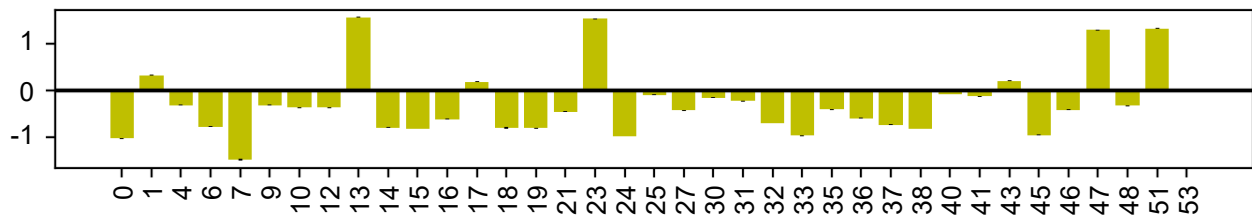
Peaks overlapping HLH-1 TF ChIP-seq sites



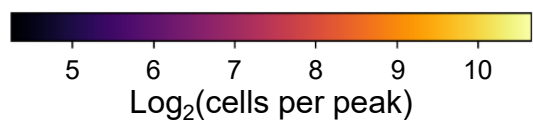
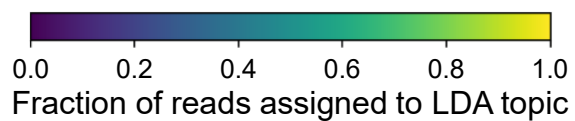
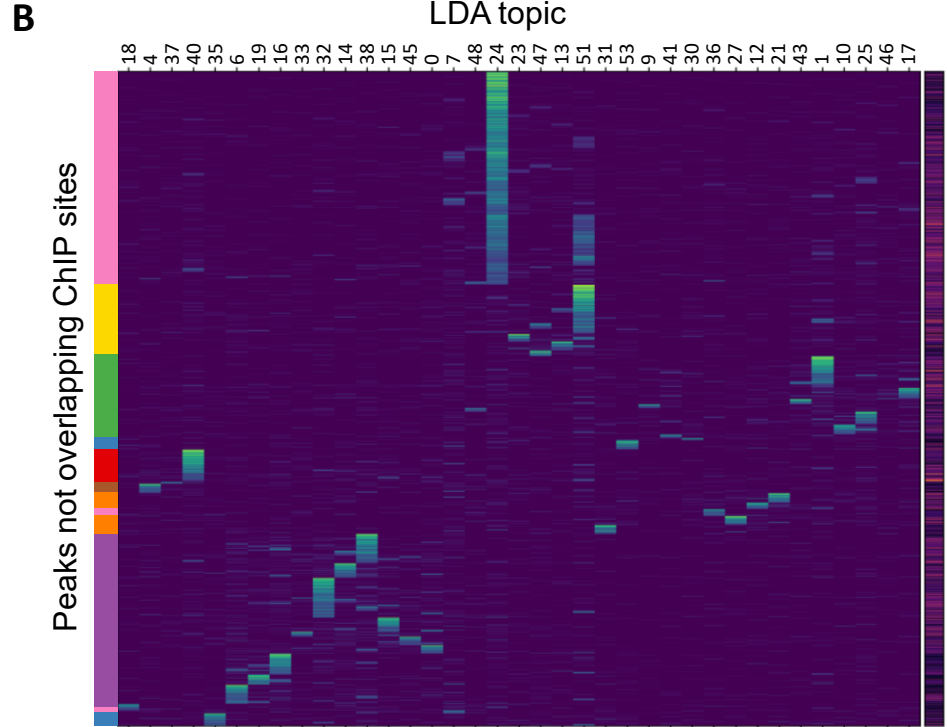
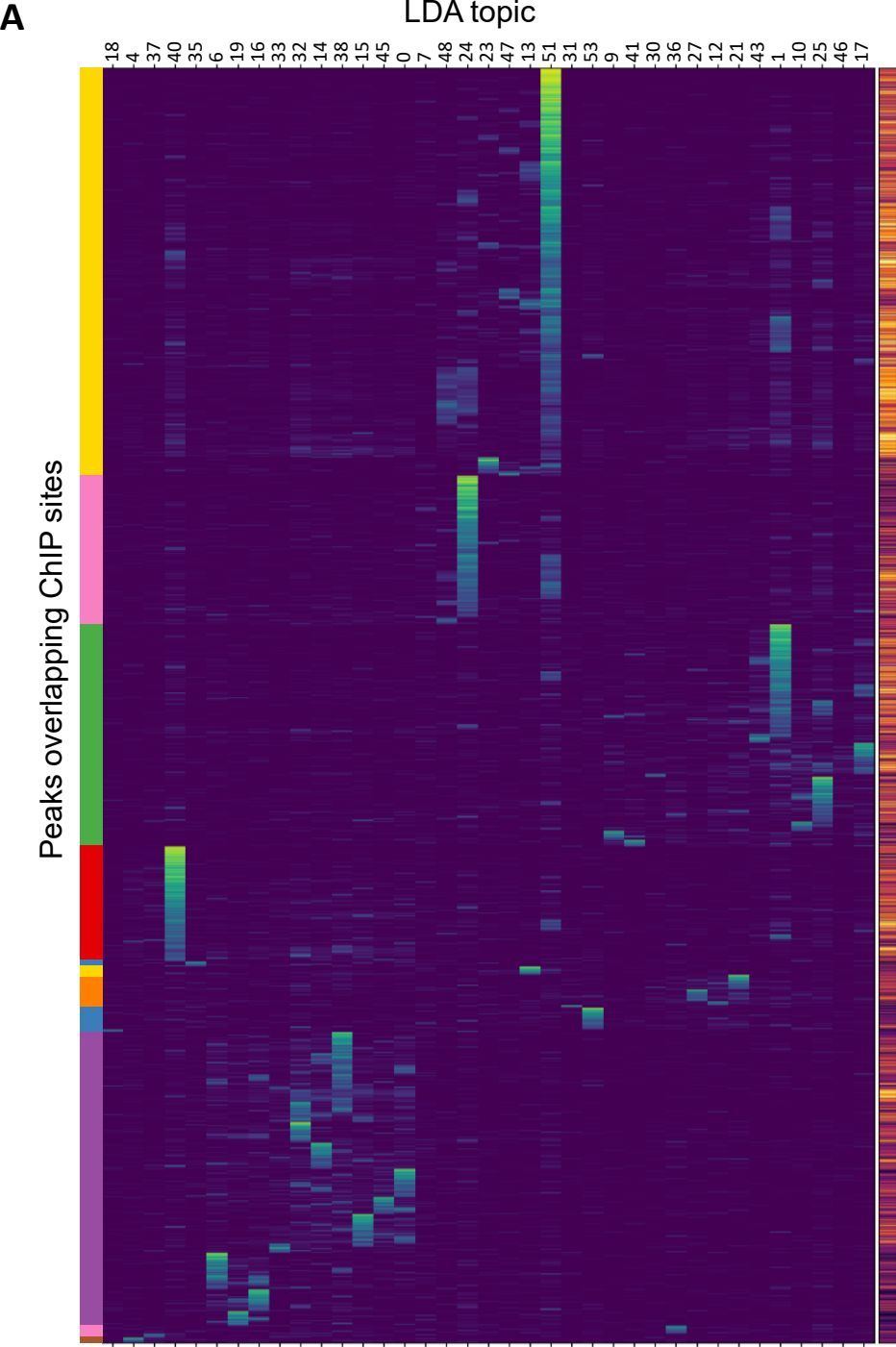
Peaks overlapping ELT-1 TF ChIP-seq sites



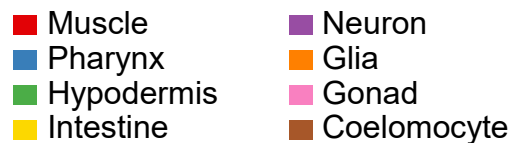
Peaks overlapping ELT-2 TF ChIP-seq sites

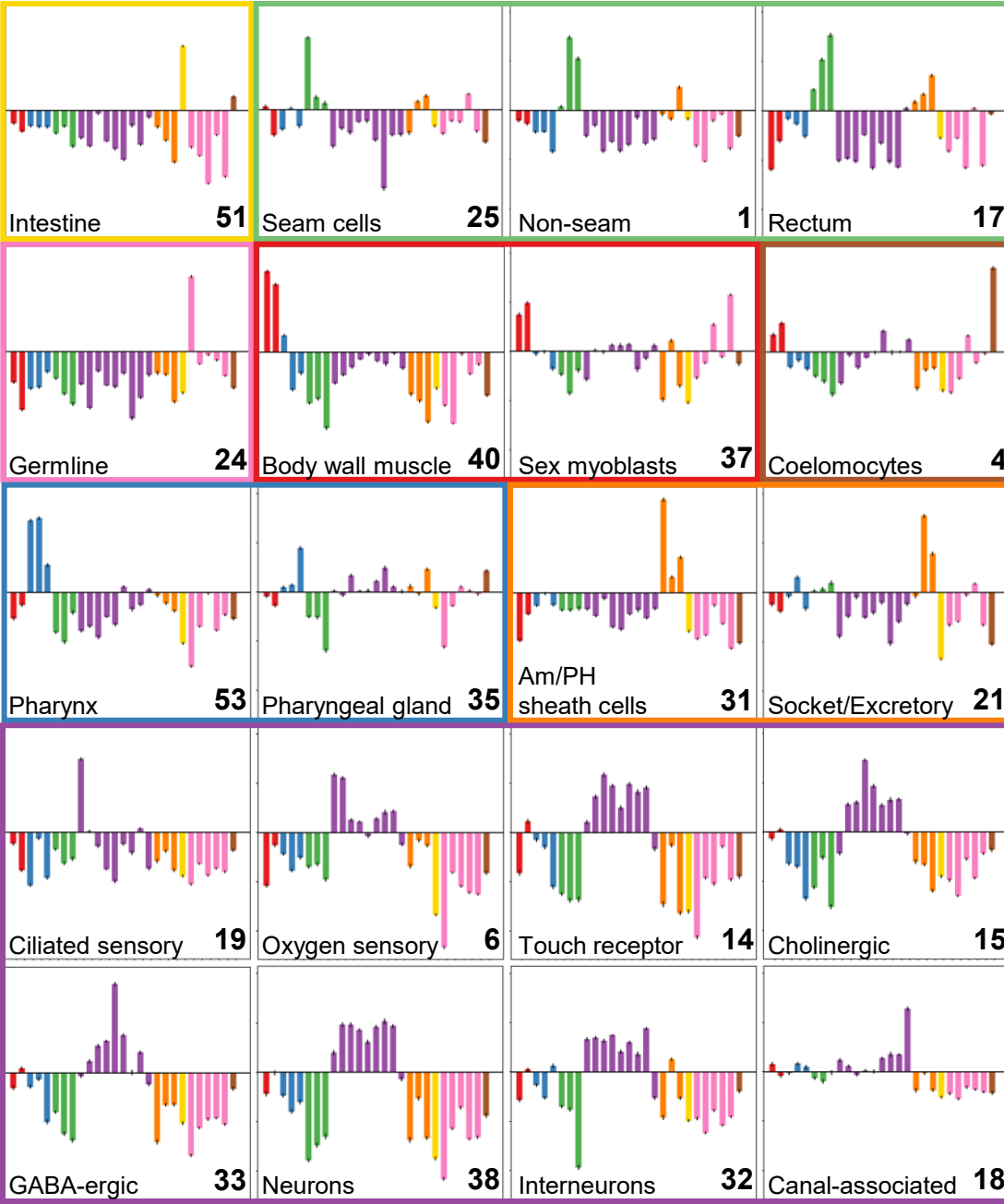


LDA topic number

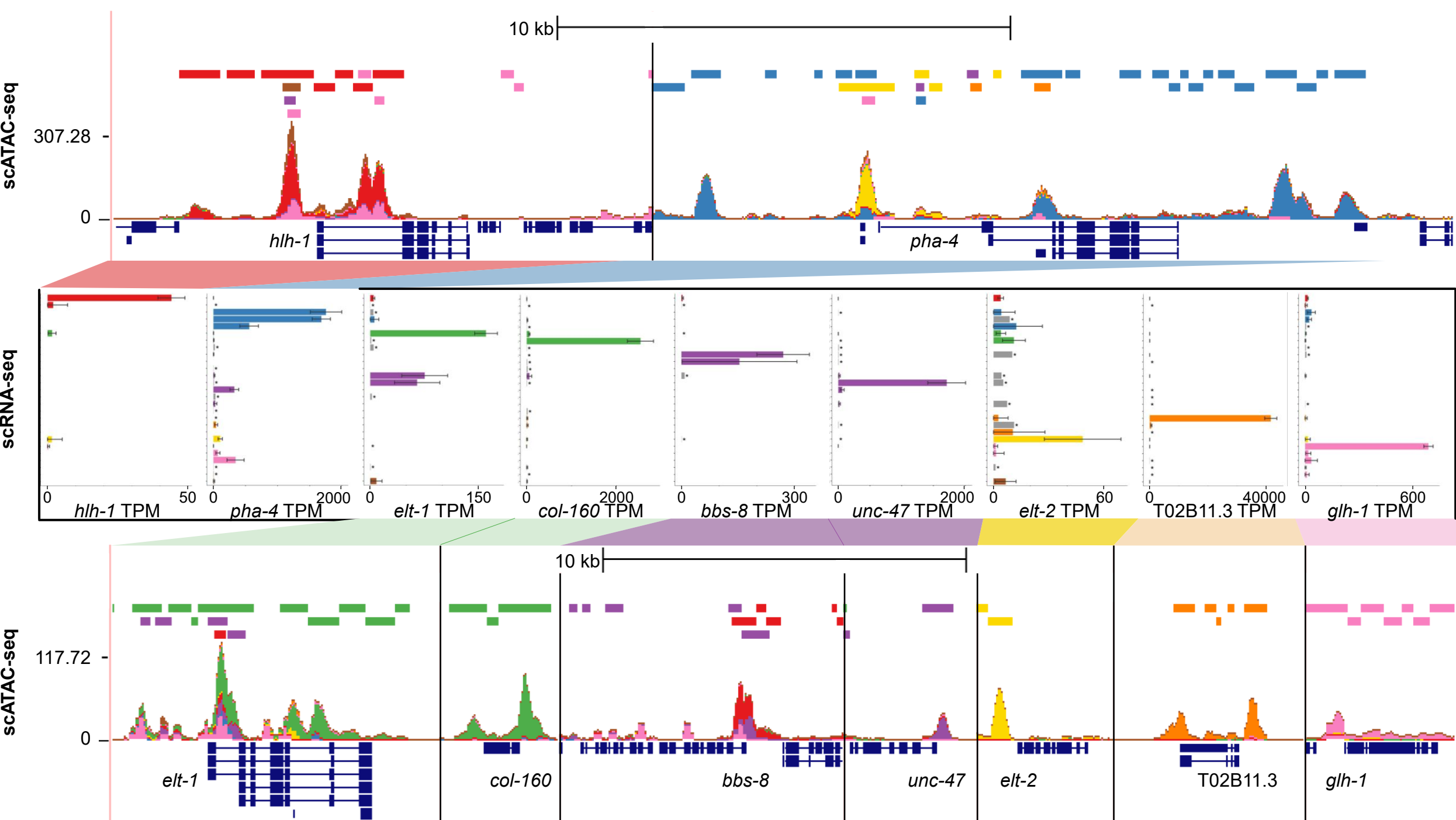


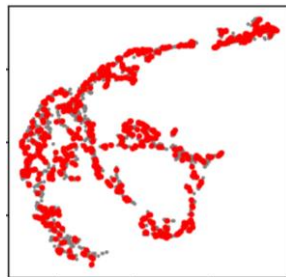
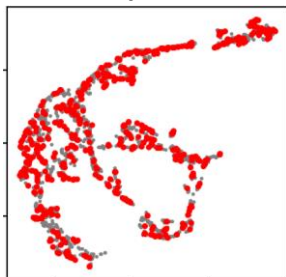
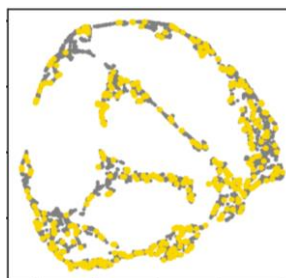
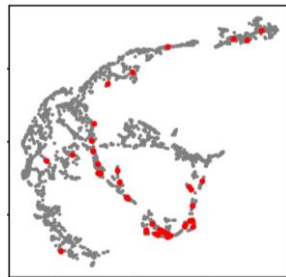
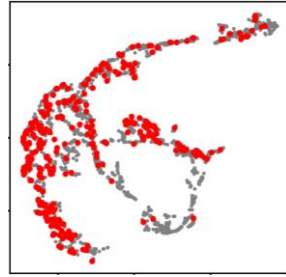
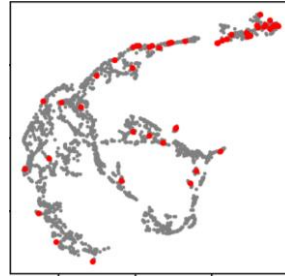
Topic tissue assignment



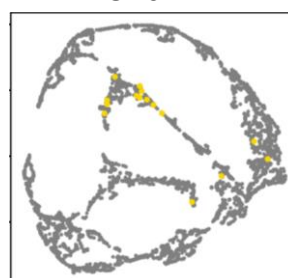
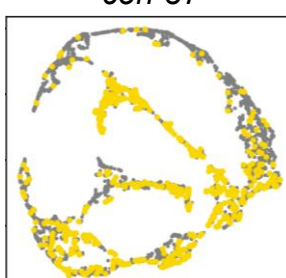
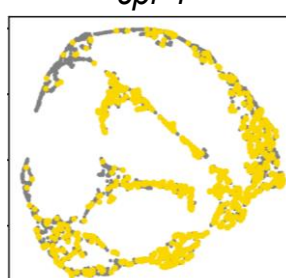
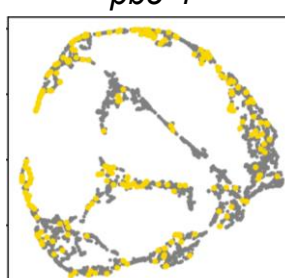
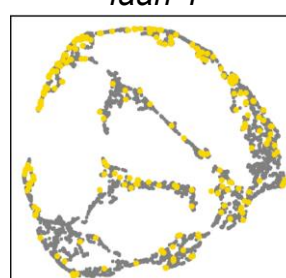


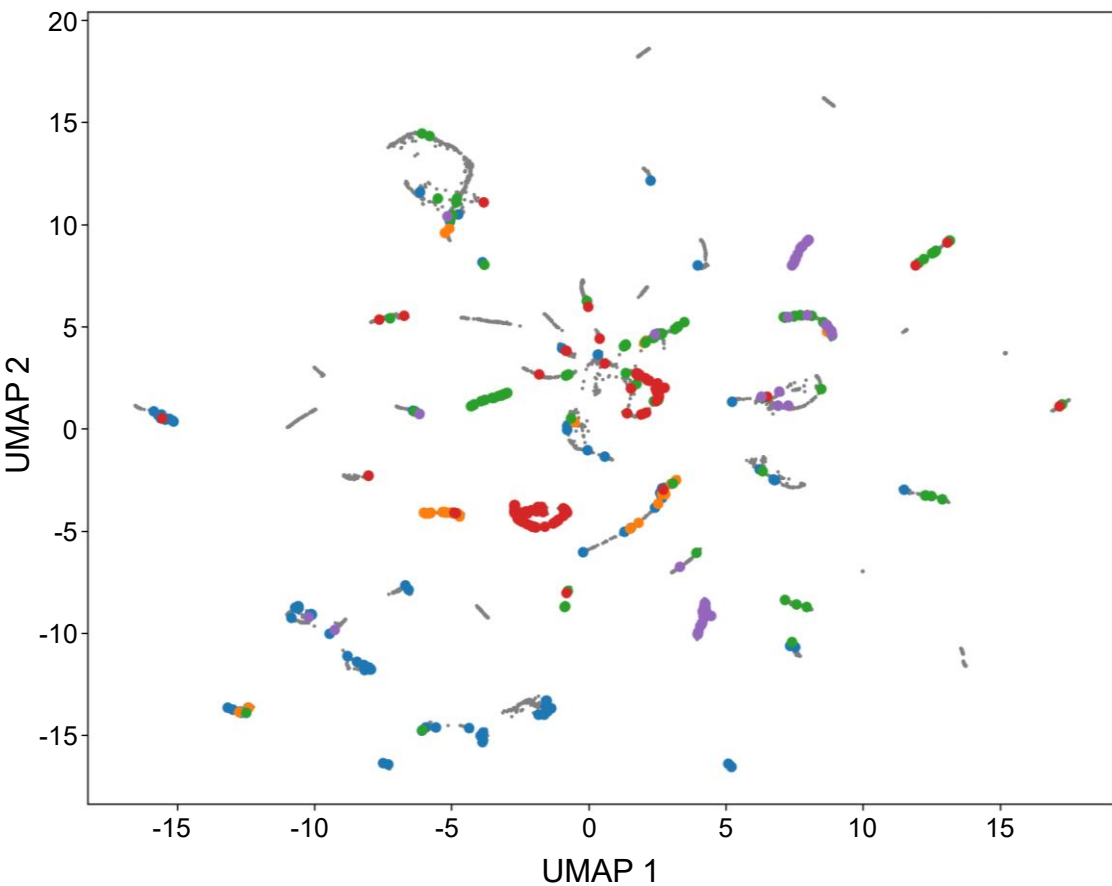
- █ Body wall muscle
- █ Intestinal/rectal muscle
- █ Pharyngeal muscle
- █ Pharyngeal epithelia
- █ Pharyngeal gland
- █ Seam cells
- █ Non-seam hypodermis
- █ Rectum
- █ Ciliated sensory neurons
- █ Oxygen sensory neurons
- █ Touch receptor neurons
- █ Cholinergic neurons
- █ GABAergic neurons
- █ Pharyngeal neurons
- █ *flp-1(+)* interneurons
- █ Other interneurons
- █ Canal associated neurons
- █ Am/PH sheath cells
- █ Socket cells
- █ Excretory cells
- █ Intestine
- █ Germline
- █ Somatic gonad precursors
- █ Distal tip cells
- █ Vulval precursors
- █ Sex myoblasts
- █ Coelomocytes



A*hlh-1**myo-3***Tissue Marker***end-1**elt-2***B***tni-3**lin-39**cwn-1***Anterior****Posterior**

ZC204.12

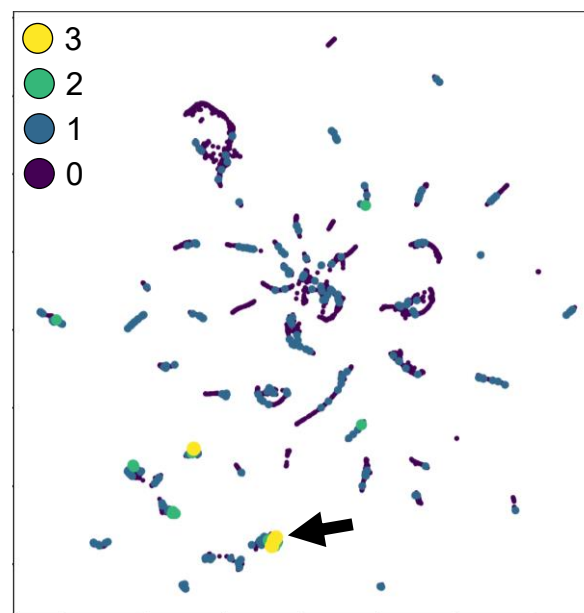
*ceh-37**cpr-1**pbo-4**faah-1*

A**Neuron type**

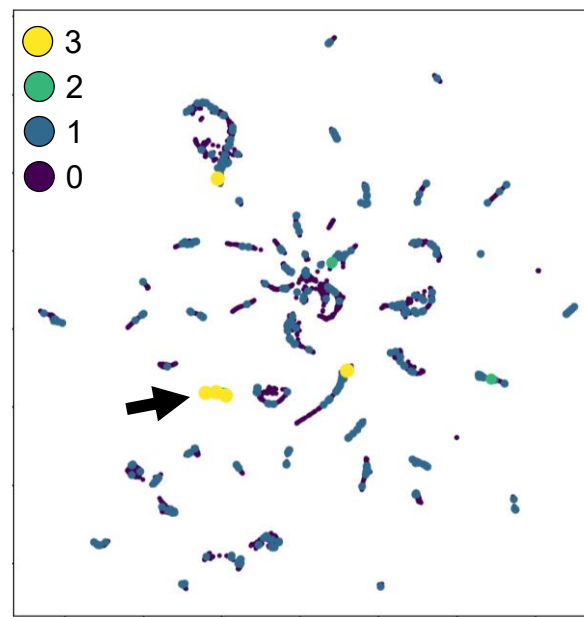
- ciliated sensory
- oxygen sensory
- touch sensory
- GABA-ergic
- cholinergic

Marker gene

bbs-8
gcy-32
mec-7
unc-30
ceh-24

B**Neuron: ASE**

Type: chemosensory
Markers: *gcy-6*, *gcy-7*,
gcy-14, *gcy-5*, *gcy-22*,
gcy-4

**Neuron: AQR, PQR, URX**

Type: oxygen sensory
Markers: *gcy-32*, *gcy-35*,
gcy-36, *gcy-37*