Inferring predictive genetic models and regulatory elements by deep learning of cross-species single-cell gene expression landscapes

Jiaqi Li
Jingjing Wang (jingjingw@zju.edu.cn)
Zhejiang University
Peijing Zhang
Renying Wang
Yuqing Mei
Zhongyi Sun
Lijiang Fei
Mengmeng Jiang
Lifeng Ma
Weigao E
Haide Chen
Xinru Wang
Yuting Fu
Daiyuan Liu
Xueyi Wang
Jingyu Li
Qile Guo
Yuan Liao
Chengxuan Yu
Danmei Jia
Jian Wu
Shibo He
Huanju Liu
Jun Ma
Kai Lei
Jiming Chen
Xiaoping Han (xhan@zju.edu.cn)
Zhejiang University
Guoji Guo (ggj@zju.edu.cn)
Zhejiang University https://orcid.org/0000-0002-1716-4621

Research Article

Keywords: single-cell, deep learning, TF, CRE, cross-species

Posted Date: April 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1544073/v1

Inferring predictive genetic models and regulatory elements by deep learning of cross-species single-cell gene expression landscapes

Jiaqi Li[1,2†], Jingjing Wang[1,2†*], Peijing Zhang[1,2†], Renying Wang[1†], Yuqing Mei[1], Zhongyi Sun[1], Lijiang Fei[1], Mengmeng Jiang[1,2], Lifeng Ma[1], Weigao E[1], Haide Chen[1,2], Xinru Wang[1], Yuting Fu[1], Daiyuan Liu[1], Xueyi Wang[1], Jingyu Li[1], Qile Guo[3], Yuan Liao[1,4], Chengxuan Yu[1], Danmei Jia[1], Jian Wu[5], Shibo He[6], Huanju Liu[7], Jun Ma[7], Kai Lei[8], Jiming Chen[6], Xiaoping Han[1,4*], Guoji Guo[1,2,3,4,9*]

[1]Center for Stem Cell and Regenerative Medicine, and Bone Marrow Transplantation Center of the First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, 310000, China.

[2]Liangzhu Laboratory, Zhejiang University Medical Center, Hangzhou, Zhejiang 311121, China.

[3]Zhejiang University-University of Edinburgh Institute, Zhejiang University School of Medicine, Zhejiang University, Hangzhou 314400, China.

[4]Zhejiang Provincial Key Lab for Tissue Engineering and Regenerative Medicine, Dr. Li Dak Sum & Yip Yio Chin Center for Stem Cell and Regenerative Medicine, Hangzhou, Zhejiang 310058, China.

[5]Division of Hepatobiliary and Pancreatic Surgery, Department of Surgery, First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang 310003, China.

[6]College of Control Science and Engineering, Zhejiang University, Hangzhou, Zhejiang 310027, China.

[7]Women's Hospital and the Institute of Genetics, Zhenjiang University School of Medicine, Hangzhou, Zhejiang 310058, China.

[8]Westlake Laboratory of Life Sciences and Biomedicine, Key Laboratory of Growth Regulation and Translational Research of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, Zhejiang 310027, China

[9]Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou, Zhejiang 310058, China.

[†]These authors contributed equally to this work.

[*]Correspondence author. Email: ggj@zju.edu.cn (G.G.); xhan@zju.edu.cn (X.H.); jingjingw@zju.edu.cn (J.W.)

Despite extensive efforts to sequence different genomes, genetic models to interpret gene regulation and cell fate decisions are lacking for most species. Here, we performed whole-body single-cell transcriptomic analysis of zebrafish, *Drosophila*, and earthworm. We then mapped cell landscapes covering eight representative metazoan species to study gene regulation through evolution. With uniformly constructed cross-species datasets, we developed a deep learning-based strategy, Nvwa, to predict gene expression landscapes and decipher *cis*-regulatory elements (CREs) at the single-cell level. We systematically compared cell type-specific transcription factors (TFs) and CREs to reveal conserved genetic regulation among vertebrates and invertebrates. Our work provides a valuable resource and a novel strategy for studying regulatory language in diverse biological systems.

Previously, we used Microwell-seq to construct organism-wide cell landscapes for humans and mice[1,2]. In this study, we aimed to build cell landscapes for zebrafish, *Drosophila*, and earthworm using a whole-body single-cell RNA-seq (scRNA-seq) strategy that could eliminate tissue-specific batch effects (Fig. 1a). Our analyses included dissociated 24 hours post fertilization (hpf), 72hpf, and 90-day zebrafish whole bodies (635,228 cells) (Fig. 1b, Extended Data Fig. 1a), 11- and 16-day *Drosophila* whole bodies (276,706 cells) (Fig. 1c, Extended Data Fig. 2a), and 6- and 8-mouth earthworm whole bodies (95,020 cells) (Extended Data Fig. 3a). The single-cell transcriptomic data were processed using published pipelines[3], and the information regarding data quality was summarized in Supplementary Table 1.

Unsupervised clustering of landscape data revealed 105 major zebrafish cell types (Fig. 1b), 87 major *Drosophila* cell types (Fig. 1c), and 62 major earthworm cell types (Extended Data Fig. 3a) with distinct gene expression programs (Supplementary Table 1). We annotated each cell type based on the expression levels of canonical cell type-specific markers. Totally, 105 zebrafish cell types were divided into 11 major cell lineages, including endothelial, epithelial, erythroid, germ, hepatocyte, immune, muscle, neuron, secretory, mesenchymal, and other cells (Extended Data Fig. 1b,c). Specifically, with high expression of *piwil1* and *hmgb2b*, cluster 7 (C7), C95, C96, and C97 were defined as germ cells. C1, C23, C41, C46, C48, C68, and C92 were defined as neurons according to *elavl3* and *snap25a* expression, and C18, C21, C37, and C57 were identified as neural progenitor cell according to *neurod1* and *hes6* expression. Epithelial cells were composed of multiple clusters including enterocyte cell (*fabp6*), epithelial cell (*epcam*), goblet cell (*arg2* and *krt18*), intestinal cell (*cftr*), ionocyte (*atp1a1a.4*), and pancreatic cell (*prss1*). Muscle cells were composed of cardiomyocyte (*myl7* and *mb*), mucosal muscle cell (*myh11b*), and smooth muscle cell (*acta2*, *myl6*, *tagln2*, and *myl9a*). We also annotated endothelial cells (C24), erythroid cells (C34 and C65), and neurosecretory cells (C51) according to *plk2b*, *hbae3*, and *scg3* expression, respectively (Extended Data Fig. 1c). Of note, about 96.2% cell types were consistent with available tissue-specific annotations[4] (Extended Data Fig. 1d). Finally, we performed subclustering analysis for each of the 105 major cell types and predicted a total of 1,285 cell-type subclusters in the hierarchy (Fig. 1d). For example, neurons (C1) could be further annotated into neural progenitor cell,

44 notochord, and retinal ganglion cell according to *otpa*, *neurod1*, *slit3*, and *irx4a* gene
45 expression in a higher resolution sub-clustering (Extended Data Fig. 1e-f).

46 In *Drosophila* cell landscape, the 87 cell types were divided into 12 major cell
47 lineages (Extended Data Fig. 2b,c), including epithelial, neuron, hemocytes, follicle,
48 gut, germ, male accessory gland (MAG), malpighian tubule (MT), muscle,
49 proliferating, fat body, and other cells. C1, C14, C21, and C81 were defined as
50 neurons according to known markers such as *noe*, *elav*, and *Appl.* C41 and C68 were
51 further identified as photoreceptor cells with high expression of *ninaE* and *trpl*. C11,
52 C27, C62, and C85 were annotated as MAG using *EbpII* and *Acp95EF*. We also
53 annotated hemocytes (C10, C39, C75, and C76), MT cells (C34 and C69), fat body
54 cells (C5, C74 and C77), and muscle cell (C36, C46, C51, and C58) according to *Hml*
55 and *Pxn*, *Mur18B*, *Fbp1*, and *Mlc* expression, respectively (Extended Data Fig. 2c).
56 We identified multiple gut-related cell types such as anterior enterocyte (marked by
57 *alphaTry* and *Amy-p*), middle enterocyte (marked by *Vha100-4*), posterior enterocyte
58 (marked by *lambdaTry*), enteroendocrine cells (marked by *AstC* and *AstA*), and
59 enteroblast (marked by *Sox100B*). In addition, we also compared our single-cell
60 landscape with a parallel fly cell atlas project[5] using MetaNeighbor. Among 87
61 *Drosophila* cell types, about 93.1% were consistent with tissue-specific annotations
62 (Extended Data Fig. 2d). Finally, the subclustering analysis was performed for each of
63 the 87 major cell types and predicted a total of 1,085 subclusters in the hierarchy (Fig.
64 1e). For example, central nerve cell (C1) could be further annotated into

glutamatergic neurons, cholinergic neurons, and Kenyon cells according to *VGlut*, *VAChT*, and *sNPF* gene expression in a higher resolution sub-clustering (Extended Data Fig. 2e,f).

For earthworm cell landscape, 62 cell types were divided into 8 major cell lineages (Extended Data Fig. 3a,b), including digestive gland, epithelial, neuron, coelomocytes, muscle, erythrocruorin, germ, and other cells. For example, digestive gland cells were identified mainly based on *FIBC*, and coelomocytes were identified based on *LYS* and *CHIT1*. Muscle cells and neuronal cells were defined according to markers *ACT1/2* and *7B2*, respectively (Extended Data Fig. 3b). Finally, the subclustering analysis predicted a total of 462 subclusters in the earthworm hierarchy, and these subclusters made sense functionally (Extended Data Fig. 3c-e). For example, neuron (C24) could be further annotated into distinct subclusters according to *ACHA2*, *NEURM*, and *NF70* gene expression (Extended Data Fig. 3d,e).

The landscape resource of zebrafish, *Drosophila*, and earthworm is available at http://bis.zju.edu.cn/nvwa/. Of note, we observed significant immune gene activity in structural cells (nonimmune cells), such as zebrafish epithelial cells (C2, C33, C44, C47, C53, and C66), *Drosophila* MT cells (C69), and earthworm coelomocytes (Extended Data Fig. 4a-c). The subclustering analysis of *Drosophila* MT cells showed that Cluster 0 (MT_*irk2*-high) encodes *Listericin*, the induction of which was cooperatively regulated by the product of *PGRP-LC* (Extended Data Fig. 4d,e, and Supplementary Table 1), suggesting that MT cells might be involved in the defense

response to Gram-negative bacteria[6]. We and another group have shown that mammalian structural cells, including epithelial, endothelial, and mesenchymal cells, might possess immune signatures[1,7].

Studying and modeling the gene regulation pattern at the organism level has always been a challenge in biology. With consistent single-cell mRNA-seq platform, our cell landscape data resource provided an unprecedented opportunity to investigate the genetic regulation of cell taxonomies across species. Therefore, we aimed to dissect the cell type-specific genetic regulation network and assessed the conservation of genetic regulation across species through data mining and machine learning. To obtain high-quality cells, we set a higher cutoff to construct high-quality cell landscapes (Extended Data Fig. 5a-c and Supplementary Table 2). We used similar cutoffs to select data from the Human Cell Landscape[1], the Mouse Cell Atlas[2] (Extended Data Fig. 5d,e, and Supplementary Table 2), and published atlases of *Ciona*[8], *C. elegans*[9], and planarians[10]. In total, we obtained 480 cell types from eight species (Supplementary Table 2), covering major cell lineages, including epithelial, immune, neuron, mesenchymal, muscle, secretory, erythroid, germ, endothelial, and proliferating lineages. We then used the Markov affinity-based graph imputation of cells (MAGIC) algorithm[11] to denoise the cell count matrices, fill in missing genes, and improve expressed gene numbers (Extended Data Fig. 5f,g).

To examine the cross-species comparability of cell types, pairwise SAMap analyses[12] were performed on the eight transcriptomic datasets. We compared the results of cross-species comparison among three types of datasets, including MAGIC,

single-cell, and pseudo-cell, and found consistent results that vertebrate cell types were conserved (Extended Data Fig. 6a). Based on MAGIC datasets, 85.9% (122 of 142) homologous cell-type pairs could be re-identified based on single-cell and pseudo-cell datasets (Extended Data Fig. 6b). In total, across eight species, there are 1,209 homologous cell-type pairs (Extended Data Fig. 6c-e), of which 36.5% (441 of 1,209) could be defined by MetaNeighbor analyses using 1-to-1 orthologous gene. To reduce the false-positive rate of our results, we set strict thresholds to construct cross-species mapping (Methods). Notably, vertebrate cell types were conserved, particularly for immune, mesenchymal, neuron, epithelial, endothelial, and germ cells (Extended Data Fig. 6f). In *Drosophila* cross-species mapping, we also found that almost all cell types from the same cell lineage showed strong links. To furtherly validate the results of cross-species mapping, we performed functional enrichment analysis on enriched gene pairs between homologous cell types. We found that enriched gene pairs shared consistent functions, such as in the muscle and neuron (Extended Fig. 6h), which is consistent with our previous study[13].

To assess the regulatory conservation and divergence of cell types between vertebrates and invertebrates at the regulatory level, transcription factor (TF) specificity scores were calculated for each species (Fig. 2a-h). Generally, we identified a total of 2,342 cell lineage-specific TFs among eight species (Supplementary Table 3). According to the conversion of homologous genes among eight species, we observed more conserved features from homologous TFs (Extended Data Fig. 7a). As shown in Fig. 2i, vertebrates' neuronal cells shared common TFs

such as *NEUROG/NEUROD* family, *OLIG1/2*, *POU3/4* family, and *SOX* family. Homologous TFs cover 91.42% (64 of 70), 98.75% (79 of 80), and 75% (78 of 104) in human, mouse, and zebrafish, respectively. Although human/vertebrates shared more conserved regulators compared to human/invertebrates (Extended Data Fig. 7b), we found that *NEUROG*1 and *NEUROG2* are conserved genes across eight species. As showed in the Extended Data Fig. 7c, there was a comprehensive homologous link from lower to higher organisms. Our results suggested that conserved regulatory programs are underlying neuronal cells. Moreover, we observed conserved TF markers for vertebrate immune cells (Extended Data Fig. 7d). For example, *LYL1*, *BATF*, *IKZF1*, and *STAT4* were highly associated with T cells, and *PLEK*, *IRF5*, and *SPI1* were highly associated with macrophage cells. *Drosophila* hemocytes shared common homologous TFs with human immune cells (Extended Data Fig. 7e). We identified 11 conserved TFs in *Drosophila* hemocytes such as *Blimp-1* (ortholog of *PRDM1*). *PRDM1* has been reported as a key TF with differences in expression during mammalian normal B cell differentiation[14]. In summary, our cross-species screening of conserved lineage regulators provides detailed information on conserved genetic regulation.

At the center of the genetic network, TFs recognize specific DNA sequences to control chromatin and transcription. However, it remains challenging to determine how regulatory elements relate to cell type-specific regulation. To understand the regulatory elements encoded in the genome, we aspired to develop a deep-learning based strategy, Nvwa (the name of a mother god in ancient Chinese legend). For an input DNA

sequence, Nvwa was able to predict its corresponding gene's expression and interpreted the learned sequence rules in individual cells. Furthermore, Nvwa identified the associations between specific cell types and deep learning derived *cis*-regulatory elements (CREs) (Fig. 3a). Briefly, Nvwa could learn major cellular equilibrium states solely using gene regulatory sequences at the single-cell level (Fig. 3a).

We first trained a set of Nvwa models for eight species independently, and evaluated whether Nvwa could predict single cell gene expression correctly. The accuracy of Nvwa expression predictions were assessed using average area under the receiver operating characteristic (AUROC) and area under the precision-recall curve (AUPR) on the held-out test data. Nvwa robustly predicted gene expression with an overall AUROC of 0.78 and AUPR of 0.59 across eight species (Extended Data Fig. 8). For example, the model trained in *Drosophila* achieved an average AUROC of 0.82 and AUPR of 0.69 among 77,337 cells in the held-out gene set (Extended Data Fig. 8). By comparing the performance between cells, we found that cells of the germ lineage always had the highest expression predictive correctness, with *C. elegans* germ cells reaching an average AUROC of 0.93 (Extended Data Fig. 8a). Of note, comparison results from single-cell datasets showed that Nvwa outperformed standard architectures including Basset, DeepSEA, Beluga, and Basenji in predicting single-cell expression (Extended Data Fig. 9a). Additionally, Nvwa model accuracies could be further improved with multiple genome training (Extended Data Fig. 9b). Furthermore, our model predictions recapitulated the cell relationship, including cell-type similarity and diversity, with predictions being substantially more similar to

the same cell type (Extended Data Fig. 9c). The cell-type specificity was further confirmed by the *t*-SNE embedding and average mutual information (AMI) score of the predicted expression landscape on the held-out genes (Extended Data Fig. 10). Overall, these evaluations confirmed that Nvwa could correctly predict gene expression at the single-cell level from DNA sequence.

Nvwa could be further extended to predict genome-wide transcriptional activity signals by scanning sequence along the held-out chromosome. Inspecting Nvwa genome-wide predictions, we observed that they were correlated with experimental functional genomics data. Take mouse neuronal cells for an example, Nvwa genome-wide predictions received average Spearman correlation coefficients (SCC) of 0.57 with DNase-seq data, average SCC of 0.47 with single-cell chromatin accessibility data, and average SCC of 0.67 with Histone ChIP data (Extended Data Fig. 11a). Furtherly, we performed the permutation test to estimate the overlap between Nvwa prediction and experimental data (Extended Data Fig. 11b). Nvwa highlighted regions were significantly concordant to peaks regions of experimental functional genomics data, as well as ENCODE CRE regions (Bonferroni adjusted p-value < 0.05). By visualizing the genome browser tracks, we observed consistency between Nvwa predictions and experimentally defined signals in multiple cell types and species (Fig. 3b and Extended Data Fig. 12). In addition, we found that the inconsistency may be caused by the independent sources of evidence. For example, Nvwa correctly predicted transcriptional signals at the minus strand of 5S rDNA regions (Fig. 3b and Extended Data Fig. 12a), while these highly duplicated tandem

196 repeats[15,16] were difficult to detect in the high-throughput sequencing experiments[17].
197 Our analysis externally validated the robustness of Nvwa predictive performance, and
198 indicated the central role of functional DNA sequences.

199 To understand why Nvwa can predict single-cell gene expression correctly, we
200 examined the sequence patterns that model learned. As expected, proximal promoter
201 regions were most informative to Nvwa model by systematical shifting of the input
202 sequences (Extended Data Fig. 13a). We then extracted the deep learning-based CREs
203 from each first-layer convolution filter (Filter) based on feature map and
204 TF-MoDISco[18] methods. In *Drosophila*, 83.3% (60 of 72) motifs discovered in
205 TF-MoDISco were identical to feature map-based Filters (Extended Data Fig. 13b).
206 We calculated the cell types specificity of CREs, which were quantified using the
207 influence score (prediction change in individual cells by in silico nullification of the
208 Filters). Of note, Filters, same as TFs, were also involved in cell-type identity and
209 cellular activity (Fig. 3c). These results inspired us to further analyze the model filters
210 and their relationship to cell types specific CREs and TFs.

211 Nvwa derived CREs could be assigned to known TF binding sites (TFBSs). For
212 example, by querying the RcisTarget database[19] with TomTom algorithm[20], 65.63%
213 (84 out of 128), 63.28% (81 out of 128) and 60.94% (78 out of 128) Filters could be
214 annotated to TFBSs for human, mice, and *Drosophila*, respectively. We also observed
215 that annotated Filters showing high similarity with known TFBS (Fig. 3d, Extended
216 Data Fig. 13c). In the cross-validation analysis, the majority of annotated Filters has
217 high reproducibility and information content, indicating the robustness of Nvwa

interpretation (Extended Data Fig. 13d). Some unannotated Filters with high influence scores possibly captured short sequence patterns. TATA box like F114 in zebrafish-specific model and CGCG box like F24, F29, and F35 in *Drosophila*-specific model were further identified to have overall positive influence on transcriptional activity (Extended Data Fig. 13d, Supplementary Table 3).

Besides the biological annotation, the cell type specificity of Nvwa derived CREs were further inspected. We found that the cell-type specific Filters were consistent with known roles of the corresponding TFs (Fig. 3e,f, Extended Data Fig. 14, and Supplementary Table 4). For example, mouse-specific model showed that F6 (*Gata1*), F28 (*Gata1*), F84 (*Gata1*) and F59 (*Bcl11a*) were enriched in erythroid cells, and F48 (*Foxg1*) was in neuronal cells. In zebrafish, we identified that F38 (*foxq1a* and *foxq1b*) were enriched in epithelial cells, F64 (*sox10*) and F101 (*foxp2*) in neuronal cells, and F44 (*irf4a* and *irf4b*) in immune cells (Extended Data Fig. 14). For *Drosophila*-specific model, F7 (*dati/rn/FoxP*), F54 (*l(1)sc*), and F85 (*FoxP*) were identified to be enriched in neuronal cells, F42 (*cad/Sox100B*) for gut cells, F44 (*lmd*) and F54 (*nau*) in muscle cells (Fig. 3f). A previous single-cell transcriptome and chromatin accessibility analysis also decoded *Dati* gene and motif (AAAAAA) as important regulators of *Drosophila* neuronal cell types[21]. We also identified Filter regulons (a set of genes enriched the corresponding CRE) in *Drosophila*, which validated that target genes regulated by the same Filter had similar cell lineage-specific expression patterns (Fig. 4a,b). Altogether, Nvwa leveraged the deep

239 learning derived CREs with TFs related to specific cell types, which enabled us to
240 screen cell-type specific regulators directly from sequence (Extended Data Fig. 14).
241 To further explore more complex sequence rules associated with cell type, we
242 performed global motif analysis on Nvwa models. We found that Filter pairs show a
243 higher correlation with cell type-specificity in vertebrates (Wilcoxon rank-sum test,
244 p-value < 0.05) (Fig. 4c,d, Extended Data Fig. 15, and Supplementary Table 5). To
245 further inspect possible rules in *Drosophila*, we pruned the cooperative Filter
246 interactions with significant target genes overlapped (p-value < 0.01, normal
247 distribution one-tailed test). We identified Filter syntax rules including positive
248 influence enhances, negative influence enhances, influence offset, and negative
249 influence offset at single-cell level (Extended Data Fig. 16a,b). For example,
250 cooperative Filter interactions between F85 (*FoxP*) and F82 (*pho*) significantly offset
251 the positive influence of F85 and negative influence of F82 (both p-value < 1e-20,
252 Wilcoxon rank-sum test). Our results showed the applicability of Nvwa in identifying
253 complex regulatory rules, and provided a practical solution to decode major cellular
254 equilibrium states in genome sequence.
255 To further analyze the cross-species genetic network (Fig. 2a), we compared the
256 conservation and divergence of deep-learning based CREs across species. We
257 identified a total of 663 cell type-specific Filters among eight specie-specific models
258 (Supplementary Table 6). And 94.9% cell-type specific Filters were identified to be
259 homologous to at least one Filter from other specie-specific models (Fig. 4e).
260 Homologous Filters tend to preserve similar cell-type specificity across species

(Extended Data Fig. 16c,d). For example, F101 (*foxp2*) in zebrafish, F7 and F85 (*FoxP*) in *Drosophila*, and F117 (*fkh-2*) in *C. elegans* were enriched in neuron lineage, indicating the conservation of *Fox* binding domain and their function[22] (Fig. 4f,g). F54 (*EGR2*) in human and its homologous F29 (*Egr2*) in mouse, F85 (*egr2a* and *egr2b*) in zebrafish were enriched in immune lineage, which was consistent with our previous TF module analysis. The cross-species comparison of deep-learning derived CREs showed promising potential in revealing the enrichment of conserved regulators underlying specific cell types.

In this study, we performed whole-body single-cell transcriptomic analysis to measure the entire cellular equilibrium states of zebrafish, *Drosophila*, and earthworm. Cross-species analysis of eight metazoan cell landscapes suggested the cell taxonomies conservation and diversity during evolution. To distinguish convergent evolution from concerted evolution[23,24], we screened cell lineage-specific TFs underline cell type specificity. Importantly, we developed a deep learning-based framework, Nvwa, to predict expression solely from genome sequences and reveal key regulatory elements at the single-cell level. We leveraged the deep learning derived CREs with TFs related to specific cell states and screened cell-type specific regulators directly from sequence. Overall, our work reveals the role of conserved regulatory programs underlying major cellular equilibrium states, and will inspire future multi-species regulatory landscape projects.

Methods

Biological samples

Wild-type *Danio rerio* (zebrafish) strain AB was raised and maintained in standard zebrafish units at Core Facilities, Zhejiang University School of Medicine. The w^{1118} strain of *Drosophila melanogaster* (*Drosophila*) was cultured under the laboratory conditions at around 25°C and 60% relative humidity under an alternating light/dark photoperiod. Live *Eisenia andrei* (earthworm) originated from Guangxi province in China.

Fabrication of the microwell device

The diameter and depth of the microwells were 28 and 35 μm, respectively. First, a silicon plate containing 100,000 microwells was manufactured by Suzhou Research Materials Microtech Co., Ltd. (Suzhou, China). The silicon microwell plate was then used as a mold to make a polydimethylsiloxane (PDMS) plate with the same number of micropillars. Before the experiments, a disposable agarose microwell plate was made by pouring 5% agarose solution onto the surface of the PDMS plate. Both the silicon and the PDMS plates were reusable. One silicon microwell plate allows almost permanent use.

Synthesis of barcoded beads

Magnetic beads (20–25 mm in diameter) coated with carboxyl groups were provided by Suzhou Knowledge & Benefit Sphere Tech. Co., Ltd. (Suzhou, China; http://www.kbspheretech.com/). The barcoded oligonucleotides on the surface of the

beads were synthesized by three rounds of split pooling. All the sequences used were the same as those reported previously. All the sequences have been listed in Supplementary Table 1.

For each batch of bead synthesis, 300 to 350 μl carboxyl magnetic beads (50 mg/ml) were washed twice with 0.1 M 2-[N-morpholino] ethanesulfonic acid (MES). The beads were then suspended in a final volume of 635 μl of 0.1 M MES. 1-Ethyl-3-(3-dimethyl aminopropyl) carbodiimide hydrochloride (EDC; 3.08 mg) was added to the beads, and 6.2 μl beads were then placed in each well of a 96-well plate. Amino-modified oligonucleotides (2.5 μl, 50 μM in 0.1 M MES) were then added to each well. After vortexing the mixture and incubating it for 20 min at ambient temperature, a 0.5 μl mix (6 mg EDC in 100 μl of 0.1 M MES) was distributed into each well. After an additional round of vortexing and incubation for 20 min at ambient temperature, an additional 0.5 μl mix (6 mg EDC in 100 μl of 0.1 M MES) was distributed into each well. After vortexing and incubation for 80 min at ambient temperature, the beads were collected in 1 ml of 0.1 M phosphate-buffered saline (PBS) containing 0.02% Tween 20. After centrifugation, the supernatant was carefully removed. The beads were then washed twice in 1 ml TE (pH 8.0).

In the second split-pool, the beads were washed with water and divided among the wells of another 96-well plate containing polymerase chain reaction (PCR) mix (1× Phanta Master Mix, Vazyme) and 5 μM oligonucleotides. The oligonucleotides in each tube encoded a sequence with reverse complementarity to linker 1, a unique barcode and a linker 2 sequence. The PCR program was as follows: 94°C for 5 min;

five cycles of 94°C for 15 s, 48.8°C for 4 min, and 72°C for 4 min; and a 4°C hold. The third split-pool procedure was the same as the second one. The PCR program was as follows: 94°C for 5 min, 48.8°C for 20 min and 72°C for 4 min, and a 4°C hold. The beads were mixed sufficiently between denaturation (95°C) and primer annealing (48.8°C) in every cycle. The oligonucleotides used in each tube encoded a linker 2 reverse-complementary sequence, a unique barcode, a UMI sequence and a poly-T tail. All oligonucleotides were synthesized by Sangon Biotech Co., Ltd. with high-performance liquid chromatography purification. To remove the chains without the third barcoded sequence, the beads were collected and suspended in 200 μl exonuclease I mix (containing 1× exonuclease I buffer and 1 U/μl exonuclease I) and incubated at 37°C for 15 min (the beads were mixed by a rotary mixer). After being washed with 200 μl TE-TW and 200 μl of 10 mM Tris-HCl (pH 8.0), the beads were resuspended in 1 ml double-distilled water. To remove complementary chains, the beads were placed in a 95°C water bath for 6 min and separated using a magnet, removing the supernatant quickly, two times. The beads could be stored in TE-TW [10 mM Tris (pH 8.0), 1 mM EDTA, 0.01% Tween 20] for 4 weeks at 4°C.

Cell preparation

Zebrafish whole bodies were minced into pieces (~1 mm) at room temperature using scissors. The tissue pieces were transferred to a 15-ml centrifuge tube and suspended in 3 ml 1x TrypLE containing collagenase I (0.25 mg/ml), collagenase II (0.25 mg/ml), collagenase IV (0.25 mg/ml) and collagenase V (0.25 mg/ml). During dissociation, the tissue pieces were pipetted up and down gently several times until no

tissue fragments were visible. The tissue pieces were digested at 30°C for 30 minutes. The dissociated cells were centrifuged at 500×g for 5 min at 25°C and then resuspended in 3 ml DPBS. After passage through a 40-μm strainer (Biologix), the cells were washed twice, centrifuged at 300×g for 5 min at 25°C, and resuspended at a density of 1×105 cells/ml in DPBS containing 2 mM EDTA.

For every 10 *Drosophila*, 30 μl Drosophila medium was added, centrifuged and minced into pieces (~1 mm) at room temperature using scissors. The tissue pieces were transferred to a 1.5-ml Eppendorf tube, 1 ml Drosophila medium was added, and the tube was centrifuged at 5000 rpm for 3 min at 25°C. The supernatant was discarded, then 1 ml Drosophila medium was added, and the tube was centrifuged at 5000 rpm for 2 min at 25°C, and the supernatant was discarded. Then, 500 μl of mixed enzyme (10 mg/ml collagenase I in 1x TrypLE) was added to the pellet and digested for 50 minutes at 1000 rpm in a metal bath at 28°C. Then, 20 μl of 10 mg/ml DNase I was added to the tube and digested for 10 minutes at 1000 rpm in a metal bath at 28°C and then pipetted up and down 100 times. The suspension was transferred to a 15-ml centrifuge tube, 5 ml PBS was added, and the tube was centrifuged at 500×g for 5 min. The supernatant was completely discarded, and the residue was resuspended with DPBS containing 0.05% BSA.

Earthworm whole bodies were washed using PBS and minced into pieces (~1 mm) at room temperature using scissors. The tissue pieces were transferred to a 1.5-ml Eppendorf tube and suspended in 1 ml 1x TrypLE containing collagenase I (5 mg/ml) and digested for 1.5-2 hours at 1000 rpm in a metal bath at 30°C. During dissociation,

the tissue pieces were pipetted up and down gently several times until no tissue fragments were visible. The suspension was transferred to a 15-ml centrifuge tube, 5 ml PBS was added, and the tube was centrifuged at 500×g for 5 min. The supernatant was completely discarded, and the residue was resuspended with PBS containing 2 mM EDTA

Cell collection and lysis

The cell concentration should be carefully controlled during Microwell-seq. Both cell and bead concentrations were estimated using a haemocytometer. The beads were resuspended in PBS before loading. The proper cell concentration is ~200,000/mL (with 10% of the wells occupied by single cells). The proper bead concentration is ~1,000,000/ml (with every well occupied by single beads). An evenly distributed cell suspension (~500mL) was pipetted onto the microwell array. To eliminate cell doublets, the plate was inspected under a microscope. Cell doublets were reduced by pipetting over the region of high cell density. After the expected number of cells were evenly and tightly dropped in the microwells, the supernatant containing the excess cells would be aspirated away and the bead suspension (~500mL) was then loaded into the microwell plate, and the plate was placed on a magne, with beads being quickly shaken into the microwells. Excess beads were washed away slowly. Cold lysis buffer [0.1 M Tris-HCl (pH 7.5), 0.5 M LiCl, 1% sodium dodecyl sulfate (SDS), 10 mM EDTA, and 5 mM dithiothreitol] was added to the surface of the plate and removed after 12 min incubation. The beads were then collected, transferred to an RNase-free tube, and washed once with 1 ml of 6× SSC, once with 500 µl of 6× SSC,

and once with 200 µl of 50 mM Tris-HCl pH 8.0. Finally, ~90,000 beads were collected in a 1.5 ml tube.

Reverse transcription

In this procedure, the instructions from the Smart-seq2 protocol were followed. Briefly, 20 µl RT mix was added to the collected beads. The RT mix contained 200 U SuperScript II reverse transcriptase, 1× Superscript II first-strand buffer (Takara), 40 U Murine RNase inhibitor (Vazyme), 1 M betaine (Sigma), 6 mM MgCl2 (Ambion), 2.5 mM dithiothreitol, 1 mM deoxynucleoside triphosphate, and 1 µM TSO LNA primer. The sequences information for the primers was included in Supplementary Table 1. The beads were incubated at 42°C for 90 min with mixing on a rotary mixer (10 rpm) and then washed with 200 µl TE-SDS (1× TE + 0.5% SDS) to inactivate reverse transcriptase.

Exonuclease I treatment

The beads were washed once with 200 µl TE-TW and once with 200 µl of 10 mM Tris-HCl (pH 8.0), resuspended in 100 µl exonuclease I mix containing 1× exonuclease I buffer and 50 U exonuclease I (NEB), and incubated at 37°C for 60 min with mixing on a rotary mixer (10 rpm) to remove oligonucleotides that did not capture mRNA. The beads were then pooled and washed once with TE-SDS, once with 1 ml TE-TW, and once with 200 µl of 10 mM Tris-HCl (pH 8.0).

Second Strand Synthesis

The beads were resuspended and incubated in 200 uL 0.1 M NaOH at room temperature for 30 seconds twice to thoroughly denature the mRNA-cDNA hybrid

and remove mRNAs. Following denaturing, the NaOH was removed and beads were washed twice with 1mL TE-TW and once with 200 μL of 10 mM Tris-HCl (pH 8.0). The beads were resuspended in 20 μL dn-TSO oligo mix containing 5mM dn-TSO oligo, 10 mM Tris-HCl (pH 8.0) and 3mM $MgCl_2$. The beads were then incubated in 95°C water bath for 30 seconds followed by mixing on a rotary mixer (10 rpm) at room temperature for 10 min to promote the binding of dn-TSO oligos (The sequences information is included in Supplementary Table 1) to multiple sites of the single chains on beads. After that, the supernatant was aspirated away, the beads were washed once with 100 μL of 10 mM Tris-HCl (pH 8.0) without being resuspended, and then combined with a 50ul mastermix consisting of 1x RT buffer（Thermo）, 12% PEG8000 solution, 1mM dNTPs, and 5 μL Klenow Exo- (Vazyme). Second-strand synthesis was carried out by incubating the beads for 1 hour at 37°C on a rotary mixer (10 rpm). The beads were then washed twice with 200 μL TE-TW and once with 10 mM Tris-HCl (pH 8.0).

cDNA amplification

The beads from the same plate were transferred into one PCR tube. To each tube, 12.5 μl PCR mix [1× HiFi HotStart Readymix (Kapa Biosystems) and 0.1 μM TSO_PCR primer] was added. The PCR program was as follows: 98°C for 3 min; six cycles of 98°C for 20 s, 65°C for 45 s, and 72°C for 6 min; 72°C for 5 min; and a 4°C hold. After pooling all PCR products, 50 μL PCR mix [1× HiFi HotStart Readymix and 0.4 μM TSO_PCR primer (The sequences information is included in Supplementary Table 1) was added to each DNA sample. The PCR program was as follows: 95°C for

3 min; 12 cycles of 98°C for 20 s, 67°C for 20 s, and 72°C for 3 min; 72°C for 10 min; and a 4°C hold. Finally, 0.9X VAHTS DNA Clean beads (Vazyme) were used to purify the cDNA library.

Transposase fragmentation and selective PCR

The purified cDNA library was fragmented using a customized transposase that carries two identical insertion sequences. The customized transposase was included in the TruePrep Homo-N7 DNA Library Prep Kit for Illumina (Vazyme) or TruePrep Homo-N7 DNA Library Prep Kit for MGI (Vazyme). The fragmentation reaction was performed according to the instructions provided by the manufacturer. We used our own P5 primer (The sequences information is included in Supplementary Table 1) and VAHTS RNA Adapters set 3-set 6 for illumine (Vazyme) or our P7 primers （N8××，the sequences are listed in Supplementary Table 1） to specifically amplify fragments that contain the 3’ ends of transcripts. Other fragments will form self-loops, impeding their binding to PCR primers. The PCR program was as follows: 72°C for 3 min; 98°C for 1 min; five cycles of 98°C for 15 s, 60°C for 30 s, and 72°C for 3 min; 72°C for 5 min; and a 4°C hold. The PCR product was purified using 0.9X VAHTS DNA Clean beads. Then, 25 μL PCR mix (1×HiFi HotStart Readymix and 0.2 μM 2100 primer) was added to each sample. The PCR program was as follows: 95°C for 3 min; five cycles of 98°C for 20 s, 60°C for 15 s, and 72°C for 15 s; 72°C for 3 min; and a 4°C hold. To eliminate primer dimers and large fragments, 0.55-0.15X VAHTS DNA Clean beads were then used to purify the cDNA library. The size distribution of the products was analysed on an Agilent 2100 bioanalyser, and a peak in the 400 to

700 bp range was observed. Finally, the samples were subjected to sequencing on the Illumina HiSeq or MGI DNBSEQ-T7. For MGI sequencing, we applied the protocal provided by VAHTS Circularization Kit for MGI (Vazyme) to obtain single-stranded circular cDNA available for DNB (DNA Nanoball) generation. We also replaced the official R1 sequencing primers with our custom R1 sequencing primers A&B to ensure the completion of the sequencing.

Pre-processing of Microwell-seq sequence data

Microwell-seq datasets were processed using the protocols described by[2]. Reads from three species were aligned to the *Danio rerio* GRCz11 genome, *Drosophila melanogaster* BDGP6.28 genome, and earthworm GWHACBE00000000 genome from Genome Warehouse using STAR (v3.1)[25]. The DGE data matrices were obtained using the modified Drop-seq tools (https://github.com/ggjlab/mca_data_analysis/tree/master/preprocessing/Drop-seq_tools-1.12/) and the corresponding protocol is available at http://mccarrolllab.org/dropseq/. The DGE data containing the top 10,000 cells sorted by the total number of transcriptions were obtained after this pre-procession.

Correction of the RNA contamination

For quality control, we filtered out cells with the detection of fewer than 500 transcripts and 200 genes. During scRNA-seq data analysis, we commonly assumed that all RNAs were endogenous to each individual cell. Some contaminating non-endogenous RNAs have also been recognized to be present also even within datasets of the highest quality[26]. While during applying the Microwell-seq technology,

there was a probability that a certain amount of cell free RNAs from the input solution, admixed to a cell in a well, were also sequenced. Although the ratio of these cell-free RNAs was very slow compared with the whole RNA content, it is still important to recognize and correct them. To solve this problem, we strictly removed the batch gene background. We assumed that for each batch of experiments, cell barcodes with fewer than 300 UMIs correspond to empty beads exposed to free RNA during cell lysis, RNA capture and washing steps. Genes with extensive expression in all beads were considered batch genes. The batch gene background value was defined as the average gene detection for all cellular barcodes with fewer than 300 UMIs multiplied by the median of the fold difference between the detected gene expression of a cell and the average detected gene expression for beads with fewer than 300 UMIs; this value was then rounded to the nearest integer. We subtracted the batch gene background for each cell from the DGE before performing the cross-tissue comparison. We used the background-removed matrix to perform cross-tissue analyses and all cells over three species.

Filter out potential doublets

ScRNA-seq data are commonly influenced by technical artefacts known as doublets, where two or more different cells receive the same barcode, resulting in an aggregated transcriptome. We used the R package DoubletFinder (v2.0.3)[27] to detect the potential doublets from each individual library. Approximately 5% of cells were labelled doublets and were removed.

Clustering analysis for single-cell datasets on all dataset bases

In zebrafish and *Drosophila*, the gene expression matrices for all single cells were merged together and fed to Scanpy[28] for clustering analysis. We chose 50 PCs for PCA and computed the neighborhood graph of cells. We then used the Leiden clustering to cluster cells with a resolution of 2. Finally, 105 clusters for zebrafish and 87 clusters for *Drosophila* were produced. *t*-SNE was applied to visualize the single-cell transcriptional profile in 2D space. The Wilcoxon rank-sum test was used by running the "rank_genes_groups" function in Scanpy to find differentially expressed genes in each cluster. We annotated each cell type by marker genes with extensive literature reading.

In earthworm, the gene expression matrix for all single cells was merged together and fed to Seurat (v3.9.9)[29]. After quality control filtering, data were normalized using the "SCTransform" function with default parameters. Then, processed data was fed to Scanpy. Subsequently, standard cell clustering and visualization were performed. We obtained the homologs of 21,716 earthworm genes from Shao et al., study[30]. According to the functions of homologs, we finished cell type identification for earthworm cell landscape (Extended Data Fig. 1).

For sub-clustering of clusters, we used the Seurat pipeline described previously with the default resolution of 0.8 to process cells from each cluster and identified a total of 1,285, 1,085 and 462 sub-clusters for zebrafish, *Drosophila*, and earthworm, respectively.

scRNA-seq data imputation

The matrix was normalized with $log2(\frac{CPM_{i,j}}{10}+1)$ transform, where $CPM_{i,j}$ refers to counts per million for gene *i* in cell *j*, and we kept only genes with nonzero expression values in at least 3 cells. Then, we performed the Markov affinity-based graph imputation of cells (MAGIC) algorithm (v2.0.3)[11] for imputation of the normalized matrix, with the recommended settings from its GitHub repository.

Cell clustering for single-cell datasets on high-quality dataset basis

We started collecting high-quality (average of approximately 1,000 genes/cells) data from background-removed single-cell digital gene expression matrices (DGEs) of human, mouse, zebrafish, and *Drosophila*. Seurat (v3.9.9)[29] was used as a tool for cell clustering on a high-quality dataset basis. After quality control filtering, data were normalized using the "SCTransform" function with default parameters for humans (~134,000 cells), mouse (~179,000 cells), zebrafish (~241,000 cells), and *Drosophila* (~77,000 cells). The top 3,000 highest standard deviations were obtained using the "SCTransform" function as highly variable genes. Subsequently, standard cell clustering and visualization were performed. Globally, we identified 97 human cell types, 98 mouse cell types, 88 zebrafish cell types, and 59 *Drosophila* cell types (Supplementary Table 2). In earthworm, we collected 29,609 high-quality cells (average of approximately 370 genes/cells). Standard cell normalization, clustering, and visualization were performed. normalized the read counts of each cell by Seurat. 30 cell types were identified in earthworm (Supplementary Table 2).

In addition, we collected 36 *Ciona* larva cell types[8] identified in our previous study[13], 28 *C. elegans* cell types[9] and 44 planarian cell types[10]. The MAGIC algorithm was adopted as a unified approach to denoise the cell count matrices, fill in missing genes and improve expressed gene numbers (Extended Data Fig. 4f,d).

Collection and of orthologous genes

The planarian transcriptome was downloaded from the PlanMine database[31] (planarian, dd_Smed_v6) and the *Ciona* transcriptomes were downloaded from Ghost database[32] (*Ciona*, KH gene models ver.2012). The protein coding sequences (CDSs) were predicted by TransDecoder (v5.3.0)[33] with default parameters. The earthworm CDSs were downloaded from the National Genomics Data Center (http://bigd.big.ac.cn/gwh/) (earthworm, GWHACBE00000000)[30]. Human, mouse, zebrafish, *Drosophila*, and *C. elegans* orthologous pairs were obtained from Ensembl by BioMart. Other orthologous pairs were predicted by OrthoFinder (v2.2.6)[34] with CDS files as the input. Here, we considered only 1-to-1 orthologous pairs for further analysis, while the one-to-many or many-to-many orthologous pairs were discarded.

MetaNeighbor analysis

To systematically assess the transcriptional similarity between cell types within species, we ran MetaNeighbor analysis[35] for each pair of cell atlases using python package pyMN[36]. Area Under the Receiver Operating Characteristic Curve (AUROC) scores were used to quantify the similarity of cell-type pairs. Hierarchical clustering trees based on AUROC scores were used to reveal the relationships within species (Fig. 1d,e, Extended Data Fig. 3c, and, Fig. 3c).

Cross-scale analysis

For the systematic cross-scale comparison, we downloaded two published tissue-wide single-cell datasets of zebrafish[4] and *Drosophila*[5]. We adopt pyMN for comparison and further label transferring. Heatmaps showed the correspondence between our single-cell whole-body landscapes and tissue-wide datasets (Extended Data Fig. 1d and Extended Data Fig. 2d). And we also provided these "Transfer_annotation" and "Mean_AUROC_Transfer_annotation" in the Supplementary Table 1.

Cross-species analysis

SAMap[12] was used to performed the cross-species analyses among eight organisms. Based on the self-assembling manifold (SAM) algorithm, SAMap enables reliably comparing species across long evolutionary distances, and identifies homologous cell types with shared expression programs across phylogenetically remote species. We first ran 'map_genes.sh' to construct a gene-gene bipartite graph between two species based on the BLAST hits between genes. And further screened homologous genes among eight species with a percentage of identical matches threshold of 30 (Supplementary Table 2). Alignments for each cell types were calculated by the 'get_mapping_scores' function. The mapping strength between cell types was measured by alignment score, which was defined as the average number of mutual nearest cross-species neighbors of each cell relative to the maximum possible number of neighbors. Enriched gene pairs from the aligned cell-type pairs were retrieved by 'gpf.find_all' function with an alignment score threshold of 0.1. Homologous cell-type pairs were identified with an alignment score threshold of 0.1 and the

number of enriched gene pairs threshold of 3. We identified 1,209 homologous cell-type pairs among eight species (Supplementary Table 2). In addition, to assure results correctness, we set strict thresholds with an alignment score threshold of 0.3 and the number of enriched gene pairs threshold of 15. The strict results were used to construct the cross-species mapping (Extended Data Fig. 6f,g).

Node-supporting TFs and Filters

Based on hierarchical clustering trees (Fig. 3c), cell types were divided into different sub-groups. We used "rank_genes_groups" function (Wilcoxon rank-sum test) in Scanpy to identify node-supporting TFs. Node-supporting TFs were selected for particular sub-groups of cell types as those TFs that were involved in top 100 markers in all the descendant tips of that particular node. In the same way, selected Filters supporting each node were indicated in red.

Cell type-specific TFs

In this study, TFs of humans, mice, zebrafish, *Drosophila* and *C. elegans* were downloaded from AnimalTFDB[37]. *Ciona* TFs were downloaded from Ghost database[32]. Planarian TFs were defined as the 516 genes annotated by the GO terms "transcription factor binding" or "transcription factor activity", downloaded from PlanMine[31]. Based on identified TFs in the above seven species, 1-to-1 orthologous TFs were selected as earthworm candidate TFs. In addition, we collected candidate TFs from Shao et al., study[30]. In total, 330 candidate TFs were defined in earthworm.

We applied the same approach as described in our previous study[13]. to infer cell type-specific TFs for eight species. Based on the TF expression status, Jensen–

Shannon divergence (JSD) was calculated. The TF specificity score was defined as $1-\sqrt{JSD}$. We then calculated the Z-score-normalized TF specificity score to predict the essential TFs in each cell type (Fig. 2 a-h).

Data processing for machine learning

The genome sequences of human (GRCh38), mouse (GRCm38), zebrafish (GRCz11) were downloaded from Ensembl. The genome sequence of *Ciona* (Joined-scaffold (KH) genome) was downloaded from Ghost database[32]. The genome sequence of *Drosophila* (BDGP6.28 genome) was downloaded from FlyBase database[38]. The genome sequence of earthworm (GWHACBE00000000) was downloaded from Genome Warehouse (http://bigd.big.ac.cn/gwh/). The genome sequence of *C. elegans* (GCA_000002985.3 genome) was downloaded from WormBase database[39]. The genome sequence of planarian (S2F2 genome) was downloaded PlanMine database[31]. The putative promoters were considered as 10 kilo-base (kb) sequence centered at the transcription start site (TSS). The promoter sequences were obtained according to gene coordinates using pysam, a Python interface to samtools[40].

Due to the zero-inflated negative binomial (ZINB) distribution characteristics of single-cell data, the gene expression patterns of the previously generated high-quality MAGIC datasets were binarized into labels. Considering the numbers of protein-coding genes and expressed genes among eight species, the different label cutoffs were used for expressed (label 1) and unexpressed (label 0) genes, with ~8,000 and ~5,000 expressed genes for vertebrates and invertebrates respectively

(Extended Data Fig. 4f). The corresponding label (gene expression vector) and input (one-hot encoding promoter sequence) were paired to generate the full dataset for deep learning.

For each species, the datasets for deep-learning have been divided into three sets, including training set, validation set, and testing set. For humans and mice, we left out all genes on chromosome 8 (chr8) for testing, and genes on the other chromosomes for training and validation (randomly split left out 1,000 genes). This approach[41] strictly excluded the sequence overlap between the training and testing sets. But for the other 6 species, due to the imbalance of gene numbers on different chromosomes or scaffolds, we randomly split left out 1,000 genes for testing, 1,000 genes for validation, and the remaining genes for training.

By uniformly processing the single cell expression data, we constructed high quality training datasets for machine learning. The data label can be accessed on Nvwa website (http://bis.zju.edu.cn/nvwa/).

Model architecture

We initialized the deep learning model architecture with convolutional blocks, fully connected feed-forward blocks and recurrent neural blocks. We applied the Tree of Parzen estimators (TPE) to optimize the hyperparameters of the model, including promoter region of input sequence, number of layers and hyperparameters of each neural block. We ranked the models based on the objective loss and manually inspected major architecture through systematic benchmarks.

650 The final model, which considered the region from 6.5 kb upstream to 6.5 kb
651 downstream of the TSS, was comprised of two blocks. The first was a convolutional
652 neural network (CNN) block and the second was a fully connected feed-forward
653 network (FFN) block.

654 Briefly,

655 $$f_t(x) = sigmoid(FFN_t(GAP(ReLU(Conv(x)))))$$

656 where $f_t(x)$ represents the predicted probability for cell target t given input x.

657 The CNN block contained two convolution layers. The first convolution layer
658 scanned a set of motif detectors along the sequence, transforming genomic sequences
659 to feature maps. Each motif detector was a 4×m weight matrix (m=7 in our model),
660 which could be considered a position weight matrix (PWM) of length m. In the
661 second convolution layer, each weight matrix reflected the degree of interaction
662 between motifs over the feature maps of the previous layer. After convolution
663 operations, outputs were then transformed by the rectified linear activation function
664 (ReLU), which clamped negative values to zero. The global average pooling (GAP)
665 layer computes the bin average of the rectified outputs across the sequence.

666 The FNN block consists of two linear transformations with a ReLU activation in
667 between. In the multitask learning (MTL) framework, hidden neurons of FNN block
668 were split according to cell type relationship (biological cell type annotation). MTL
669 framework of Nvwa was able to overcome the noise that might exist in a single cell
670 task and automatically learn to preferentially utilize sequence features from similar
671 and different cells. We fed the results into the FFN to produce predictions for the cell

672 expression state and scaled the predictions to expression probability by a sigmoid
673 function.

674 $$Sigmoid(x) = \frac{1}{1 + e^{-x}}$$

675 **Training Nvwa**

676 At the beginning of training, all parameters were initialized to random values using
677 "Kaiming initialization". A minibatch of N (N=32 in our model) randomly selected
678 input sequences was then fed through the network, generating predictions for each
679 individual cell state.

680 We used the mean of binary cross entropy loss (BCELoss) as the objective
681 function to measure the discrepancy between each and its corresponding target.

682 Specially,

683

684 $$BCELoss = -\sum_{i}\sum_{t} y_t^i \cdot log\, f_t(x_i) + \left(1 - y_t^i\right) \cdot log\left(1 - f_t(x_i)\right)$$

685 where i indicates the index of input samples and t indicates the index of targets, y_t^i
686 indicates the target label for sample i, cell t and $f_t(x_i)$ represents the predicted
687 probability for target t given input x_i.

688 For controlling overfitting and sparsity of our model parameters, L1 and L2
689 regularization were used as penalty. The final objective function is defined as the sum
690 of BCELoss and regularization terms.

691 $$Objective = BCELoss + \gamma_1 L1loss + \gamma_1 L2loss$$

Where γ_1, γ_2 are defined as hyper parameters controlling L1 and L2 regularization in our model.

To train the model, we minimized the objective function by standard backpropagation, which optimizes the learnable parameters using Adam optimizer. The number of minibatches used in training is determined manually. However, to avoid overfitting, we stopped training early when the model converged on the validation set.

All models were implemented using PyTorch (https://pytorch.org) 1.5.0+cu101, CUDA 10.1, the Anaconda3 distribution of Python and trained on computer cluster equipped four NVIDIA GTX1080Ti GPUs.

Evaluation of prediction performance

The expression prediction performance was evaluated on a held-out test gene set, which had never been seen by the training model to avoid information leakage. For each task, we computed AUROC and the AUPR scores to evaluation of prediction performance from the predictions. And we measured average AUROC and AUPR for each cell lineage in each species-specific model. To evaluate the predicted cell relationship in test set, we computed Pearson correlation between predicted and observed expression profiles measured by log fold change over cell-average. Then, the predicted gene expression of single cell was projected into *t*-SNE embedding using the standard procedure in Scanpy[28]. Adjusted mutual information (AMI) was applied to quantitatively measure the clustering similarity of predicted cell atlas against the original cell atlas.

Benchmark of prediction performance

We performed several benchmarks to validate and compare model prediction performance on human and *Drosophila* specific datasets, the representative scenario of vertebrate and invertebrate. First, we applied two different approaches to generate random datasets, including randomly generating input promoter sequence and shuffling the gene expression label. Models trained on these random datasets yielded the null prediction performance of AUROC (Extended Data. Fig. 8a), which indicating the well trained Nvwa model on real datasets captured the transcription regulatory mapping encoded in the genome.

We compared prediction performance to SVM Classifier. Fimo (v5.0.4)[42] was used to find motifs in the core promoter sequence using JASPAR (v2020-core)[43] under threshold of 1e-6. Then, OneVsRest SVM Classifier was trained using scikit-learn (v0.24.1). Nvwa significantly outperformed SVM Classifier, indicating CNN may learning a better regulatory representation of promoter sequence than validated motifs in JASPAR. Standard deep learning archfitectures, including DeepSEA[41], Beluga[44], Basset[45] and Basenji[46], were benchmarked in the same species-specific datasets. We compared the AUROC scores to measure predictive performance for each task.

Multiple genome training procedure

We selected orthologous genes from closely related species to augment the training set. Orthologous gene pairs were obtained from Ensembl release-102 by BioMart and only high quality orthologous gene pairs among closely related species were obtained.

For the *C. elegans* specific model, we downloaded *C. brenneri* PRJNA20035, *C. briggsae* PRJNA10731, *C. latens* PRJNA248912, and *C. nigoni* PRJNA384657 genome build from WormBase[47], with 74.84%, 77.10%, 75.80% and 75.11% target gene identical to *C. elegans* genes, respectively. For the zebrafish model, we downloaded *Astyanax mexicanus*, *Ictalurus punctatus*, *Denticeps clupeoides*, *Clupea harengus*, *Sphaeramia orbicularis*, *Salarias fasciatus*, and *Gouania willdenowi* genomes build from Ensembl, with 75.21%, 74.21%%, 72.26%, 71.41%, 68.42%, 69.22%, and 69.94% target gene identical to zebrafish genes, respectively. Then, the promoter sequences of closely related species were processed and paired with gene expression label in the training set. Note that those held-out test gene set should not be augmented to prevent information leakage. This multiple genome training strategy augmented the training set and improved the overall predictive performance of the model in the held-out test set, indicating that the closely related species may have similar gene regulation patterns.

Predict genome-wide signals

We extracted 13 kb sequences window across the whole leave-out chr8 of the human and mouse genomes with 1 kb increments. The "bedtools makewindows -w 13000 -s 1000" command was used to generate these windows and "bedtools getfasta" command was used to extract the sequences. The mouse-specific Nvwa model was used to predict on both the plus and minus strands of 1 kb-increment window, and the predicted genome-wide signals of each single cell were aggregated to cell type level. We generated null distributions by predicting on random locations of genome. The

regulatory peak regions were called on each 13 kb sequences window by comparing Nvwa predicted signals with null distribution (threshold p-value < 0.05, Wilcoxon rank-sum test).

In addition, to access Nvwa model on all chromosomes and different genomes, we applied our Human and *Drosophila* specific models to predict transcriptional signals on whole genome reference.

Comparison with functional genomic tracks

Comparison of the genome-wide predictions of Nvwa model with functional genomic tracks provides external validation of model predictive performance. Besides, the comparison of functional genomic data highlighted what the Nvwa model recognized, and gained insight into the biological interpretation. To retain comparable adult mouse samples, we download corresponding Bigwigs files from Mouse sci-ATAC-seq Atlas[48] and ENCODE project. Further, the mm9 gene coordinate was updated to mm10 using CrossMap (v0.5.2)[49].

DeepTools (v2.0)[50] "multiBigwigSummary" and "plotCorrelation" commands were applied to calculate the spearman correlations between Nvwa predictions and functional genomic tracks. ChIPseeker[51] "enrichPeakOverlap" function with parameters "nShuffle=5000" was used to perform statistically rigorous comparisons. And "annotatePeak" function was used for annotating peaks with nearest gene and genomic region. We visualized 400-kb region at chr8 using pyGenomeTrack (v3.6)[52] with "make_tracks_file" and "pyGenomeTracks" commands.

Benchmark of input region

We systematically altered input sequence region in *Drosophila*-specific model with 6.5 kb upstream to 6.5 kb downstream of the TSS (-6.5kb, +6.5kb), 4 kb upstream to 9 kb downstream of the TSS (-4kb, +9kb) and 9 kb upstream to 4 kb downstream of the TSS (-9kb, +4kb). After model training, saliency scores, the gradient of the class score with respect to the input, were used to evaluate the importance of input region[53]. We randomly sampled 10,000 cells to improve the computational efficiency of captum (https://captum.ai/). To visualize the most information rich region that model recognized in input sequence, we plotted the average saliency score among all the tasks and compared it to TSS location.

Model Interpretation

We applied a feature map-based method to interpret the model parameters. Similar to the node-based strategy in Basset[45], each of the first-layer convolution Filters were extracted and converted to a PWM by counting nucleotide occurrences in the feature map. We chose an activation threshold of 0.9, which means that only nucleotide occurrences that activate 0.9 of the maximum value of the Filter were counted in the PWM. Then, we calculated the information content (IC) of each Filter.

$$IC(M,B) = \sum_{i=1}^{m}\sum_{j=1}^{4} M_{i,j}\frac{M_{i,j}}{B_j}$$

Where M and B indicates the PWM (4×m position weight matrix, m=7 in our model) and background, $M_{i,j}$ is the probability of the base j at position i in M and B_j is the probability of the same letter in the background model.

We then annotated the PWM-represented Filter as known transcription factor binding sites (TFBS) by querying the RcisTarget[19] (human, mouse, *Drosophila*) or Cis-BP[54] (zebrafish, *C. elegans*, and planarian) databases using TomTom algorithm (v5.3.3)[20]. Therefore, corresponding TFs were assigned to the Filter.

To compare the first-layer convolution Filters with TF-MoDISco[51] derived motifs on saliency scores. We randomly sampled 10,000 cells and averaged the saliency scores on each task to improve the computational efficiency of TF-MoDISco. Motif similarity was quantified using the TomTom algorithm. Both feature map-based and gradient-based approaches yield similar results.

Cross-validation and Filter reproducibility

To calculate the reproducibility of learned first layer Filters in different sequence input, we implemented a standard 10-fold cross-validation procedure. PWM represented Filters of each individual model were exacted using the methods described above. Motif similarity was quantified using the TomTom algorithm[20]. With a FDR q-value cutoff of 0.05, the reproducibility of Filters was defined as the number of matching motif in each independent cross validation run.

Filter influence

Further, we computed a per cell influence profile, quantifying the fold change between predicted expression with and without each Filter. We replaced all activation values for a certain Filter with zero tensor, then fed forward the modified output through the remaining layers of the model to obtain the altered prediction vector. Specifically, influence of a certain Filter per cell can be denoted as follow,

$$Influence_{t,j} = 1 - \frac{1}{N}\sum_{i}^{N}\frac{f_t(v_{i,j}^{'})}{f_t(v_i)}$$

where i indicates index of N input samples, j indicates index of Filters and t indicates index of targets. $v_i$and $v_{i,j}^{'}$ indicates the original and modified values of Filter j after first layer convolution. $f_t(.)$ represents the output of remaining layers for target t given input v_i. The overall Filter influence score was then defined as average influence among all the cell tasks. And ETA correlation coefficient was used to measure the correlation between Filter influence and cell types.

Gene regulatory network

To reduce the false positive rates in direct target gene identification, 2 kb upstream to 2 kb downstream of the TSS (-2kb, +2kb) were considered as core promoter region for further analysis. Similar to the feature map-based method described above, we exacted and counted the activate sites in gene core promoter regions. For each Filter, direct target genes were identified using number of activate sites with p-value < 0.01 (normal distribution one-tailed test). Further, we calculated the Pearson correlation between Filter influence and target gene expression. And the downstream regulated genes were pruned to remove targeted genes with low correlation under threshold 0.1 (Fig. 4b). Then, the expression patterns per cell were visualized using "Dotplot" function in Scanpy (Fig. 4b).

Global motif interaction analysis

We were interested to ask the predictive effect of cooperative Filter interactions in our model. However, the PWM represented second-layer convolutional Filters seems to

be confusing. Therefore, we performed global motif interaction analysis by quantifying the predictive influence of each combination of every first-layer Filters. Similar to the Filter influence above, we replaced all activation values for each two Filters with zero tensor and compared the modified and original prediction. Influence for cooperative Filter pairs can be denoted as,

$$Influence_{t,j,k} = 1 - \frac{1}{N}\sum_{i}^{N}\frac{f_t(v'_{i,j,k})}{f_t(v_i)}$$

where i indicates index of N input samples, t indicates index of targets. j and k indicates index of Filter pairs. $v_i$and $v'_{i,j,k}$ indicates the original and modified values of Filter pair j, k after first layer convolution. $f_t(.)$ represents the output of remaining layers for target t given input v_i.

Further, we calculated the ETA correlation between influence of cooperative Filter interactions and cell types. We measured the overlap of direct target genes of Filter pairs using Jaccard similarity. We then pruned Filter pairs into cooperative Filter interactions, whose direct target genes significantly overlapped with p-value < 0.01 (normal distribution one-tailed test). Those redundant Filters were identified using TomTom and further excluded. And four syntax patterns of cooperative Filter interactions were identified by comparing the distribution of influence per cells.

Website construction

The main Nvwa website (http://bis.zju.edu.cn/nvwa/) used a bootstrap framework to improve overall adaptability and interactivity. Its back-end was completed by PHP, R language and MySQL. The main functions of the Nvwa website were divided into

three parts: Landscape, Nvwa, and Results. Landscape page provides *t*-SNE visualizations for global view single-cell landscapes for zebrafish, *Drosophila*, and earthworm. Specific markers for each cell types and subclusters are listed in data tables. Nvwa page provided the high-quality machine-learning datasets for eight species, which would promote further development in deep-learning genomics area. The Nvwa code for training, explaining models, and reproducing our work are also open-accessed. Results page contained detailed results of our work, such as raw TomTom analysis results in html format.

Data availability

Digital Expression Matrices can be accessed at http://bis.zju.edu.cn/nvwa/.

Code availability

The source code for running and training the Nvwa models is available at https://github.com/JiaqiLiZju/Nvwa/.

Acknowledgments

G.G. is a participant of the Human Cell Atlas Project. We thank Fei Gu, Chao Li, Kehan Li, and Hangjun Wu for support on the project. We also thank the Center of Cryo-Electron Microscope (CCEM) at Zhejiang University for the resource on the computation. This work was supported by National Natural Science Foundation of China grant 31930028, 31871473, 31922049, 91842301, 32000461, T2121004; National Key Research and Development Program grant 2018YFA0800503, 2018YFA0107804, 2018YFA0107801; Fundamental Research Funds for the Central

885 Universities (G.G.); Alibaba-Zhejiang University Joint Research Center of Future
886 Digital Healthcare.

887 **Author contributions**

888 G.G., X.H., and J. Wang conceived the study. G.G. and X.H. supervised the study.
889 Jiaqi L. designed the model. J. Wang, P.Z., Y.M., Z.S., and D.L. performed
890 computational data analyses. Jiaqi L., J. Wang, P.Z., M.J., R.W., H.L., and J.M.
891 annotated cell landscapes. R.W., M.J., X.H., H.C., Xinru W., Xueyi W., Y.L., D.J.,
892 and T.Z. designed and performed experiments. J. Wang, L.F., L.M., W.E., Y.F., J.L.,
893 C.Y., and Q.G. collected data. J. Wu., S.H., and J.C. guided to model design and
894 parameter optimization. J. Wang, Jiaqi L., P.Z., and Y.M. prepared display items.
895 G.G., J. Wang., Jiaqi L., and P.Z. wrote the initial draft of the manuscript. All authors
896 participated in discussions of results and manuscript editing.

897 **Competing interests**

898 Authors declare that they have no competing interests.

899

900 **References**

901 1. Han, X. *et al.* Construction of a human cell landscape at single-cell level. *Nature* **581**, 303-309
902 (2020).
903 2. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **173**, 1307 (2018).
904 3. Satija, R., Farrell, J.A., Gennert, D., Schier, A.F. & Regev, A. Spatial reconstruction of
905 single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).
906 4. Jiang, M.M. *et al.* Characterization of the Zebrafish Cell Landscape at Single-Cell Resolution.
907 *Frontiers in Cell and Developmental Biology* **9**(2021).
908 5. Li, H. *et al.* Fly Cell Atlas: a single-cell transcriptomic atlas of the adult fruit fly. *bioRxiv*,
909 2021.07.04.451050 (2021).

6. Buchon, N., Silverman, N. & Cherry, S. Immunity in Drosophila melanogaster--from microbial recognition to whole-organism physiology. *Nat Rev Immunol* **14**, 796-810 (2014).
7. Krausgruber, T. *et al.* Structural cells are key regulators of organ-specific immune responses. *Nature* **583**, 296-302 (2020).
8. Cao, C. *et al.* Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nature* **571**, 349-354 (2019).
9. Cao, J.Y. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661-667 (2017).
10. Fincher, C.T., Wurtzel, O., de Hoog, T., Kravarik, K.M. & Reddien, P.W. Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science* **360**(2018).
11. van Dijk, D. *et al.* Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729 e27 (2018).
12. Tarashansky, A.J. *et al.* Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *Elife* **10**(2021).
13. Wang, J. *et al.* Tracing cell-type evolution by cross-species comparison of cell atlases. *Cell Rep* **34**, 108803 (2021).
14. Rui, L., Schmitz, R., Ceribelli, M. & Staudt, L.M. Malignant pirates of the immune system. *Nat Immunol* **12**, 933-40 (2011).
15. Srivastava, A.K. & Schlessinger, D. Structure and Organization of Ribosomal DNA. *Biochimie* **73**, 631-638 (1991).
16. Suzuki, H., Moriwaki, K. & Sakurai, S. Sequences and Evolutionary Analysis of Mouse 5s Rdnas. *Molecular Biology and Evolution* **11**, 704-710 (1994).
17. Zentner, G.E., Balow, S.A. & Scacheri, P.C. Genomic Characterization of the Mouse Ribosomal DNA Locus. *G3-Genes Genomes Genetics* **4**, 243-254 (2014).
18. Maslova, A. *et al.* Deep learning of immune cell differentiation. *Proc Natl Acad Sci U S A* **117**, 25655-25666 (2020).
19. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083-1086 (2017).
20. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S. Quantifying similarity between motifs. *Genome Biology* **8**(2007).
21. Janssens, J. *et al.* Decoding gene regulation in the fly brain. *Nature* **601**, 630-+ (2022).
22. Hannenhalli, S. & Kaestner, K.H. The evolution of Fox genes and their role in development and disease. *Nature Reviews Genetics* **10**, 233-240 (2009).

23. Arendt, D. *et al.* The origin and evolution of cell types. *Nat Rev Genet* **17**, 744-757 (2016).
24. Shafer, M.E.R. Cross-Species Analysis of Single-Cell Transcriptomic Data. *Front Cell Dev Biol* **7**, 175 (2019).
25. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
26. Zheng, G.X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
27. McGinnis, C.S., Murrow, L.M. & Gartner, Z.J. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Syst* **8**, 329-337 e4 (2019).
28. Wolf, F.A., Angerer, P. & Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).
29. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e21 (2019).
30. Shao, Y. *et al.* Genome and single-cell RNA-sequencing of the earthworm Eisenia andrei identifies cellular mechanisms underlying regeneration. *Nature Communications* **11**(2020).
31. Rozanski, A. *et al.* PlanMine 3.0-improvements to a mineable resource of flatworm biology and biodiversity. *Nucleic Acids Res* **47**, D812-D820 (2019).
32. Satou, Y., Kawashima, T., Shoguchi, E., Nakayama, A. & Satoh, N. An integrated database of the ascidian, Ciona intestinalis: towards functional genomics. *Zoolog Sci* **22**, 837-43 (2005).
33. Haas, B.J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494-512 (2013).
34. Emms, D.M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**, 157 (2015).
35. Crow, M., Paul, A., Ballouz, S., Huang, Z.J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat Commun* **9**, 884 (2018).
36. Fischer, S., Crow, M., Harris, B.D. & Gillis, J. Scaling up reproducible research for single-cell transcriptomics using MetaNeighbor. *Nat Protoc* (2021).
37. Hu, H. *et al.* AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res* **47**, D33-D38 (2019).

38. dos Santos, G. *et al.* FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res* **43**, D690-7 (2015).
39. Dubaj Price, M. & Hurd, D.D. WormBase: A Model Organism Database. *Med Ref Serv Q* **38**, 70-80 (2019).
40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9 (2009).
41. Zhou, J. & Troyanskaya, O.G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**, 931-4 (2015).
42. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-8 (2011).
43. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **48**, D87-D92 (2020).
44. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* **50**, 1171-1179 (2018).
45. Kelley, D.R., Snoek, J. & Rinn, J.L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26**, 990-9 (2016).
46. Kelley, D.R. *et al.* Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* **28**, 739-750 (2018).
47. Harris, T.W. *et al.* WormBase: a modern Model Organism Information Resource. *Nucleic Acids Res* **48**, D762-D767 (2020).
48. Cusanovich, D.A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324 e18 (2018).
49. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006-7 (2014).
50. Ramirez, F., Dundar, F., Diehl, S., Gruning, B.A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research* **42**, W187-W191 (2014).
51. Yu, G., Wang, L.G. & He, Q.Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382-3 (2015).
52. Ramirez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat Commun* **9**, 189 (2018).
53. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. 1-8 (ICLR, 2014).

54. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).

Figure

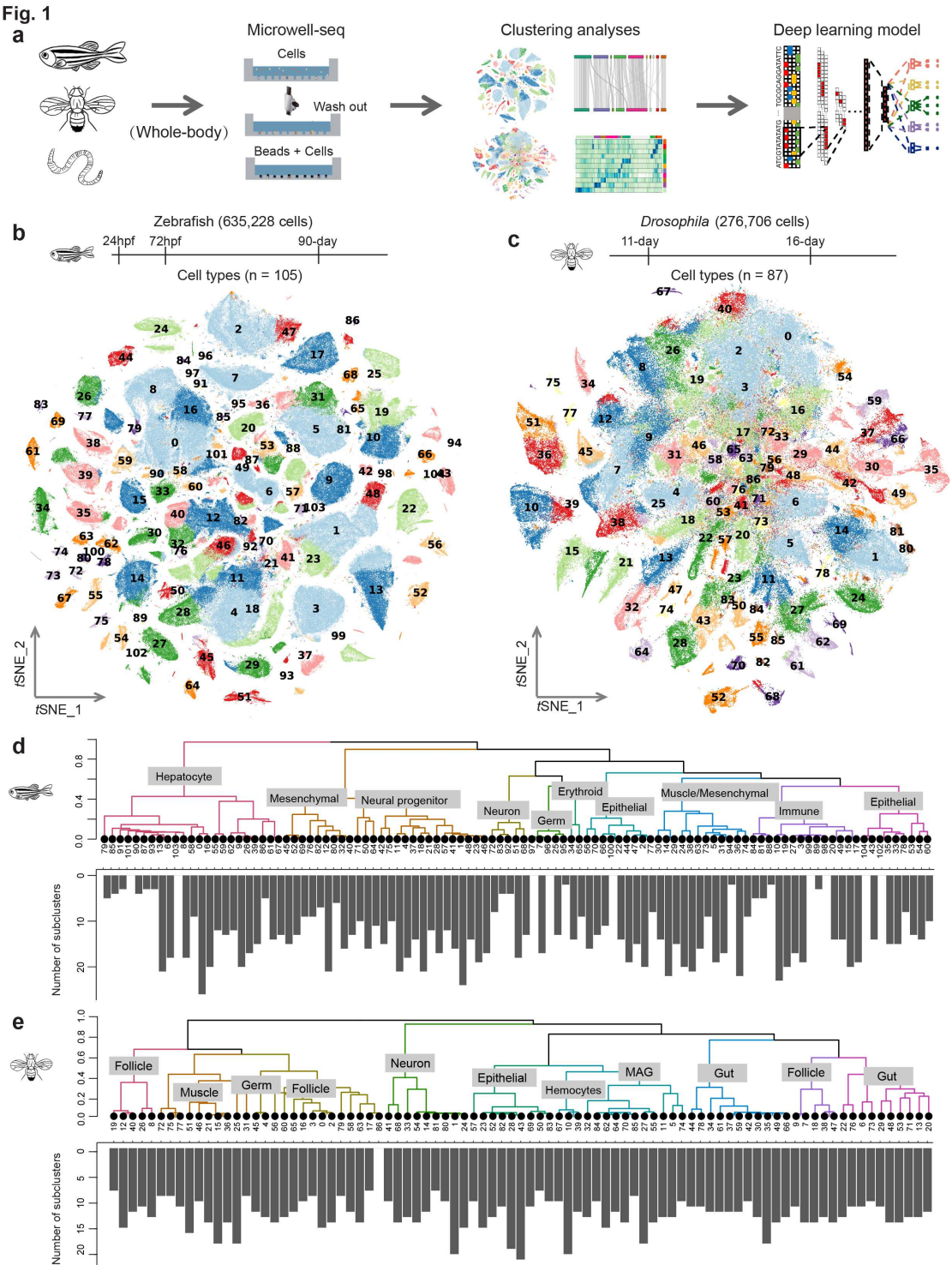


Fig. 1. Zebrafish and *Drosophila* cell landscapes were constructed using Microwell-seq. (**a**) Overview of the whole-body single-cell mRNA-seq experimental

and bioinformatics analysis workflow. (**b**) *t*-SNE analysis of 635,228 single cells from zebrafish. In the *t*-SNE map, 105 cell types are labeled by different colors. (**c**) *t*-SNE analysis of 276,706 single cells from *Drosophila*. In the *t*-SNE map, 87 cell types are labeled by different colors. (**d**) The hierarchical clustering tree (top) showing the similarity (AUROC scores) among zebrafish 105 cell types, and the histogram plot (bottom) showing the subtypes of each cell type. (**e**) The hierarchical clustering tree (top) showing the similarity (AUROC scores) among *Drosophila* 87 cell types, and the histogram plot (bottom) showing the subtypes of each cell-type. The similarity refers to the AUROC scores from MetaNeighbor analyses.

Fig. 2

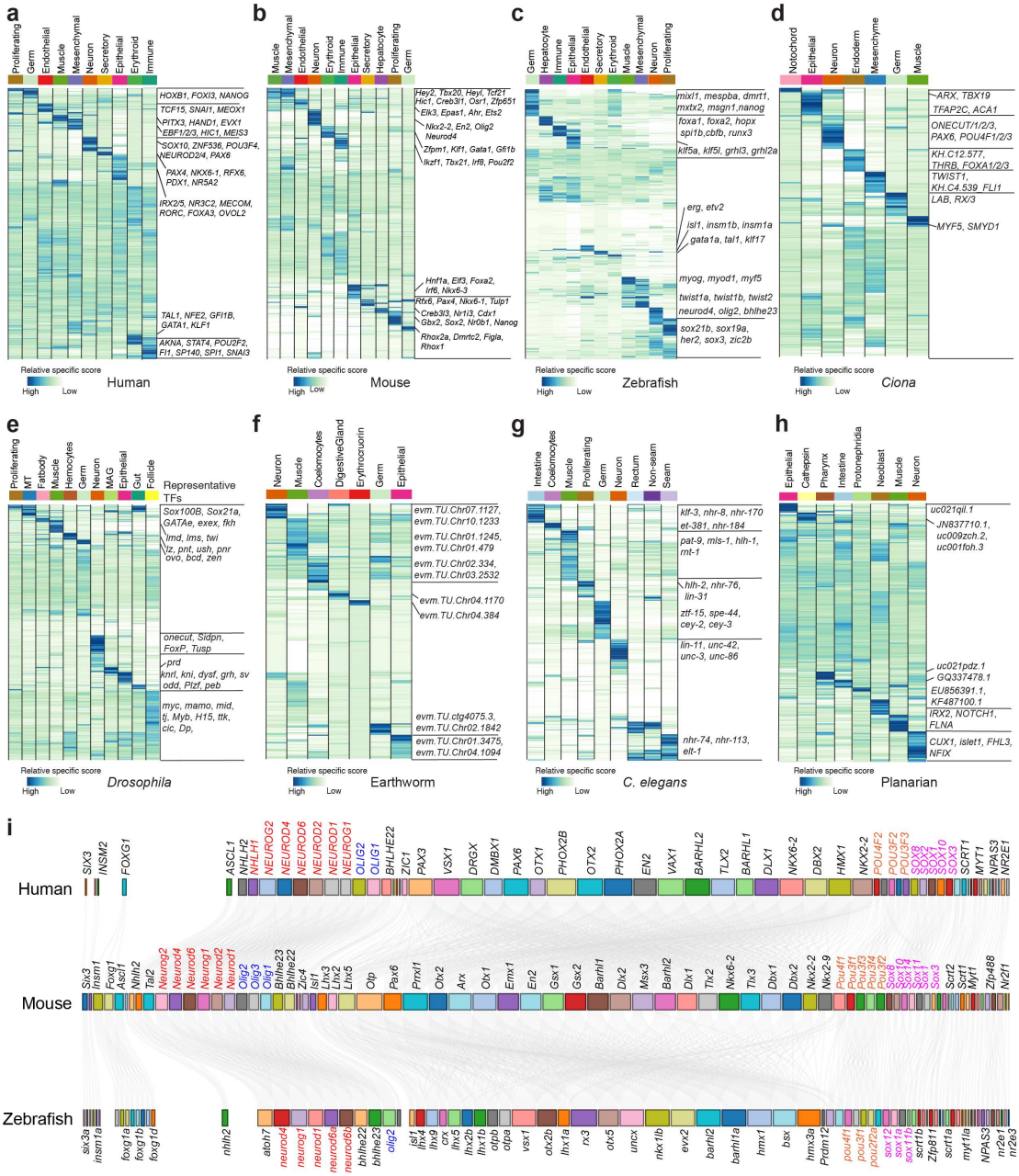


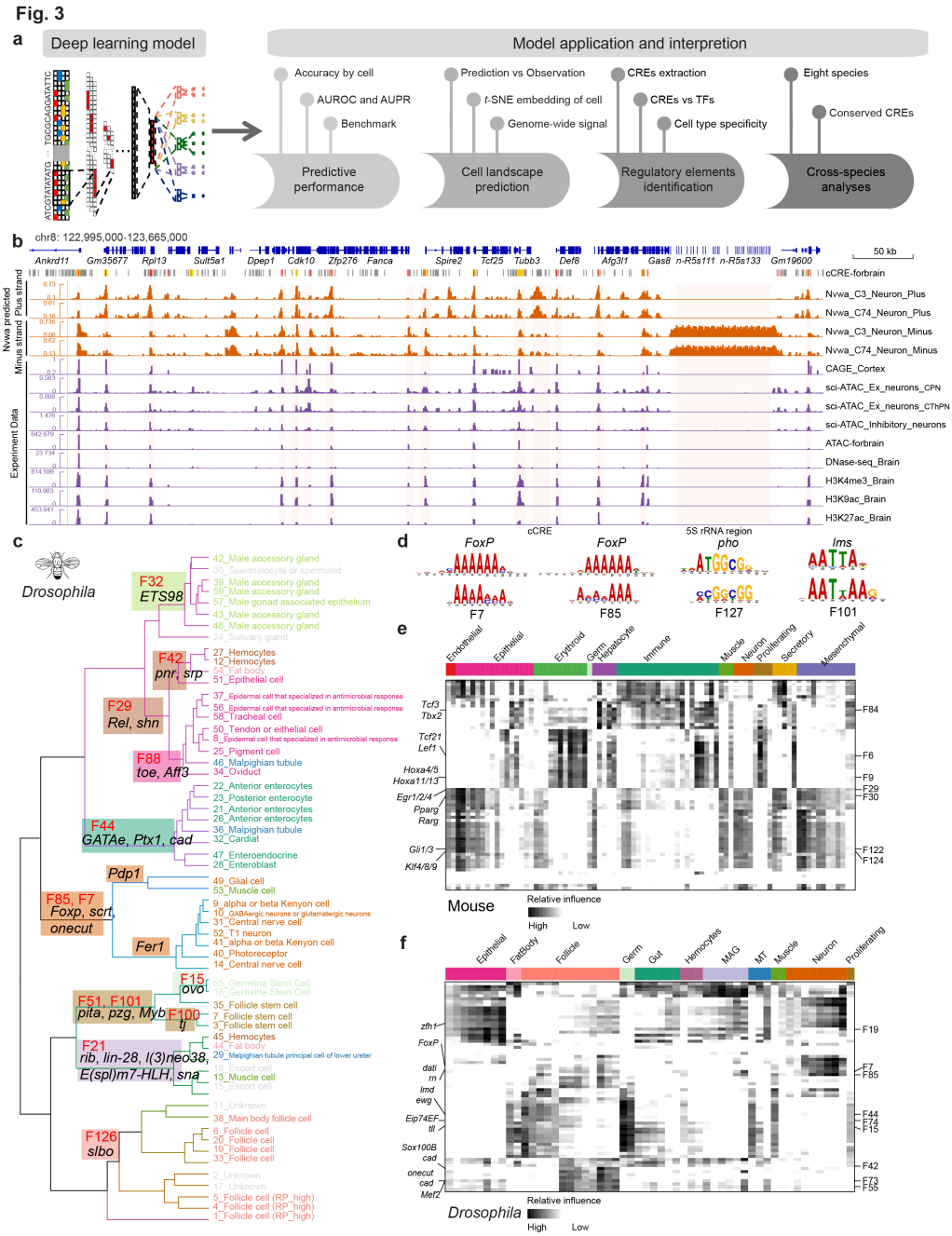
Fig. 2. Comparison of regulatory TFs within and across species using single-cell RNA sequencing data. (**a-h**) Heatmaps showing the specific scores of TFs in humans (**a**), mice (**b**), zebrafish (**c**), *Ciona* (**d**), *Drosophila* (**e**), earthworm (**f**), *C. elegans* (**g**), and planarians (**h**). Each row represents one TF, and each column represents one cell lineage. The representative TFs in each cell lineage are presented in the right panel. (**i**)

1031 Sankey diagram showing homologous relationships among vertebrates' neuron-related

1032 TFs.

1033

1034



1035

1036 **Fig. 3. Application and interpretation of the deep-learning model framework.** (**a**)

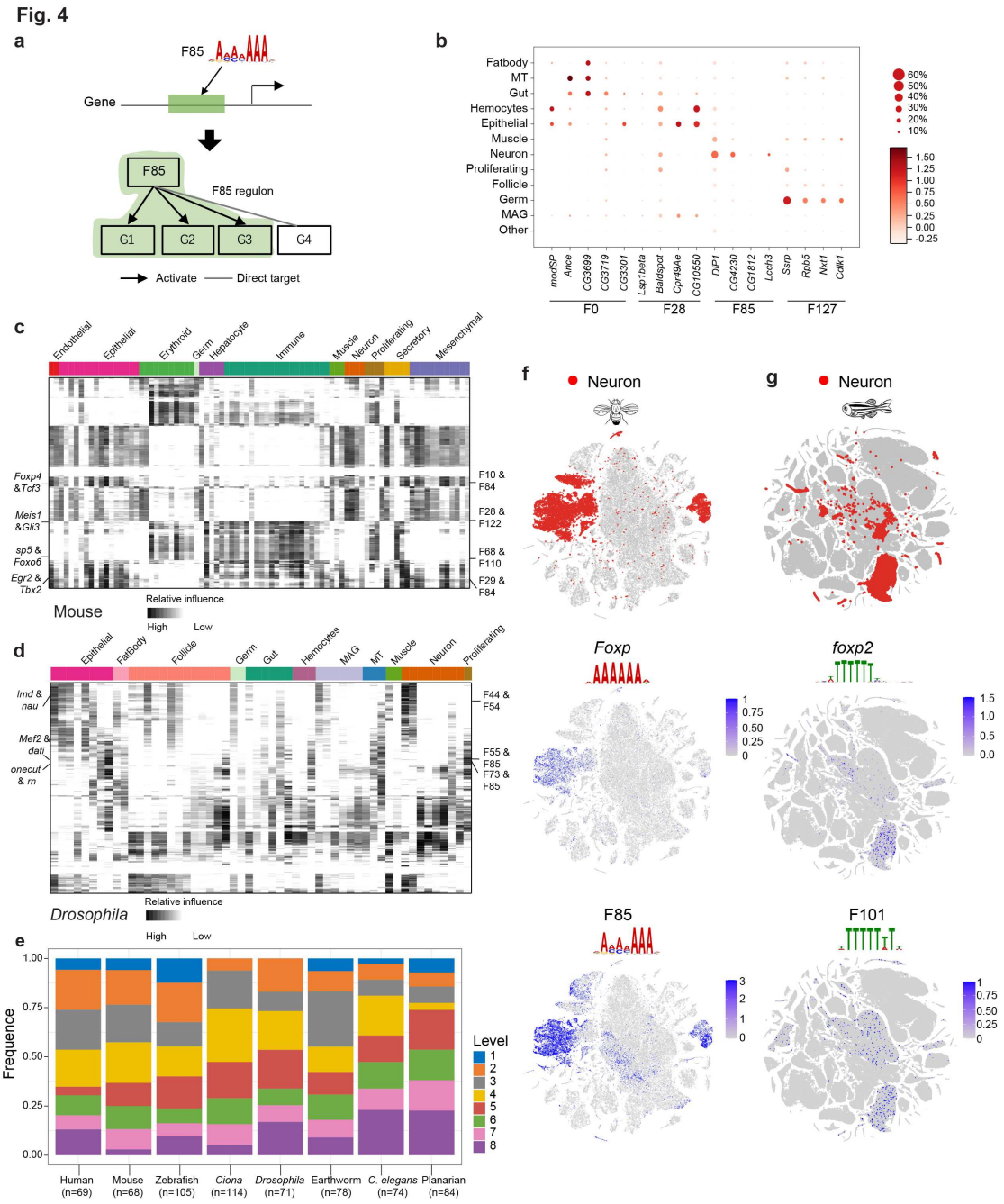
1037 Overview of the analysis workflow of Nvwa. (**b**) Visualization of the predicted

1038 genome-wide transcriptional signals and experimentally functional genomics data at

1039 mouse held-out chr8:122,995,000-123,665,000 region in neuronal cells. ENCODE,

1040 scATAC, and Nvwa predictions are indicated in the side panel. ENCODE cCRE and 5S

rRNA region are highlighted. (**c**) Node-supporting TFs and Filters in *Drosophila* cell-type hierarchical tree inferred with MetaNeighbor. Selected TFs and Filters supporting each node are indicated, with Filters in red. (**d**) Examples of known TFBS compared with the PWMs of Nvwa first-layer Filters F7, F85, F127, and F101 in *Drosophila*. (**e-f**) Heatmaps showing the specific scores of Filters in mice (**e**) and *Drosophila* (**f**). Each row represents one Filter, and each column represents one cell. The representative Filters in each cell lineage are presented in the right panel, and the corresponding TFs are presented in the left panel.



1050

1051 **Fig. 4. Interpretation of the deep learning model framework.**

1052 (**a**) Workflow of Filter regulon identification. (**b**) Dotplot visualizing the expression

1053 patterns of direct target genes for Filter F0, F28, F85, and F127 in *Drosophila*. (**c-d**)

1054 Cooperative syntax of Nvwa first-layer Filter pairs. Heatmaps showing the positive

1055 cell type-specific influence scores of Filter interactions in mice (**c**) and *Drosophila* (**d**).

Each row represents a Filter pair, and each column represents one cell. The representative TF pairs in each cell lineage are presented in the left panel, and the corresponding Filter pairs are presented in the right panel. (**e**) The distribution of cell lineage-specific TFs at eight conservative levels across humans, mice, zebrafish, *Ciona*, *Drosophila*, earthworm, *C. elegans*, and planarians. The number and corresponding colour refer to the conservative level. Level 1 means that Filters have no homologous Filters in other species, and Level 8 means that Filters have homologous Filters in all other seven species. (**f**) *t*-SNE visualization of the cell type-specific Filter (F85) influence and the corresponding TF expression in *Drosophila*. (**g**) *t*-SNE visualization of the cell type-specific Filter (F101) influence and the corresponding TF expression in zebrafish.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NGLE58036R2SupplementaryFiles.pdf