 doi:10.20944/preprints202205.0378.v1

Article

Single-cell atlas of the *Drosophila* leg disc identifies a long non-coding RNA associated with distal leg development

Joyce Tse [1,2], Tsz Ho Li [1,2], Jizhou Zhang [1,2], Alan Lee [1,2], Ivy Lee [1,2], Zhe Qu [1,2], Xiao Lin [1,2], Jerome Hui [1,2], and Ting-Fung Chan [1,2,*]

[1] School of Life Sciences, The Chinese University of Hong Kong, Hong Kong
[2] State Key Lab of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong
* Correspondence: tf.chan@cuhk.edu.hk

Abstract: The *Drosophila* imaginal disc has been an excellent model for the study of developmental gene regulation. In particular, long non-coding RNAs (lncRNAs) have gained widespread attention in recent years due to their important role in gene regulation. Their specific spatiotemporal expressions further support their role in developmental processes and diseases. In this study, we explored the role of a novel lncRNA in *Drosophila* leg development by dissecting and dissociating w^{1118} third-instar larval third leg (L3) discs into single cells and single nuclei, and performing single-cell RNA-sequencing (scRNA-seq) and single-cell assays for transposase-accessible chromatin (scATAC-seq). Single-cell transcriptomics analysis of the L3 discs across three developmental timepoints revealed different cell types and identified *lncRNA:CR33938* as a distal specific gene with high expression in late development. This was further validated by fluorescence *in-situ* hybridization (FISH). The scATAC-seq results reproduced the single-cell transcriptomics landscape and elucidated the distal cell functions at different timepoints. Furthermore, overexpression of *lncRNA:CR33938* in the S2 cell line increased the expression of leg development genes, further confirming its important role in development.

Keywords: *Drosophila*; leg imaginal disc; lncRNA; development; scRNA-seq; scATAC-seq

1. Introduction

Long non-coding RNAs (lncRNAs) are defined as RNAs longer than 200 nucleotides and not translated into functional proteins. Human GENCODE (v40) identifies 17,748 lncRNA genes, which roughly equates to the number of protein coding genes (19,988) signifying the importance of lncRNAs. The majority of lncRNAs are transcribed by RNA polymerase II and are often 5′-end 7-methyl guanosine (m7G) capped, 3′-end polyadenylated, and spliced similarly to mRNAs. They are often classified based on their position relative to neighboring genes (divergent, convergent, intergenic, antisense, sense, enhancer, intronic and miRNA host), transcript length (long intergenic, very long intergenic, and macroRNA), association with annotated protein-coding genes, association with other DNA elements, protein-coding RNA resemblance, association with repeats, association with a biochemical pathway, sequence and structure conservation, biological state, association with subcellular structures, and function [1], [2]. As key regulators of gene expression at the epigenetic, transcriptional, and post-transcriptional levels, they are implicated in various biological processes and diseases. The contribution of lncRNAs to organ development in several mammalian species has revealed a transition of broadly expressed lncRNAs towards an increasing number of spatiotemporal-specific and condition-specific lncRNAs [3]. The role of lncRNAs in cancer has been studied extensively, but they are also involved in many other human diseases from neurological disorders to cardiovascular issues [4]. Notably, lncRNA expression is generally spatiotemporal specific, indicating the unique functions and probable pharmacological targeting of lncRNA.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0378.v1

Drosophila melanogaster (fruit fly) is an ideal model organism to study developmental and cellular processes in higher eukaryotes, including humans, because a wide range of genetics tools can be applied and its genome has been extensively studied [5]. In fact, the *D. melanogaster* genome is 60% homologous to that of humans and nearly 75% of human disease-causing genes are believed to have functional homologs in the fruit fly [6]. Furthermore, its short generation time, high fecundity, and low maintenance as well as the abundance of publicly available fly stocks and databases also make *D. melanogaster* an appealing model organism.

Despite different taxonomic origins, the *Drosophila* larval leg disc, which develops into the adult leg, is an ideal model for studying the complex vertebrate limb because it is relatively simple and amenable to genetic manipulations. Research on fly imaginal discs has revealed the tissue compartments and organ-specific regulator genes critical to development, and has generated established models for the study of cellular interactions and complex genetic pathways [7]. Moreover, the easy accessibility of imaginal discs further supports their utility.

Advances in the past decade on single-cell RNA sequencing (scRNA-seq) and related computational analysis pipelines have allowed scientists and bioinformaticians to understand the cellular heterogeneity of tissues at an unprecedented level, from manually selecting a single-cell under the microscope to plate-based and droplet-based high throughput methods with multimodal capabilities [8]. Since the publication of the first single-cell transcriptome study based on a next-generation sequencing platform, the number of publications on scRNA-seq addressing development, disease, and bioinformatics tool improvement has exponentially grown [9]–[12]. Many of these publications have focused on developmental biology, often involving single-cell studies, as it represents a crucial period during which cells first begin to differentiate [13]. Single-cell transcriptomics studies on *Drosophila* larval imaginal wing and eye-antennae discs have emerged since 2018 [14]–[19] and shown that single cells could be mapped to the distinct subregions of their respective imaginal discs, thus confirming the spatial expression of genes determined by previous immunostaining methods.

While the Fly Cell Atlas recently performed single-nucleus RNA-sequencing (snRNA-seq) on adult *Drosophila* legs [20], single-cell transcriptomics and epigenomics studies on the developing leg imaginal disc remain lacking due to the challenges of its dissection compared to the larger wing and eye-antennae discs. We thus report the single-cell transcriptomic and epigenomic landscapes of w^{1118} third leg discs (L3) across three stages of development of third-instar larvae. We identified and validated a novel, highly expressed lncRNA in the distal epithelial cells that changes its spatial expression at various stages of development and confirms its importance in leg development.

2. Results

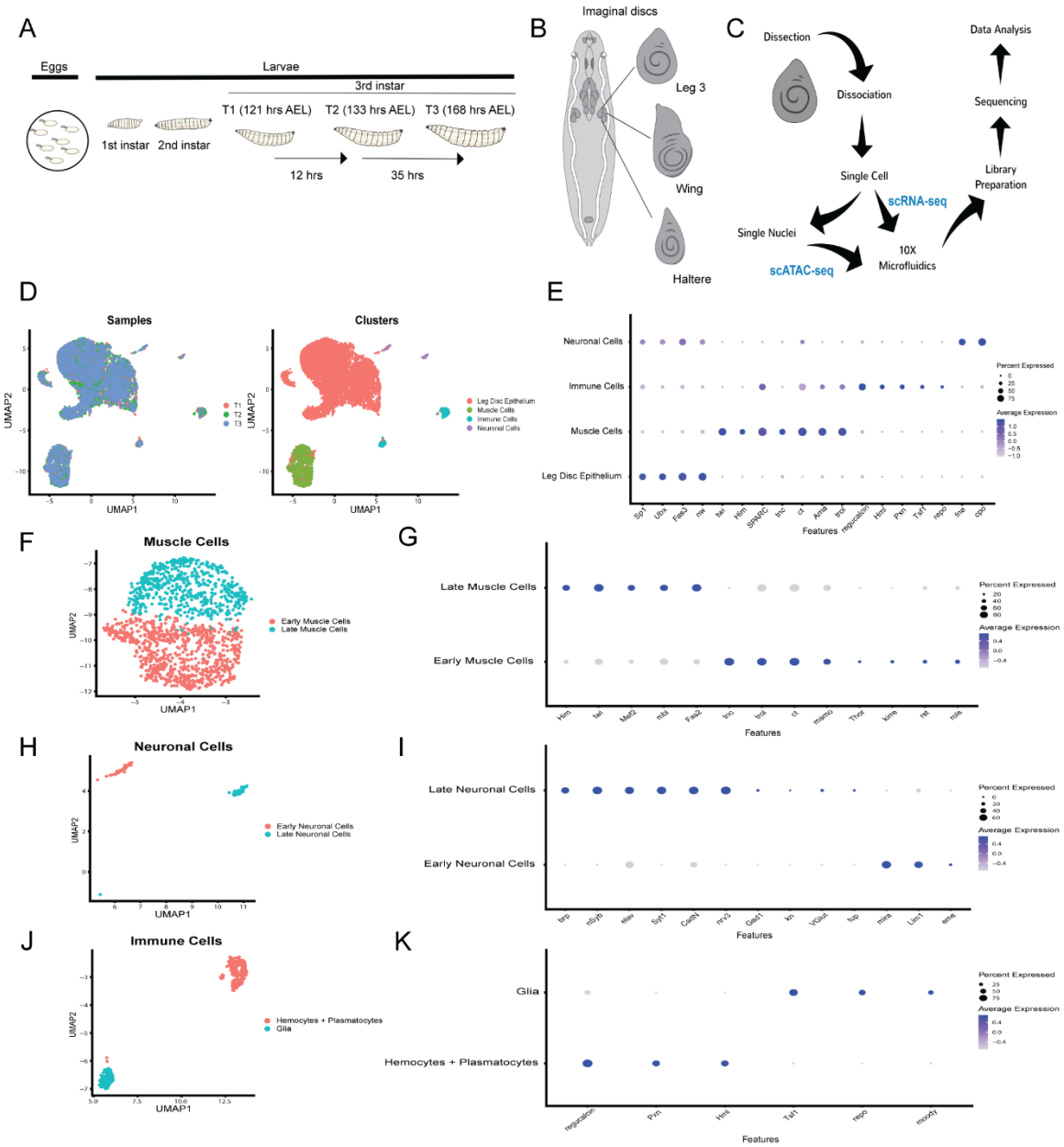
2.1. Generation of a transcriptomic cell atlas of the developing leg imaginal disc

2.1.1. Single-cell RNA-sequencing identifies four main cell types in L3 discs

To study the cellular heterogeneity of developing L3 discs, collected embryos were dissected for L3 discs at 121 hours (T1), 133 hours (T2), and 168 hours (T3) after egg laying (AEL) (**Figure 1A**) for scRNA-seq. Sequencing statistics showed similar data quality amongst the three samples, including percentage of mapped reads, percentage of mapped reads aligned to genes, number of cells, and mean reads per cell (**Table S1**). The L3 disc was identified as a trio of discs on either side of the larval body that differed from the wing and haltere discs in morphology and patterning (**Figure 1B**). Cell preparation workflow involved dissection and dissociation of L3 discs into single cells, after which a portion of the cells were used for scRNA-seq and the remaining cells having their nuclei isolated for single-cell assay for transposase-accessible chromatin (scATAC-seq) (**Figure 1C**). Both assays used the 10x Genomics platform and the prepared libraries

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0378.v1

were subjected to sequencing and subsequent data analysis. The integrated dataset overlayed T1, T2, and T3 individual samples and identified four distinct clusters (**Figure 1 D**). The largest cluster represented the leg disc epithelium, which expressed epithelial markers *Fasciclin 3* (*Fas3*) and *narrow* (*nw*) (**Figure 1E**). Expression of *Sp1* and *Ultrabithorax* (*Ubx*) confirmed that the cells originated from L3 discs. The second largest cluster represented muscle cells, which expressed the muscle markers *twist* (*twi*), *Holes in muscle* (*Him*), *Secreted protein, acidic, cysteine-rich* (*SPARC*), *tenectin* (*tnc*), *cut* (*ct*), *Amalgam* (*Ama*), and *terribly reduced optic lobes* (*trol*). The identity of the immune cell cluster was determined by expression of *regucalcin*, *Hemolectin* (*Hml*), *Peroxidasin* (*Pxn*), *Transferrin 1* (*Tsf1*), and *reversed polarity* (*repo*). The smallest cluster represented the neuronal cells, which expressed *found in neurons* (*fne*) and *couch potato* (*cpo*).



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0378.v1

ring-like pattern at its center within a trio of discs on bilateral sides of the larvae which also included the wing and haltere discs. (C) Flowchart of scRNA-seq and scATAC-seq experiments. Dissected leg discs were dissociated into single cells, a portion of which were used for scRNA-seq with the remaining cells having their nuclei isolated for scATAC-seq. Both scRNA-seq and scATAC-seq used the 10x Genomics Chromium Controller and proceeded with their respective library preparation protocols, sequencing, and data analysis. (D) UMAP visualizations of the scRNA-seq data showing that T1, T2, and T3 overlay each other, although the four identified cell types were quite segregated. (E) Dot plot showing the known marker genes of the respective cell types identified in the UMAP visualization. (F) Muscle cell subset of the scRNA-seq data showing differentiation between early and late muscle cells. (G) Dot plot showing the known marker genes of the early and late muscle cells identified in the UMAP visualization. (H) Neuronal cell subset of the scRNA-seq data showing differentiation between early and late neuronal cells. (I) Dot plot showing the known marker genes of the early and late neuronal cells identified in the UMAP visualization. (J) Immune cell subset of the scRNA-seq data showing differentiation between hemocytes (and plasmatocytes) and glia. (K) Dot plot showing the marker genes of hemocytes (and plasmatocytes) and glia identified in the UMAP visualization.

2.1.2. Subclustering of the main cell types reveals cell subtypes

The muscle cell cluster was composed of early and late muscle cell subclusters (**Figure 1F**). The early cells expressed *tenectin* (*tnc*), *terribly reduced optic lobes* (*trol*), *cut* (*ct*), *maternal gene required for meiosis* (*mamo*), *Thor*, *kin of irre* (*kirre*), *roughest* (*rst*), and *rolling pebbles* (*rols*) (**Figure 1G**). The late cells expressed *Holes in muscle* (*Him*), *twist* (*twi*), *Myocyte enhancer factor 2* (*Mef2*), *muscleblind* (*mbl*), and *Fasciclin 2* (*Fas2*). Early muscle cells increased expression of late muscle cell marker *Fas2* over time in terms of both expression level and the number of cells that expressed this gene (**Figure S1**). Late muscle cell marker *Mef2*, a skeletal muscle differentiation transcription factor, similarly increased expression in the late muscle cell subcluster over time in terms of both expression level and the number of cells that expressed the gene. The heatmap of the most upregulated genes in the early and late muscle cells showed a distinction in upregulated genes between the two subclusters (**Figure S2**).

The neuronal cell cluster was also composed of early and late neuronal cell subclusters (**Figure 1H**). The early cells expressed *miranda* (*mira*), *LIM homeobox 1* (*Lim1*), and *empty spiracles* (*ems*) (**Figure 1I**). The late cells expressed *bruchpilot* (*brp*), *neuronal Synaptobrevin* (*nSyb*), *embryonic lethal abnormal vision* (*elav*), *Synaptotagmin 1* (*Syt1*), *Cadherin-N* (*CadN*), *nervana 3* (*nrv3*), *Glutamic acid decarboxylase 1* (*Gad1*), *knot* (*kn*), *vesicular glutamate transporter* (*VGlut*), and *tailup* (*tup*). The heatmap of the most upregulated genes in the early and late neuronal cells showed a clear distinction between the two subclusters (**Figure S3**).

The immune cell cluster was composed of glia and hemocytes (including plasmatocytes), which are the phagocytes found in invertebrates **(Figure 1J)**. Glial cells expressed *Transferrin 1* (*Tsf1*), *reversed polarity* (*repo*), and *moody*, while the hemocytes and plasmatocytes expressed *regucalcin*, *Peroxidasin* (*Pxn*), and *Hemolectin* (*Hml*) (**Figure 1K**). These markers were highly specific to their respective cell subtypes and the heatmap of the most upregulated genes in the glia and hemocytes (including plasmatocytes) showed a clear distinction between the two subclusters (**Figure S4**).

The leg disc epithelium cluster was subclustered into six cell subtypes, including the distal, medial, and proximal cells as well as stem cell-like cells, such as those of the proximal-distal-axis (PD axis) and anterior-posterior-axis (AP axis), and cells of undetermined fate (**Figure 2A**). The distal cells expressed the markers *aristaless* (*al*), *C15*, and *Distal-less* (*Dll*) (**Figure 2B**). The medial cells expressed *dachshund* (*dac*) and *Dll*, while the proximal cells expressed *teashirt* (*tsh*) and *homothorax* (*hth*). The PD axis cells expressed

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0378.v1

vestigial (*vg*), *spalt-related* (*salr*), and *spalt major* (*salm*), and the AP axis cells expressed *hase und igel* (*hui*).

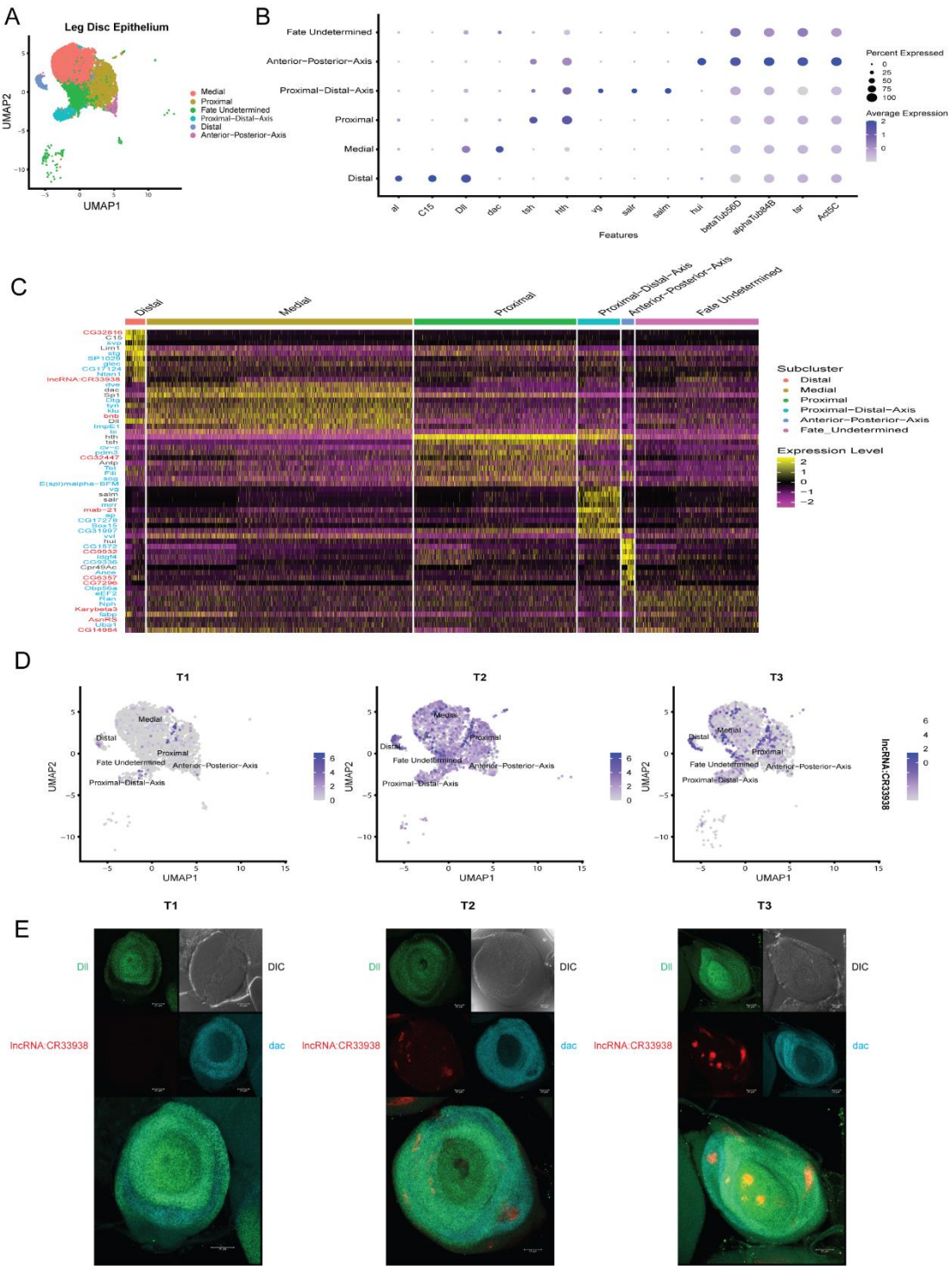


Figure 2. Subclustering of the epithelial cell cluster with identified cell subtypes along the PD axis and a distal-specific *lncRNA:CR33938*. (A) Leg disc epithelium subset of the scRNA-seq data showing differentiation of cells along the PD axis of the fly leg. (B) Dot plot showing the known marker genes of the proximal, medial, and distal cells as well as the earlier stem-cell like cells of the PD axis. (C) Heatmap of the top ten most upregulated genes for each cell subtype (subcluster) of the epithelial cell cluster, where black represents known marker genes, blue represents genes of known function as potential markers, and red represents genes of unknown functions. *LncRNA:CR33938* was

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0378.v1

identified as one of the most upregulated genes in the distal cells. (D) Feature plots showing the expression levels of *lncRNA:CR33938* in different epithelial subclusters across T1, T2, and T3. (E) Validation of the scRNA-seq *lncRNA:CR33938* identified using FISH showing negligible expression during T1, epithelium wide expression in T2, and mainly distal-specific expression in T3.

2.2. Identification and characterization of a novel long non-coding RNA

2.2.1. Identification of a long non-coding RNA of unknown function in distal cells

The most upregulated genes in each leg disc epithelium subcluster is shown in a heatmap (**Figure 2C**). The genes colored black represent known markers for their respective subclusters, those colored blue represent genes with known functions as potential markers for their respective subclusters, and the genes colored red represent genes with unknown functions as potential markers for their respective subclusters. lncRNA (*lncRNA:CR33938*) is unique because the 10x 3' gene expression kit detects polyA-tailed transcripts, which mostly include mRNAs. However, *lncRNA:CR33938* expression was observed in this study (**Figure 2D**). Upon splitting the integrated data into its respective samples (T1, T2 and T3), *lncRNA:CR33938* expression was negligible in T1, appeared more widespread in T2 and became specific to the distal cells in T3.

2.2.2. Experimental validation of *lncRNA:CR33938* expression in L3 discs

Fluorescence *in-situ* hybridization (FISH) of *lncRNA:CR33938* in T1, T2, and T3 L3 discs was performed alongside region-delineating controls *Dll* and *dac* (**Figure 2E**). Expression of only *Dll* represented the distal cells, while co-expression of *Dll* and *dac* or only *dac* represented the medial cells. *LncRNA:CR33938* expression in T1 L3 discs did not occur. *LncRNA:CR33938* expression in T2 L3 discs was present in the proximal, medial, and distal cells, while *lncRNA:CR33938* expression in T3 L3 discs was most prominent in distal cells, although medial cells also showed more limited expression. These FISH results corroborated the scRNA-seq data.

2.2.3. Conservation of *lncRNA:CR33938* in insect species

The conservation state of *lncRNA:CR33938* across 124 insect species revealed that the lncRNA had a high conservation level in exon regions (**Figure 3A**). Moreover, comparison with the conservation state of all 2258 lncRNAs annotated in the reference annotation suggested that *lncRNA:CR33938* was more conserved than 90% of the other lncRNAs (**Figure 3B**). These concordances reflected a critical regulatory role of *lncRNA:CR33938* in insect development.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0378.v1

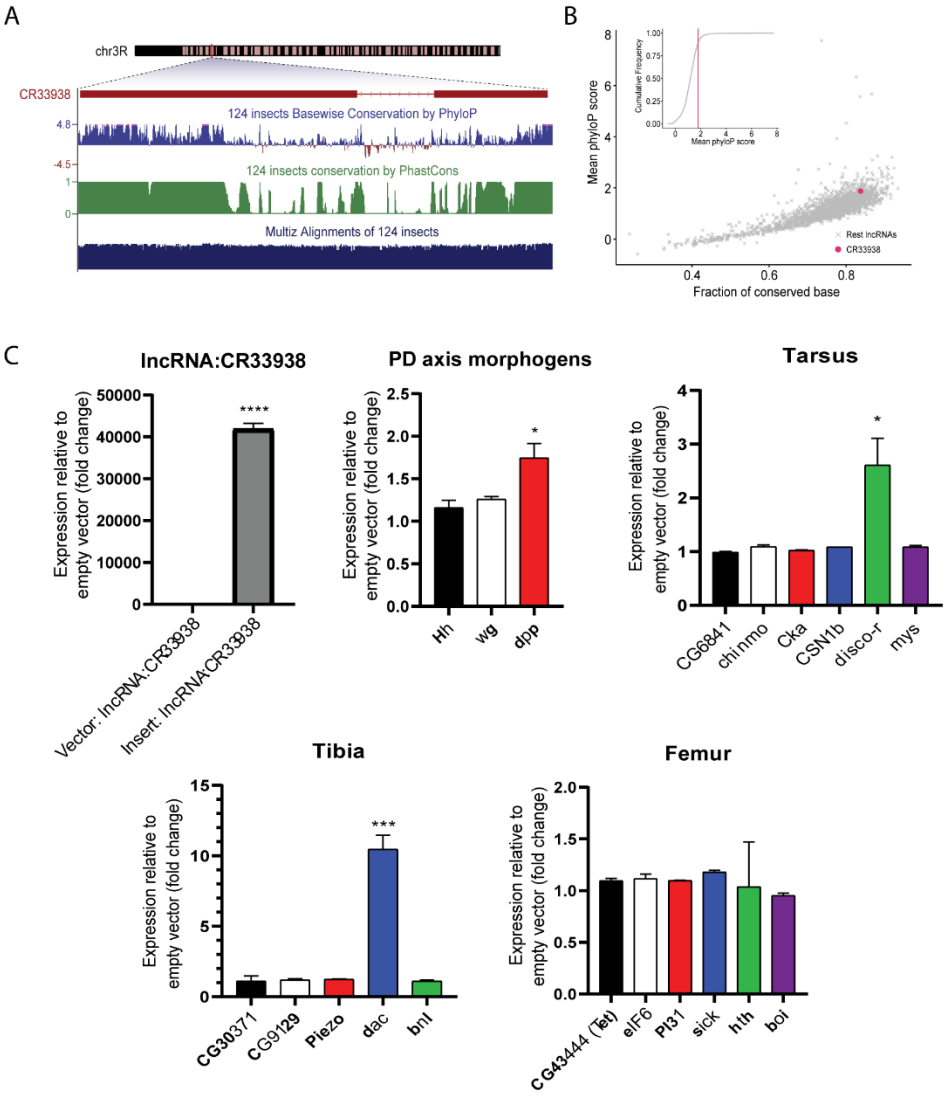


Figure 3. ***LncRNA:CR33938*** **is conserved among insects and overexpression in** ***Drosophila*** **S2 cells increased expression levels of genes involved in leg development.** (A) *lncRNA:CR33938*, identified on chromosome 3, is conserved within insects. (B) The fraction of conserved bases of *lncRNA:CR33938* across insects is greater than 0.8. (C) Overexpression of *lncRNA:CR33938* in S2 cells produced an increase in the expression of leg development genes, including PD axis genes, distal leg tarsal *disco-r*, and medial leg tibial *dac* according to qPCR. There was no effect on proximal leg femur genes. *, ***, **** equate to *p*-values of less than 0.05, 0.001 and 0.0001, respectively.

2.2.4. Overexpression of *lncRNA:CR33938* in S2 cells

Transient overexpression of full length *lncRNA:CR33938* in S2 cells produced an approximately 40,000-fold increase in expression level compared to the empty vector control (**Figure 3C**) according to qRT-PCR. Correspondingly, expression of the PD axis genes (*Hh*, *wg*, and *dpp*) showed an increasing trend upon *lncRNA:CR33938* overexpression. While there was no effect on the expression of genes controlling proximal leg femur growth, expression of distal leg tarsal *disco-r* and medial leg tibial *dac* significantly increased with *lncRNA:CR33938* overexpression. This corroborated the scRNA-seq and FISH data that *lncRNA:CR33938* more greatly affected (and was normally expressed in) the distal end of the leg.

2.3. Generation of an epigenomic cell atlas of the developing leg imaginal disc

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0378.v1

2.3.1. scATAC-seq identified similar cell types as scRNA-seq

The sample-wide integrated scATAC-seq dataset showed an overlaying of the T1, T2, and T3 individual samples (**Figure 4A**) and identified twelve distinct clusters based on differences in chromatin accessibility (**Figure 4B**). A heatmap distinguished the proportion of cells in each cluster at each timepoint and showed differences in chromatin accessibility and cell composition across three timepoints (**Figure 4C**). For example, cluster 6 (C6) showed greater than 80% of the cells in T1, a small proportion of the cells in T2, and nearly no cells in T3. Similarly, T1 had many cells from cluster 12 (C12) and cluster 2 (C2).

Upon integration of the chromatin accessibility data with the gene expression data, seven cell types identified in scRNA-seq were transferred to the scATAC-seq clusters (**Figure 4D**). These cell types corresponded to the cell subtypes of the PD axis of the leg disc epithelium (proximal, medial, and distal cells) as well as those of the muscle, neuronal, and immune cells.

The cell type identities were confirmed by an inferred gene score of chromatin accessibility for a list of known marker genes specific to the cell types (**Figure 4E**). Similar to the gene expression data in scRNA-seq, high gene scores of *Sp1* and *Ubx* confirmed that the cells originated from L3 discs. All cells that composed the leg disc epithelium (proximal, medial, and distal cells) showed markers *Fas3* and *nw*. The presence of *Dll* only (without *dac*), *al*, and *C15* confirmed the identity of the distal cells. *Dll* (with *dac*) and *dac* only confirmed the identity of the medial cells. Similarly, *tsh* and *hth* were markers for the proximal cells, while *Him* and *twi* represented the muscle cells. The neuronal cells showed high gene scores for *nervana 3* (*nrv3*) and *complexin* (*cpx*), and the immune cells produced high scores for *Hml* (for hemocytes) and *Pxn* (for plasmatocytes) markers.

The most enriched motifs for each cell type are shown as a heatmap (**Figure 4F**). The GATA motifs were evident in the immune cells, with several GATA family members observed. The PRDM9 and HIF2a.bHLH motifs were highly enriched in the medial and distal cells, respectively. While the NRF motif was enriched in the proximal cells, its enrichment was more evident in the neuronal cells. The muscle cells were enriched in many motifs, including Maz, KLF14, ZNF, Egr2, Olig2, Egr1, KLF10, and Klf9. The neuronal cells were also enriched for many motifs, including MyoD, Myf5, E2A, PAX5, MyoG, Tcf12, EKLF, Ascl1, and NRF.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0378.v1

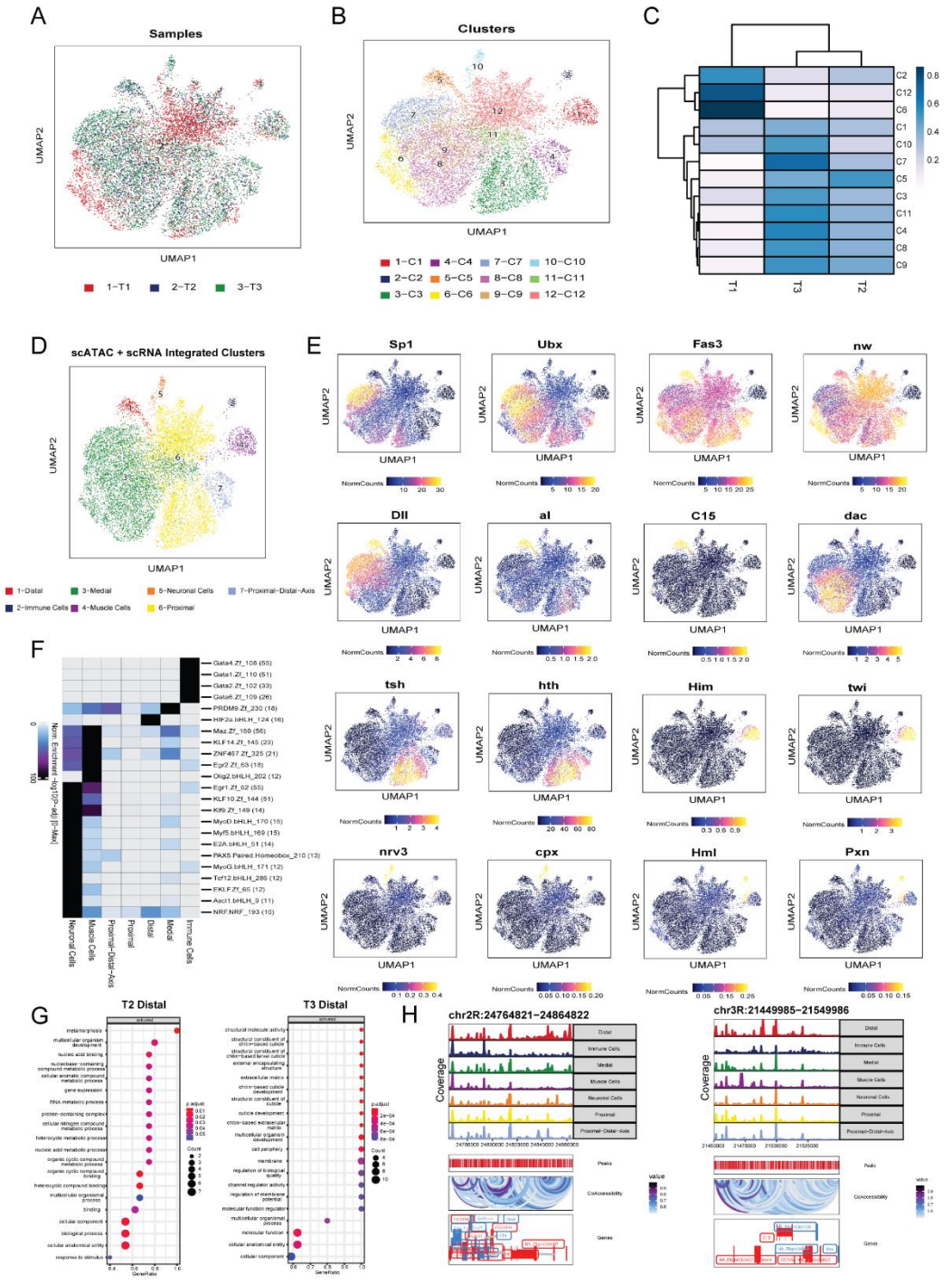


Figure 4. scATAC-seq revealed the same major cell types of *Drosophila* L3 disc as in scRNA-seq. (A) UMAP visualization of the scATAC-seq data, showing that T1, T2, and T3 overlay one another. (B) UMAP visualization revealing the different clusters identified prior to integration with scRNA-seq. (C) Heatmap showing the proportion of cells of each cluster within each sample (T1, T2, and T3). The color scale represents the cell proportion within each cluster. (D) UMAP visualization of clusters after integration of scATAC-seq data with scRNA-seq data. Most cell types and cell subtypes were remapped, including the proximal, medial, and distal cells as well as the muscle, neuronal, immune, and stem-cell like cells of the PD axis. (E) Feature plots showing the known marker genes, and the respective cell types and cell subtypes identified in the UMAP visualization after scRNA-seq data integration. (F) Heatmap of important motifs in each cluster. (G) Gene Ontology analysis of the chromatin accessible genes of T2 and T3 relative to those of T1 showed

 doi:10.20944/preprints202205.0378.v1

many metabolic processes occurring in early T2, while many chitin-based cuticle development processes occurred in late T3. (H) Genome tracks of distal marker genes (*Dll* and *C15*) revealed high co-accessibility in neighboring genes.

2.3.1. Chromatin accessibility differentiated the T1, T2 and T3 distal cell functions

Gene set enrichment of T2 genes relative to T1 and T3 genes relative to T1 showed differences in cellular processes (**Figure 4G**). The T2 distal cells were more involved in metabolic processes, while the T3 distal cells had a larger role in chitin-based larval cuticle development.

Fragment coverage within the 40,000 base pairs on either side of the distal cell marker gene *Dll* showed increased coverage in the distal cells with marked co-accessibility in neighboring genes (**Figure 4H**). Distal cell marker gene *C15* similarly displayed increased coverage in the distal cells with marked co-accessibility in neighboring genes.

3. Discussion

We used scRNA-seq and scATAC-seq to explore the *Drosophila* L3 disc transcriptomic and epigenomic landscapes, respectively, at three timepoints of development. The multi-omics datasets corroborated each other and showed similar cell types that delineated the various regions of the leg disc, namely, those along the PD axis. Moreover, scRNA-seq identified an experimentally validated late-stage distal-specific and conserved lncRNA (*lncRNA:CR33938*) that, upon further characterization by overexpression studies, promoted distal leg growth gene expression. In addition, differences in chromatin accessibility determined by scATAC-seq indicated the disparate functions of early- and late-stage distal cells.

Given that the three legs of *Drosophila* differ in their developmental programs, their underlying differences cannot be ignored when studying leg disc development [21]. Subsequently, we specifically isolated the third leg disc to provide a more coherent single-cell atlas.

Simultaneous multi-omics library preparation methods, where the same cell or nuclei are used for different single-cell assays, were not available at the time these experiments were completed. As a result, the same cell suspension was used for both scRNA-seq and scATAC-seq to minimize biological variation. Furthermore, the limited number of cells extracted per leg disc prevented the execution of multiple experiments of biological replicates, in which one experiment consisted of one replicate. Rather, a single assay comprised of many biological replicates was conducted for each time point.

The computationally determined assignment of cell types to clusters depended upon the most upregulated genes in each cluster and prior information about cell type-specific marker genes. In addition to the prominent distal, medial, and proximal cell types, cells that did not express explicit marker genes denoting specific cell types represented early developing cells with undetermined fate, which we referred to as “stem-cell like cells.”

We found a large cluster of epithelial cells and a smaller cluster of muscle cells in the L3 discs, which corroborated previous studies that have shown the presence of many epithelial cells and accompanying muscle cells in the wing discs of third-instar larva [15]–[17]. Previous studies have suggested that the epithelial cells of the wing disc can be mapped to distinct subregions, including the pouch, hinge, notum, and peripodial membrane [16]. Similarly, the epithelial cells of the leg disc could be mapped to distinct proximal, medial, and distal subregions. Regarding muscle cells, research has shown that they can be subcategorized into direct and indirect flight muscles [17]. Given this finding, we also subcategorized the L3 disc muscle cells based on early versus late muscle development genes.

Our results also demonstrated the presence of neuronal and immune cells in L3 discs, which corroborates the recent single-nucleus transcriptomics study on the adult fruit fly leg by the Fly Cell Atlas showing the presence of various differentiated neurons as well

 doi:10.20944/preprints202205.0378.v1

as hemocytes and glial cells [20]. This illustrated that the neuronal cells in the developing leg disc have not yet differentiated, though they can subsequently differentiate.

Despite the publication of several works on single-cell transcriptomic landscapes of the *Drosophila* wing and eye-antennae imaginal discs [14]–[19], this study is the first to describe the transcriptomic and epigenomic landscape of the leg disc, specifically the third leg disc, at single-cell resolution. The Fly Cell Atlas study determined the single-nucleus transcriptomic atlas of the adult fruit fly leg, but it was based on fully differentiated tissues [20]. Conversely, our work was based on developing tissue and characterized the importance of an identified lncRNA.

lncRNAs tend to have lower expression levels than protein coding genes [22]. The detection of *lncRNA:CR33938* by our polyA-tailed single-cell transcriptomics assay indicated that it had a robust level of expression and suggested that it had an important physiological function in leg development given that lncRNA expression is environment specific [22].

Our work also highlighted the spatiotemporal function of *lncRNA:CR33938* and supported its importance to late-stage distal leg development. We used a previously published list of larval stage genes that establish the PD axis of the *Drosophila* leg [23]. Prior to this study, *lncRNA:CR33938* did not have an annotated function, but our study indicated that it may influence late-stage distal, and perhaps mid-stage medial, leg growth. In future research, we suggest spatiotemporally modulating the expression of *lncRNA:CR33938* in *Drosophila* to decipher its effect on leg phenotype.

4. Materials and Methods

Fly maintenance and stocks

The w^{1118} *Drosophila melanogaster* fly line was obtained as a gift from Prof. Edwin Chan's lab at The Chinese University of Hong Kong. All flies were maintained at room temperature in regular light-dark cycles in vials containing standard cornmeal agar medium (Nutri-fly, #66-112).

Fly breeding schedule for T1, T2, and T3

Male and female w^{1118} flies were allowed to mate for 2 hours at room temperature in a clear plastic cup with an attached petri dish containing apple juice agar. After this time had elapsed, embryos were transferred from the apple juice agar plate to a vial containing standard cornmeal agar medium. They were then allowed to grow for 121, 133 or 168 hours, corresponding to T1, T2, and T3, after which the third leg discs of these third-instar larvae (L3) were dissected.

Third leg disc dissection and single cell dissociation

At least 70 L3 discs were dissected for T1, and at least 50 L3 discs each were dissected for T2 and T3. These discs were collected in an Eppendorf tube containing phosphate-buffered saline (PBS) with 0.04% bovine serum albumin (BSA) on ice. After pipetting out the PBS from briefly centrifuged samples, we added TrypLE Select Enzyme (10X) (ThermoFisher, #A1217702). The discs were then incubated in a thermomixer shaken at 500 rpm for 25 minutes at 37°C (with the tube being flicked every five minutes). S2 medium (Gibco, #21720) supplemented with 10% fetal bovine serum and 2% penicillin/streptomycin were then added to stop the dissociation reaction. Finally, the isolated single cells were washed and resuspended in PBS + 0.04% BSA.

DNA library preparation and sequencing

The complementary DNA (cDNA) libraries for T1, T2, and T3 were prepared according to the 3' scRNA-seq library preparation protocol (v3.1) of 10x Genomics. In summary, a microfluidics chip was used to produce GEMs (Gel Bead-in-Emulsions), which are droplets that each contain a single microbead with attached oligonucleotides that include a

 doi:10.20944/preprints202205.0378.v1

unique cell barcode, a single cell, and reverse transcription reagents. When the single cell lyses within the intact GEM, the cellular polyA-tailed transcript sequences become exposed, reverse transcription occurs, and each cDNA transcript within the same cell receives the same cell barcode with a different UMI (unique molecular identifier). Subsequently, the droplets lyse and the cell-barcoded cDNA from all cells are pooled and amplified. cDNA library construction involved fragmentation, end-repair, A-tailing, double-sided size selection, and sample index incorporation. Quality control and qualitative analysis of the final library was performed on an Agilent Bioanalyzer DNA High Sensitivity chip. Sequencing of the libraries was completed on the Illumina NovaSeq6000 platform by Novogene.

scRNA-seq raw data processing, quality assessment and filtering

The raw paired-end sequencing data files (Fastq) were processed using the Cell Ranger pipeline v4.0.0 with default settings. Read alignment and UMI counts were based on a BDGP6 genome reference fasta file and annotated by a BDGP6.28 gtf file developed by Ensembl. Cell-UMI count tables were loaded into Seurat v4.0 [24]. Cells with 1,000-250,000 UMI counts and less than 5% mitochondrial genes were used as filtering gates to select cells for downstream analysis. We only kept genes with at least 20 UMI counts in all cells. Qualimap (v2.2.1) was further used to assess the percentage of mapped reads and percentage of mapped reads aligned to genes for comparison between T1, T2, and T3.

scRNA-seq data integration, clustering, and cell type identification

The T1, T2, and T3 filtered single-cell datasets were merged and integrated using Seurat (v4.0.) Batch effects between samples were corrected using Harmony (v1.0) prior to clustering analysis. PCA was used to determine the optimal dimension for dimensionality reduction, and clustering was performed based on K-nearest neighbor (KNN) graphs with a resolution of 0.02 before UMAP visualization of the single-cell data in two dimensions. The major cell types of the clusters were identified based on known marker genes, and these marker genes were listed among the most upregulated differentially expressed genes compared to other clusters. All clusters were further subclustered into constituent cells based on known marker genes. Dotplots, featureplots, heatmaps, and UMAPs were then generated. Other than the known marker genes, novel genes were also identified as potential markers.

Validation of scRNA-seq results by FISH and confocal imaging

w1118 flies were bred and T1, T2, and T3 L3 discs were dissected as described above. The discs were fixed in 3.7% paraformaldehyde on ice for 30 minutes. Then, probe hybridization was completed according to the protocol provided by Molecular Instruments. The discs were first permeabilized in a detergent solution containing sodium dodecyl sulfate and Tween-20, before custom designed probes for *Dll*, *dac*, and *lncRNA:CR33938* were hybridized to the fixed and permeabilized discs for 20 hours at 37°C. After several washes with 5X SSC-Tween-20, hairpins with different fluorophores for each probe were added and incubated for 16 hours in darkness at room temperature. The discs then underwent another several washes with 5X SSC-Tween-20 and were mounted onto a Menzel-Glaser Superforst Plus microscope slide (Thermo Scientific #J1800AMNZ) with a Hydromount mounting medium (National Diagnostics #HS-106) and a 22 x 50 mm Deckglaser microscope coverglass (VWR #630-1461). The mounts were visualized on a Leica SP8 confocal microscope and each sample was imaged every 0.25 µm along the z-axis. The confocal images were z-stacked and processed with Leica Application Suite software.

Construction of lncRNA:CR33938 expression vector for expression studies

Total RNA was extracted from *D. melanogaster* L3 discs using NucleoZOL (Macherey-Nagel #740404.200) following the manufacturer's protocol. Following RNase-free DNaseI (Thermo Scientific #EN0521) treatment and DNaseI inactivation by EDTA, the purified

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 May 2022 doi:10.20944/preprints202205.0378.v1

RNA was subjected to cDNA generation using PrimeScript II (Takara #RR036A). The cDNA concentration was measured using the Qubit High Sensitivity double-stranded DNA assay. *lncRNA:CR33938* was amplified with PCR from the cDNA using the following primers with restriction site sequences inserted: forward primer 5′- TTTGGTACCTTGAGTCCGAGAGGTT -3′ and reverse primer 5′- CGCTCTAGACTCTTTTTTTGGTAGCCTATT -3′). The amplicon and the pAc5.1/V5-His B expression vector (Invitrogen #V411020) were digested with KpnI (New England Biolabs #R3142) and XbaI (New England Biolabs #R0145) restriction enzymes and subsequently ligated using T4 DNA ligase (Invitrogen #15224017). The ligation mixture was transformed to chemically competent *Escherichia coli* (Invitrogen #C404003) and selected using 100 mg/ml of ampicillin. The sequence of the *lncRNA:CR33938* construct cloned into the expression vector was verified by Sanger sequencing at the Beijing Genomics Institute. Transfection-ready plasmid DNA was extracted using a Plasmid Miniprep kit (Invitrogen #K210011).

S2 cell culture and transfection

D. melanogaster S2 cells were provided by Prof. Jerome Hui from the School of Life Sciences of The Chinese University of Hong Kong. S2 cells were cultured in Schneider's Drosophila Medium (Gibco #21720) supplemented with 10% heat inactivated fetal bovine serum (Gibco #10270) and 1% penicillin-streptomycin antibiotic mixtures (Gibco #15140122) in a 25°C humidified incubator. The cloned pAc5.1-lncRNA:CR33938 construct was transfected into S2 cells using Effectene (Qiagen #301425) and the pAc5.1 backbone vector was used as a negative control. The cells were then incubated at 25°C for 48 hours prior to RNA extraction for qRT-PCR.

RNA extraction and qRT-PCR of S2 cells

RNA was extracted with NucleoZOL (Macherey-Nagel #740404.200). Following RNase-free DNaseI (Thermo Scientific #EN0521) treatment and DNaseI inactivation by EDTA, the purified RNA was subjected to cDNA generation using PrimeScript II (Takara #RR036A). The cDNA concentration was measured with a Qubit High Sensitivity double-stranded DNA assay. For qRT-PCR, 1 ng of template cDNA and 1xTB Green II (Takara #RR820) were added to each well of a 96-well plate (Axygen #PCR-96-FSC) and covered with an optical adhesive film (Applied Biosystems ABI #4311971) prior to execution on a BioRad CFX96 real-time PCR detection system. Primer sequences for each tested gene are listed in **Table S2**.

Nuclei isolation for scATAC-seq

The same suspension of single cells used for scRNA-seq was used for nuclei isolation for scATAC-seq. Nuclei isolation was performed according to a 10X Genomics low input protocol for scATAC-seq with some optimizations. The cell suspension was pelleted and lysed on ice in a buffer containing the detergents Tween-20 and nonidet-P40 (NP40) for 30 seconds. The isolated nuclei were then washed twice and resuspended in chilled 10X diluted nuclei buffer (provided by 10x Genomics). Trypan blue stained nuclei were observed under the microscope to assess nuclei quality.

DNA library preparation and scATAC-seq

DNA library preparation was performed according to the scATAC-seq preparation protocol (v1.1) of 10x Genomics. First, the nuclei suspensions were incubated in a transposition mix that included a transposase that preferentially fragments the DNA in open regions of the chromatin. Simultaneously, adapter sequences were added to the ends of the DNA fragments. As in scRNA-seq, a microfluidics chip was used to produce GEMs (Gel Bead-in-Emulsions), but in this case the droplets contained a single microbead with attached sequences consisting of a unique cell barcode, a single nuclei, and DNA amplification reagents. Once the DNA from each nucleus was barcoded, all nuclei were pooled for DNA library construction. Because only the histone unbound areas of the genome are

 doi:10.20944/preprints202205.0378.v1

cut by the transposase, the library consisted of DNA fragments that represented the open chromatin regions of the genome. Quality control and qualitative analysis of the final library was performed on an Agilent Bioanalyzer High Sensitivity DNA chip. The libraries were sequenced on the Illumina NovaSeq6000 platform by Novogene at PE50 with a sequencing depth of approximately 50,000 read pairs per cell.

scATAC-seq data analysis

The raw paired-end sequencing data were processed by Cell Ranger ATAC pipeline v2.0.0 with default settings, using a dm6 UCSC reference generated by the 10X Genomics mkref function. Data processing, filtering, dimensionality reduction, and clustering was performed with ArchR v1.0.1 [25]. UMAP visualizations of the scATAC-seq clusters were created before and after integration with scRNA-seq data. Determination of cell type identities were aided by manual annotation of cell type-specific marker genes based on gene scores estimated from the chromatin accessibility data. Peak calling with MACS2 v2.2.7.1 [26] was performed on each cell cluster. Identification of robust peak sets allowed prediction of enriched transcription factor motifs for each cluster. Gene ontology analysis was performed using Cluster Profiler to determine the enriched distal process in T2 and T3 relative to T1. Genome browser plots depicting co-accessibility of distal genes with nearby genes was also generated using ArchR.

Supplementary Materials: Table S1: Sequencing statistics among the T1, T2 and T2 samples; Table S2: List of qRT-PCR primer sequences of *Drosophila* leg development genes; Figure S1: Expression of muscle markers across T1, T2 and T3; Figure S2: Heatmap of top upregulated genes in the early and late muscle cells; Figure S3: Heatmap of top upregulated genes in the early and late neuronal cells; Figure S4: Heatmap of top upregulated genes in the various immune cells.

Author Contributions: Conceptualization: T.F.C. and J.H.; methodology: J.T.; validation: J.T., T.H.Li and A.Lee; formal analysis: J.T..; investigation: J.T., T.H.Li, A.Lee, I.L., Z.Q. and X.Lin; data curation: J.T. and J.Z.; writing—original draft preparation: J.T..; writing—review and editing: T.F.C. and J.H.; visualization: J.T.; supervision: T.F.C. and J.H.; project administration: J.T.; funding acquisition: T.F.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by System Information Science fund from The Chinese University of Hong Kong Faculty of Science, Donation from Mr. and Mrs. Sunny Yang, The Chinese University of Hong Kong Direct grant 4053486, and Innovation and Technology Fund (to the State Key Lab of Agrobiotechnology.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The sequencing data presented in this study are openly available in NCBI SRA with BioProject ID PRJNA831899.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] S. U. Schmitz, P. Grote, and B. G. Herrmann, "Mechanisms of long noncoding RNA function in development and disease," *Cell. Mol. Life Sci. 2016*, vol. 73, no. 13, pp. 2491–2509.

[2] G. S. Laurent, C. Wahlestedt, and P. Kapranov, "The landscape of long non-coding RNA classification: The non-coding RNA universe," *Trends Genet. 2016*, vol. 31, no. 5, pp. 239–251.

[3] I. Sarropoulos, R. Marin, M. Cardoso-Moreira, and H. Kaessmann, "Developmental dynamics of lncRNAs across mammalian organs and species," *Nature 2019*, vol. 571, no. 7766, pp. 510–514.

[4] M. Esteller, "Non-coding RNAs in human disease," *Nat. Rev. Genet. 2011*, vol. 12, no. 12, pp. 861–874.

[5] M. D. Adams *et al.*, "The genome sequence of *Drosophila melanogaster*," *Science 2000*, vol. 287, no. 5461, pp. 2185–2195.

[6] U. B. Pandey and C. D. Nichols, "Human disease models in *Drosophila melanogaster* and the role of the fly in therapeutic drug discovery," *Pharmacol. Rev. 2011*, vol. 63, no. 2, pp. 411–436.

[7] J. V. Beira and R. Paro, "The legacy of *Drosophila* imaginal discs," *Chromosoma 2016*, vol. 125, no. 4, pp. 573–592.

[8] S. Aldridge and S. A. Teichmann, "Single cell transcriptomics comes of age," *Nat. Commun. 2020*, vol. 11, no. 1, pp. 1–4.

[9] F. Tang *et al.*, "mRNA-Seq whole-transcriptome analysis of a single cell," *Nat. Methods 2009*, vol. 6, no. 5, pp. 377–382.

[10] B. Hwang, J. H. Lee, and D. Bang, "Single-cell RNA sequencing technologies and bioinformatics pipelines," *Exp. Mol. Med. 2018*, vol. 50, no. 8, p. 96.

[11] G. Chen, B. Ning, and T. Shi, "Single-cell RNA-seq technologies and related computational data analysis," *Front. Genet. 2019*, vol. 10, no. , pp. 1–13.

[12] V. Svensson, E. da Veiga Beltrame, and L. Pachter, "A curated database reveals trends in single-cell transcriptomics," *Database 2020*, vol. , pp. 1–7.

[13] A. M. Klein and B. Treutlein, "Single cell analyses of development in the modern era," *Dev. 2019*, vol. 146, no. 12, pp. 1–3.

[14] M. M. Ariss, A. B. Islam, M. Critcher, M. P. Zappia, and M. V. Frolov, "Single cell RNA-sequencing identifies a metabolic aspect of apoptosis in Rbf mutant," *Nat. Commun. 2018*, vol. 9, no. 1, p. 5024.

[15] J. Bageritz, P. Willnow, E. Valentini, S. Leible, M. Boutros, and A. A. Teleman, "Gene expression atlas of a developing tissue by single cell expression correlation analysis," *Nat. Methods 2019*, vol. 16, no. 8, pp. 750–756.

[16] M. Deng *et al.*, "Single cell transcriptomic landscapes of pattern formation, proliferation and growth in *Drosophila* wing imaginal discs," *Dev. 2019*, vol. 146, no. 18,.

[17] M. P. Zappia, L. de Castro, M. M. Ariss, H. Jefferson, A. B. Islam, and M. V Frolov, "A cell atlas of adult muscle precursors uncovers early events in fibre-type divergence in *Drosophila*," *EMBO Rep. 2020*, vol. 21, no. e49555.

[18] C. B. González-Blas *et al.*, "Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics," *Mol. Syst. Biol. 2020*, vol. 16, no. e9438.

[19] N. J. Everetts, M. I. Worley, R. Yasutomi, N. Yosef, and I. K. Hariharan, "Single-cell transcriptomics of the *Drosophila* wing disc reveals instructive epithelium-to-myoblast interactions," *Elife 2021*, vol. 10, no. e61276.

[20] H. Li *et al.*, "Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly," *Science 2022*, vol. 375, no. 6584.

[21] G. Schubiger, M. Schubiger, and A. Sustar, "The three leg imaginal discs of *Drosophila*: 'Vive la différence,'" *Dev. Biol. 2012*, vol. 369, no. 1, pp. 76–90.

[22] Q. Xu *et al.*, "Systematic comparison of lncRNAs with protein coding mRNAs in population expression and their response to environmental change," *BMC Plant Biol. 2017*, vol. 17, no. 1, pp. 1–15.

[23] N. Grubbs, M. Leach, X. Su, T. Petrisko, and J. B. Rosario, "New components of *Drosophila* leg development identified through genome wide association studies," *PLoS One 2013*, vol. 8, no. 4, p. e60261.

[24] Y. Hao *et al.*, "Integrated analysis of multimodal single-cell data," *Cell 2021*, vol. 184, no. 13, pp. 3573-3587.e29.

[25] J. M. Granja *et al.*, "ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis," *Nat. Genet. 2021*, vol. 53, pp. 403–411.

[26] J. M. Gaspar, "Improved peak-calling with MACS2," *bioRxiv*, pp. 1–16, 2018.