Thesis for the Doctor of *philosophy*

Exploring α-Arrestin Interactomes in Human and Drosophila and Investigating the Transcriptomic Landscape in Drosophila Hematopoiesis and Immune Response

KyungTae Lee

Graduate School of Hanyang University

August 2023

Thesis for the Doctor of *philosophy*

Exploring α-Arrestin Interactomes in Human and Drosophila and Investigating the Transcriptomic Landscape in Drosophila Hematopoiesis and Immune Response

Thesis Supervisor: Jin-Wu Nam

A Theis submitted to the graduate school of Hanyang University in partial fulfillment of the requirements for the degree of *Doctor of Philosophy*

KyungTae Lee

August 2023

Department of Life Science

Graduate School of Hanyang University

This thesis, written by KyungTae Lee,
has been approved as a thesis for the Doctor of *Philosophy.*

August 2023

Committee Chairman: Jiwon Shim (Signature)

Committee member: Young Kwon (Signature) Young Kwon

Committee member: Jin-Wu Nam (Signature)

Committee member: Jeong-Yeon Lee (Signature)

Committee member: Hyobin Jeong (Signature)

Graduate School of Hanyang University

Table of contents

List of Figures

List of tables

Abstract

Exploring α-Arrestin Interactomes in Human and Drosophila and Investigating the Transcriptomic Landscape in Drosophila Hematopoiesis and Immune Response

Kyung-Tae Lee
Department of Life science
The graduate School
Hanyang University

This research explores the intricacies of α-arrestins and their protein-protein interactions, as well as investigates the *Drosophila* immune system and the potential influences of non-coding RNAs on lamellocyte development.

The α-arrestins are evolutionarily conserved modulators that have been reported to control diverse signaling pathways, particularly G-protein coupled receptors. A few mammalian α-arrestins and those conserved in yeast and *Drosophila* have been studied of their protein interactors and functions. However, substantial part of interactome and biological functions of α-arrestins from diverse species remain largely uncharacterized. Employing affinity purification and mass spectrometry, we constructed protein-protein interaction networks for six human and twelve *Drosophila* α-arrestins. This analysis yielded high-confidence interactomes with hundreds of prey proteins for each species, indicating conserved and species-specific interactions. Notably, we found that the human α-arrestins ARRDC3 and *Drosophila* α-arrestins Vdup1 and CG18746 interacts with orthologous proteins involved in RNA splicing. Analysis of RNA-seq of HeLa cells under ARRDC3, TXNIP and splicing factors depleted conditions showed that perturbation of ARRDC3 influences certain type of alternative RNA splicing, a degree of which was comparable to splicing factor depleted conditions. In addition to conserved interactome, we found that human α-arrestins, TXNIP, influences chromatin structures and transcription signals by obstructing HDAC2 recruitment. Additionally, analysis of the interactome for the uncharacterized human α-arrestins ARRDC5 revealed a link to the key bone resorption regulator, V-type ATPase.

Meanwhile, we also focused on the *Drosophila* immune system, particularly hemocytes and their progenitors, prohemocytes. With the use of both Illumina short- and Nanopore long-read sequencing, we generated integrative hybrid transcriptomes, leading to the identification of novel non-coding RNAs, distinctly expressed in lamellocytes that are induced upon wasp infestation. Furthermore, we noticed a potential shift in alternative splicing and isoform usage under infested conditions, which we are further investigating at the bulk and single-cell levels. Fusion genes identified from our long-read RNA-seq data are also currently under experimental validation.

In summary, our research provides a valuable resource for understanding α-arrestins's cellular functions, discovers novel non-coding RNA markers in response to immune challenges, and illuminates the changes in alternative splicing and isoform usage in *Drosophila* larvae.

General introduction

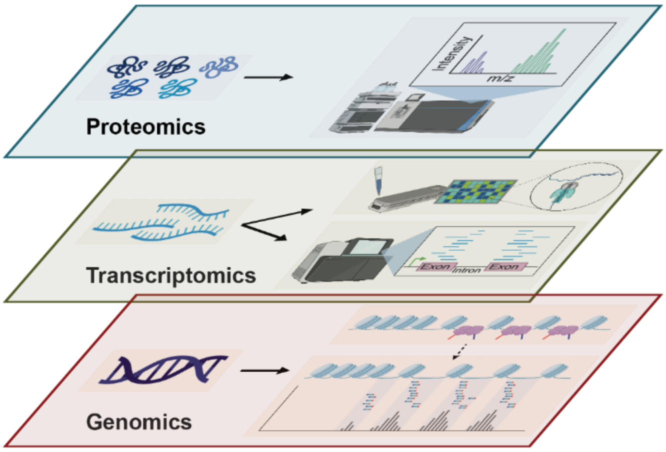
1. High-throughput multi-omics: Transforming our understanding of biological systems in modern biology

The advancement of high-throughput technologies has revolutionized biological research by enabling comprehensive analyses of biological systems at unprecedented scale and resolution. High-throughput technologies are extensively employed across various multi-omics studies including genomics, transcriptomics, and proteomics. These approaches provide valuable insights into the complex interplay of different biological molecules in diverse organisms.

Among the high-throughput technologies, introduction of next-generation sequencing (NGS) has brought paradigm shift in genomics and transcriptomics studies by offering high-throughput, cost-effective solutions for large-scale DNA and RNA sequencing. NGS have facilitated the exploration of genomic and transcriptomic landscape including genetic variants associated with diverse disease and gene expression patterns in multitude of biological systems (Goodwin, McPherson, & McCombie, 2016). At present, there are two main streams of NGS: short-read sequencing and long-read sequencing. Short-read NGS technologies have at first provided low-cost and high-accuracy data that are useful for studies of large population or biological specimens. These short-read sequencing data have limitations, however, in resolving complex genomic regions or identifying large structural variations (Goodwin et al., 2016). On the other hand, long-read NGS technologies, such as Oxford Nanopore’s long-read sequencing, can read longer DNA or RNA fragments and became a powerful tool for assembly of complex genomes and detection of large structural variations (Sedlazeck, Lee, Darby, & Schatz, 2018) (Jain et al., 2018). However, long-read NGS technologies also have some limitations such as lower-accuracy and relatively higher cost compared to short-read NGS technologies (Goodwin et al., 2016).

On the other side of the multi-omics spectrum, significant advancement has also been brought in proteomics field, especially with maturation of mass spectrometry (MS)-based techniques. MS is an analytical technique used to identify and quantify molecules, ranging from small molecules to large proteins and complex mixtures, based on their mass-to-charge ratio (m/z), enabling researchers to identify and analyze diverse proteins of interest in a great depth with higher sensitivity (Aebersold & Mann, 2016). Combined with affinity purification (AP), which is a technique used in biochemistry to isolate a specific molecule of interest, AP/MS is widely used to identify and characterize protein-protein interactions, allowing the comprehensive study of protein complexes in various biological contexts.

In this study, we employed high-throughput multi-omics data including AP/MS, short- and long-read RNA-seq in both bulk and single-cell level and assay for transposase-accessible chromatin using sequencing (ATAC-seq) to decipher protein-protein interaction network of α-arrestin family of proteins in human and *Drosophila* and transcriptomic landscape of *Drosophila* larvae under immune responses (Introduction Figure 1).



Introduction Figure 1. Employing high-throughput multi-omics data to decipher biological questions.

2. From interaction to function: The critical role of protein-protein interactions

Proteins rarely act and function in isolation. Instead, they interact with one another, forming complex networks in coordinated manner that underpin various biological functions (Alberts, 1998). The study of these protein-protein interactions (PPIs) and their resulting intricate networks has provided invaluable insights into the molecular mechanisms of biological processes to scientific community. By studying PPIs, our understanding of the structural basis of protein functions and biological principles of signaling cascades has greatly expanded and led to identification of mechanisms underlying diverse disease models and discovery of potential therapeutic targets (Vidal, Cusick, & Barabasi, 2011) (Fields & Song, 1989).

As noted in the previous section, recent advancements in high-throughput technologies in proteomics such as mass-spectrometry have enabled the systematic mapping of PPI networks on a genome-wide scale. Along with advancement in experimental techniques to generate high-throughput proteomics data, introduction of computation methodologies to analyze these data has enabled construction and characterization of comprehensive PPIs. Among them, significance analysis of interactome (SAINT) (Choi et al., 2011) is a computational tool used for the probabilistic scoring of AP/MS data. It aims to distinguish true PPIs from nonspecific background noise using the Bayesian approach. SAINTexpress (Teo et al., 2014), which is an updated version of original SAINT tool, was employed for identification of true interactors of α-arrestin in human and *Drosophila*.

Nonetheless, a few challenges remain in identification of high-confident PPIs. One primary concern is the high rate of false positives and negatives in PPI detection. To overcome this in this study of α-arrestin interactome in human and *Drosophila*, we selected SAINTexpress score thresholds that correspond to low false discovery rates, in conjunction with spectral count cutoffs that demonstrate high reproducibility between replicate experiments. The authenticity of identified PPIs involving α-arrestins has been confirmed through validations against known PPIs from a variety of sources. The validation process also included evaluating the affinities between short Iinear motifs and protein domains.

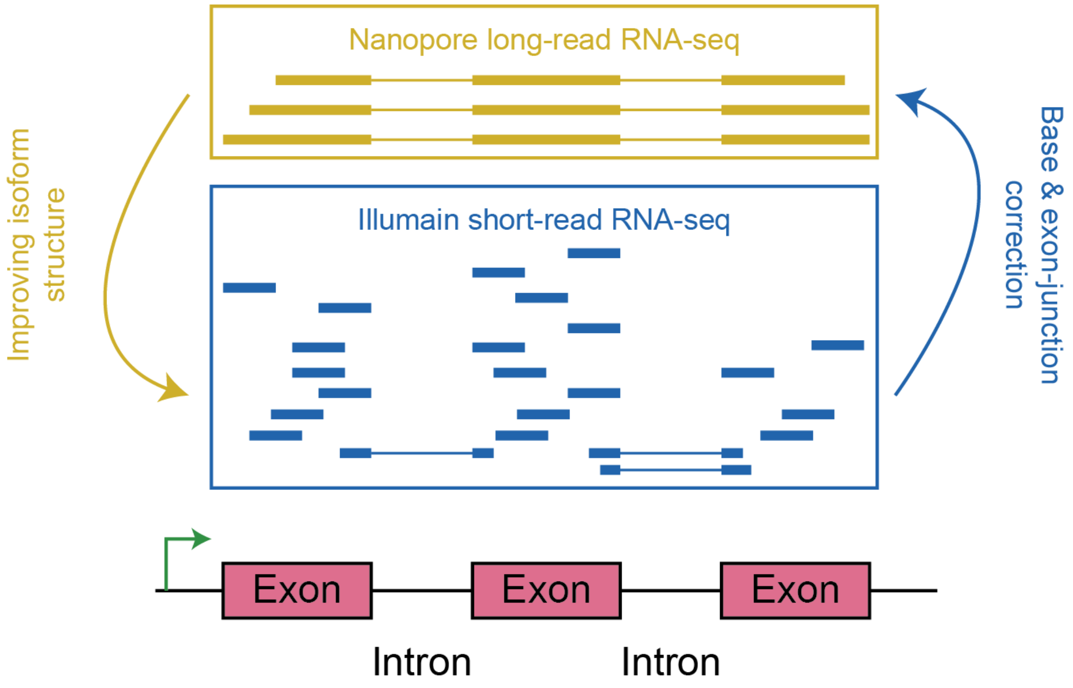
3. Harnessing short- and long-read RNA-seq data from comprehensive transcriptome assembly and lncRNA discovery

Transcriptome assembly is a process of reconstructing full-length transcripts from sequenced RNA fragments. This process is critical in transcriptomics filed as it delineates gene structures, identify isoforms, and discovers novel transcripts (Martin & Wang, 2011). Traditionally, transcriptome assembly has relied on short-read sequencing technologies, such as Illumina, which often generates millions of highly accurate but short reads (100~300 base pairs (bp)). Assembling these reads into complete transcripts is computationally challenging and can often result in false positives due to the complex nature of transcriptome, especially in Eukaryotes (Steijger et al., 2013).

The recent development of third-generation sequencing technologies, such as Oxford Nanopore Technologies (ONT), provide an opportunity to overcome these challenges of short-read sequencing as they can produce longer reads without fragmentation of RNA molecules. Sequencing reads from this third-generation sequencing technologies can span entire transcript, thereby improving the accuracy of transcriptome assemblies (Workman et al., 2019).

Long non-coding RNAs (lncRNAs) represent a class of transcripts that are at least 200 nucleotides long and not capable of producing proteins. They have been implicated in a diverse array of biological processes, including gene expression regulation, chromatin remodeling, and cellular differentiation (Quinn & Chang, 2016). Despite their biological significance, the full catalog of lncRNAs and their variant isoforms are not completely known, especially in organisms other than human and mouse. This incomplete knowledge of lncRNAs can be attributed to their typically low expression levels compared to protein-coding genes and the difficulties in detection and characterization of full-length transcript structure using short-read sequencing data (K. C. Wang & Chang, 2011).

By combining short- and long-read sequencing technologies, I leveraged the strength of both platforms, generating more accurate and comprehensive transcriptome assemblies that include a more complete repertoire of lncRNAs in hemocytes of Drosophila larvae under immune responses (Introduction Figure 2).


Nanopore long-read RNA-seq
Illumain short-read RNA-seq
Improving isoform structure
Base & exon-junction correction
Exon
Exon
Exon
Intron
Intron

Introduction Figure 2. Complementary hybrid RNA sequencing approach to identify and characterize accurate transcript isoforms.

4. Multi-faceted roles of α-arrestins in cellular processes

α-arrestins constitute a family of proteins that are highly conserved across eukaryotes, from yeast to humans (Alvarez, 2008). Initially discovered in yeast, α-arrestins have emerged as key regulatory components involved in the endocytosis and post-endocytic trafficking of plasma membrane proteins (Lin, MacGurn, Chu, Stefan, & Emr, 2008). While ß-arrestins, which are another subfamily of arrestin proteins, have been extensively studied for their role in turning off the G-protein-coupled-receptor signaling pathway such as ß-adrenergic signaling through receptor desensitization and internalization (Benovic, DeBlasi, Stone, Caron, & Lefkowitz, 1989) (Lohse, 1992) (Shenoy & Lefkowitz, 2011), α-arrestins have more recently come into focus.

These α-arrestins were first studied in conjunction with ß-arrestins in regulation of ß2AR. ARRDC3 was reported to act as an adaptor protein for ubiquitination of ß2AR by recruiting the neural precursor-cell-expressed developmentally downregulated gene 4 (NEDD4) through its conserved PPXY motifs (Nabhan, Pan, & Lu, 2010). Two subsequent studies also showed the involvement of α-arrestins, especially ARRDC3 and ARRDC4, in receptor desensitization and trafficking ß2AR, although they disagreed on when α-arrestins move into action (Shea, Rowell, Li, Chang, & Alvarez, 2012) (S. O. Han, Kommaddi, & Shenoy, 2013). Besides ß2AR, α-arrestins were reported to be involved in trafficking and sorting of other GPCR or signaling molecule through post-translational modification, especially ubiquitination, such as degradation pathway of notch receptor by ARRDC1 and ARRDC3 (Puca & Brou, 2014).

Aspects of α-arrestin in diseases and therapeutics have also been inspected in many studies. Among the mammalian α-arrestins, thioredoxin-interacting protein (TXNIP) was reported to directly interact with thioredoxin (TXN), which is essential component of system of cellular redox, to inhibit its activity as antioxidant (Patwari, Higgins, Chutkow, Yoshioka, & Lee, 2006) (Junn et al., 2000) (Nishiyama et al., 1999). TXNIP was also reported to bind Nod-like receptor pyrin domain-containing protein 3 (NLRP3) inflammasome, which regulates innate immunity and is associated with inflammatory diseases (Zhou, Tardivel, Thorens, Choi, & Tschopp, 2010). As TXNIP plays major roles in metabolism and inflammation, effects of TXNIP in the progression of diseases such as diabetes mellitus (DM) and neurological disorders have been studies in many publications (reviewed in detail in (Qayyum, Haseeb, Kim, & Choi, 2021)). Cases of suppression of tumors by α-arrestins have also been reported. Initially, suppression of cell surface adhesion molecule, ß-4 integrin (ITGß4) which was shown to promote progression of breast cancer (Lu, Simin, Khan, & Mercurio, 2008) (Diaz et al., 2005), by direct binding of ARRDC3 to ITGß4 and inducing subsequent internalization, ubiquitination and degradation was observed (Draheim et al., 2010).

While a handful of α-arrestins have been studies of their PPIs and biological functions, many of them in diverse species remains unknown of their functions and interactome. To decipher biological functions of α-arrestins that are both conserved or unique to certain species, we have conducted AP/MS to identify and characterize PPI networks of α-arrestins in human and *Drosophila*. Some of conserved and unique biological functions and protein interactors of α-arrestins have been extensively validated through experimental approaches and computational analysis of high-throughput multi-omics data.

5. *Drosophila* as a model for studying hematopoiesis

Drosophila melanogaster serves as a powerful model organism to study a wide range of biological processes including formation of blood cells, hematopoiesis (Bier, 2005). Similar to vertebrates, *Drosophila* hematopoiesis is a highly regulated process that generates various types of blood cells known as hemocytes. These hemocytes perform critical functions in development and innate immunity, exhibiting functions analogous to those of mammalian myeloid cells (Honti, Csordas, Kurucz, Markus, & Ando, 2014).

In the *Drosophila* larva, hematopoiesis occurs predominantly in a specialized organ known as the lymph gland, which is divided into several distinct zones each harboring a particular type of hemocyte (Jung, Evans, Uemura, & Banerjee, 2005). The posterior signaling center (PSC), a group of cells within the primary lymph gland lobe, serves as a niche that maintains hemocyte precursors in an undifferentiated state (Mandal, Martinez-Agosto, Evans, Hartenstein, & Banerjee, 2007).

In response to developmental cues or immune challenges such as wasp infestation or wounds, these precursor cells differentiate into mature hemocytes, including plamatocytes, crystal cells, and lamellocytes. Each of these hemocytes plays a unique role: plasmatocytes are involved in phagocytosis, crystal cells act in melanization, and lamellocytes function in encapsulation of larger parasites (Gold & Bruckner, 2015).

Study of hematopoiesis in Drosophila larvae has shed light on the molecular and cellular mechanisms underlying blood cell development and function and has revealed intriguing parallels with vertebrate hematopoiesis (Banerjee, Girard, Goins, & Spratford, 2019). These insights have paved the way for a deeper understanding of blood cell biology and diseases such as leukemia, which involves dysregulated hematopoiesis (Banerjee et al., 2019).

Chapter I

Comparative Interactome Analysis of α-arrestin Families in Human and *Drosophila*

I-1 Abstract

The α-arrestins form a large family of evolutionally conserved modulators that control diverse signaling pathways, including both G-protein-coupled receptor- (GPCR-) mediated and non-GPCR mediated pathways, across eukaryotes. However, unlike **β**-arrestins, only a few α-arrestin targets and functions have been characterized. Here, using affinity purification and mass spectrometry, we constructed interactomes for six human and twelve Drosophila α-arrestins. The resulting high-confidence interactomes comprised 307 and 467 prey proteins in human and Drosophila, respectively. A comparative analysis of these interactomes predicted not only conserved binding partners, such as motor proteins, proteases, ubiquitin ligases, RNA splicing factors, and GTPase-activating proteins, but also those specific to mammals, such as histone modifiers and the subunits of V-type ATPase. Given the manifestation of the interaction between the human α-arrestin, TXNIP, and the histone-modifying enzymes, including HDAC2, we undertook a global analysis of transcription signals and chromatin structures that were affected by TXNIP knockdown. We found that TXNIP activated targets by blocking HDAC2 recruitment to targets, a result that was validated by chromatin immunoprecipitation assays. Additionally, the interactome for an uncharacterized human α-arrestin ARRDC5 uncovered multiple components in the V-type ATPase, which plays a key role in bone resorption by osteoclasts. Our study presents conserved and species-specific protein-protein interaction maps for α-arrestins, which provide a valuable resource for interrogating their cellular functions for both basic and clinical research.

I-２ Introduction

The discovery of first arrestin protein in retinal rods contributed to a deeper understanding of photoreceptor signaling mediated by rhodopsin, which is one of the G-protein-coupled receptor (GPCR) class. As its ability to arrest the GPCR signaling pathway, the protein was first named as “arrestin” (Kuhn, Hall, & Wilden, 1984; Wilden, Wust, Weyand, & Kuhn, 1986; Zuckerman & Cheasty, 1986). Shortly after this discovery of the first arrestin protein in the retina, another arrestin protein that specifically turns off β-adrenergic signaling was identified and named “β-arrestin”. This arrestin-mediated termination of signaling from GPCRs is called “receptor desensitization” (Benovic et al., 1989; Lohse, 1992; Shenoy & Lefkowitz, 2011), one of crucial cellular process in maintaining cellular homeostasis and preventing overstimulation of signaling pathways. Further studies have revealed that β-arrestins regulate the receptor desensitization of other signaling pathways through ubiquitination and regulation of trafficking of various cargo molecules (Y. M. Kim & Benovic, 2002; Malik & Marchese, 2010; Puca & Brou, 2014).

Another class of arrestin, α-arrestin, was first studied in fungi and yeast (Andoh, Hirata, & Kikuchi, 2002) and subsequently recognized as new class of arrestins (Boase & Kelly, 2004; Herranz et al., 2005). They contain characteristic arrestin domains, arrestin_N and arrestin_C, and PPXY motifs, which are unique to the α-arrestin clan. A phylogenetic study of arrestin proteins showed that α-arrestins are the ancestral class of the arrestin family and conserved from yeast to human (Alvarez, 2008). To date, six α-arrestins, arrestin domain containing protein 1 (ARRDC1), ARRDC2, ARRDC3, ARRDC4, ARRDC5, and thioredoxin-interacting protein (TXNIP), have been found to be in the human genome. These human α-arrestins were first studied in conjunction with β-arrestins in the regulation of the β2-adrenergic receptor (β2AR) in human cells. ARRDC3 and ARRDC4 works as an adaptor protein for the ubiquitination of β2AR by recruiting the NEDD4 protein, an E3 ubiquitin ligase, through its conserved PPXY motifs(S. O. Han et al., 2013; Nabhan et al., 2010; Shea et al., 2012).In addition to their β2AR-associated roles, α-arrestins are involved in trafficking and sorting of other GPCRs and signaling molecules through post-translational modifications, including ubiquitination. For example, ARRDC1 and ARRDC3 were reported to play roles in the degradation of the Notch receptor (Puca & Brou, 2014) and in the ubiquitination of ALG-2-interacting protein X (ALIX) (Dores, Lin, N, Mendez, & Trejo, 2015). ARRDC1 contains a PSPA motif, which binds the tumor susceptibility gene 101 (TSG101) protein, an essential component of an endosomal sorting complex. ARRDC1 also recruits E3 ligases, such as WW domain-containing E3 ubiquitin protein ligase2 (WWP2), inducing ubiquitination of itself and the subsequent release of ARRDC1-associated microvesicles (Nabhan, Hu, Oh, Cohen, & Lu, 2012). Another well-known α-arrestin, TXNIP, was first named as vitamin D3-upregulated protein 1 (VDUP1) after verification that its gene is a vitamin D3 target in cancer cells (K. S. Chen & DeLuca, 1994; Qayyum et al., 2021). Since then, TXNIP had been reported to directly interact with thioredoxin, which is an essential component of the cellular redox system, to inhibit its activity as an antioxidant (Junn et al., 2000; Nishiyama et al., 1999; Patwari et al., 2006). TXNIP was also reported to inhibit glucose uptake by inducing the internalization of glucose transporter 1 (GLUT1) through clathrin-mediated endocytosis and by indirectly reducing GLUT1 RNA levels (Wu et al., 2013). Although the TXNIP is known to be localized in both cytoplasm and nucleus, biological functions of TXNIP have been mostly explored in cytoplasm but remained poorly characterized in nucleus.

A few α-arrestins appear to have evolutionarily conserved functions in both human and invertebrates. For instance, the Hippo signaling pathway, which impacts a variety of cellular processes such as metabolism, development, and tumor progression (Mo, Park, & Guan, 2014; Pei et al., 2015; Schutte et al., 2014; Y. Wang et al., 2010; Zhi, Zhao, Zhou, Liu, & Chen, 2012), was shown to be regulated by α-arrestin in both *Drosophila* (Y. Kwon et al., 2013) and human cells (J. Xiao et al., 2018). In *Drosophila*, the protein Leash was identified as an α-arrestin and shown to down-regulate Yki by promoting its lysosomal degradation, leading to a restriction in growth (Y. Kwon et al., 2013). In human cells, ARRDC1 and ARRDC3 were shown to induce degradation of the mammalian homolog of Yki, YAP1, by recruiting the E3 ubiquitin ligase ITCH in renal cell carcinoma (J. Xiao et al., 2018), suggesting functional homology between human and *Drosophila*. However, because the α-arrestins interact with multiple targets, an unbiased, comparative analysis of interactome is required to determine whether other α-arrestin from human and *Drosophila* have common and specific interacting partners, which will determine their functional homology and diversification.

A comprehensive understanding of their protein-protein interactions (PPIs) and interactomes will shed light on the underlying molecular mechanisms, reveal novel regulatory axes, and enable the identification of previously unrecognized roles of α-arrestin in cellular processes. Furthermore, extensive characterization of the α-arrestin interactome may help uncover potential therapeutic targets and provide valuable insights into the treatment of diseases associated with dysregulated signaling pathways (Diaz et al., 2005; Lu et al., 2008; Q. Y. Wang et al., 2018; Zhou et al., 2010).

In this study, we conducted affinity purification/mass spectrometry (AP/MS) of six human and twelve *Drosophila* α-arrestins. A high-confidence PPI network was constructed by selecting a cut-off for receiver operating characteristic (ROC) curves of Significance Analysis of INTeractome express (SAINTexpress) scores (Teo et al., 2014). The constructed interactomes were validated using known affinities between domains of prey proteins and the short linear motifs of α-arrestins. We also investigated orthologous relationships between binding partners from human and *Drosophila* and found that many proteins with both known and novel functions could be conserved between two species. Finally, we performed experiments to provide new insights into the functions of TXNIP and ARRDC5 that were revealed in our study. Together, our results provide a valuable resource that describes the PPI network for α-arrestins in both human and *Drosophila* and suggest novel regulatory axes of α-arrestins.

I-３ Results

I-３.１ AP/MS of α-arrestins from human and Drosophila and identification of high-confidence PPIs

Genome-scale sets of prey proteins interacting with α-arrestins (referred to herein as 'interactomes') were compiled by conducting AP/MS for six human and twelve *Drosophila* α-arrestin proteins (Figure I-1A). Proteins possibly interacting with α-arrestins were pulled down from total cell lysates of human embryonic kidney 293 (HEK293) and S2R+ cells stably expressing GFP-tagged α-arrestins (Figure I-1B; Figure I-2). All α-arrestin experiments were replicated twice, and negative control experiments were conducted multiple times. In total, 3,243 and 2,889 prey proteins involved in 9,908 and 13,073 PPIs with human and *Drosophila* α-arrestins, respectively, were initially detected through AP/MS (Figure I-1B).

To build high-confidence interactomes of α-arrestin family proteins, a probabilistic score for individual PPIs was estimated using SAINTexpress (Teo et al., 2014) and an optimal cutoff for the scores was set using positive and negative PPIs of α-arrestin from public databases and the literature (Colland et al., 2004; Dotimas et al., 2016; Draheim et al., 2010; Mellacheruvu et al., 2013; Nabhan et al., 2012; Nishinaka et al., 2004; Puca & Brou, 2014; Szklarczyk et al., 2015; Warde-Farley et al., 2010; Wu et al., 2013). The resulting ROC curves showed high area under curve (AUC) values and the SAINTexpress scores at which the false discovery rate (FDR) was 0.01 were selected as cutoffs (0.85 for human and 0.88 for *Drosophila,* Figure I-3A). Given the cutoffs, 1,306 and 1,732 PPIs involving 902 and 1,732 proteins were selected for human and *Drosophila*, respectively. Because proteins of low abundance (low spectral counts) are easily affected by a stochastic process (Lundgren, Hwang, Wu, & Han, 2010; Old et al., 2005), the minimum spectral count of PPIs was set at 6, allowing us to select PPIs with higher confidence. In fact, the spectral counts of the filtered PPIs were highly reproducible between replicates (Figure I-3B; Pearson's correlations, 0.91 for human; 0.89 for *Drosophila*). Principal component analysis (PCA) based on $\log_2$ spectral counts also confirmed a high reproducibility between replicates (Figure I-4). Moreover, we successfully detected many known interaction partners of α-arrestins such as NEDD4, WWP2, WWP1, ITCH and TSG101, which have been previously reported in the literature and PPI databases (Figure I-5) (Colland et al., 2004; Dotimas et al., 2016; Draheim et al., 2010; Mellacheruvu et al., 2013; Nabhan et al., 2012; Nishinaka et al., 2004; Puca & Brou, 2014; Szklarczyk et al., 2015; Warde-Farley et al., 2010; Wu et al., 2013). Finally, our filtered interactomes of human and *Drosophila* α-arrestins, comprised of 390 PPIs between six α-arrestins and 307 prey proteins in human, and 740 PPIs between twelve α-arrestins and 467 prey proteins in *Drosophila*, are hereafter referred to as 'high-confidence PPIs'.

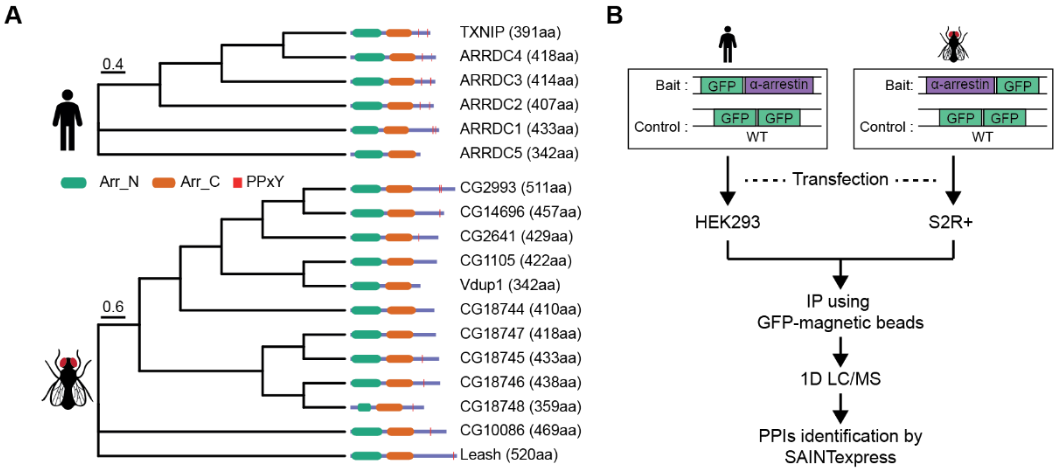


Figure I-1. AP/MS of α-arrestins from human and Drosophila

A) Phylogenetic tree of α-arrestins from human (6, top) and *Drosophila* (12, bottom) based on protein sequences. The numbers in parentheses indicate the length of each protein. aa, amino acids; Arr_N: Arrestin N domain; Arr_C: Arrestin C domain; PPxY: PPxY motif. **(B)** Shown is a schematic flow of AP/MS experiments and computational analysis.

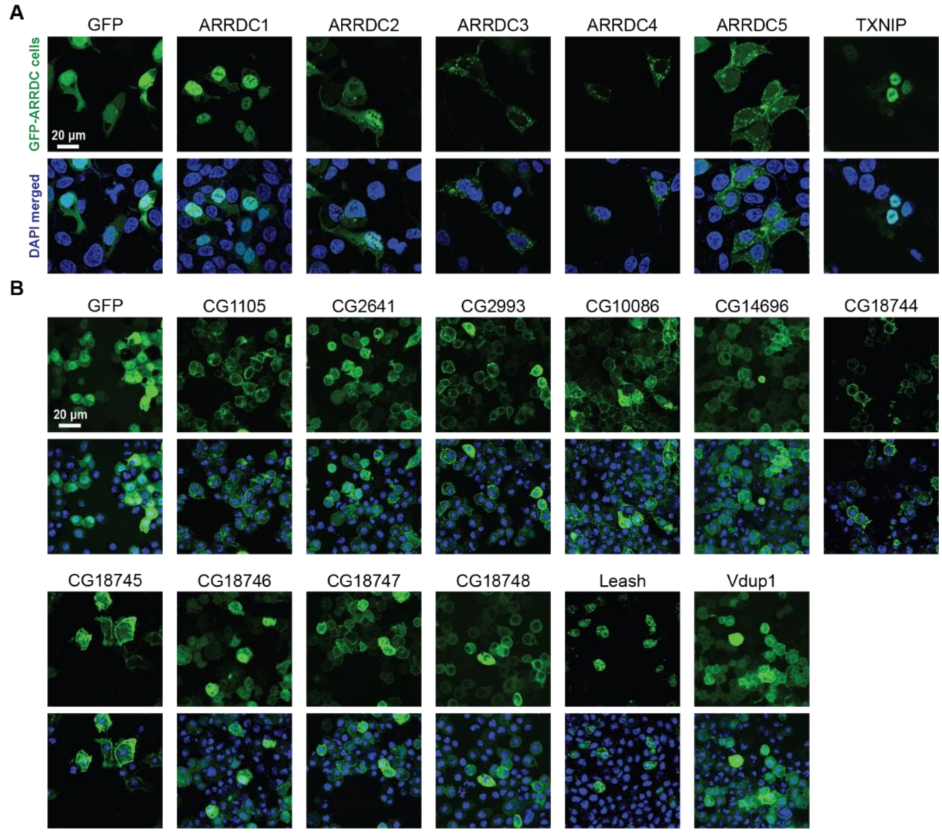


Figure I-3. Fluorescence images showing HEK293 and S2R+ cells stably expressing GFP-tagged α-arrestins

Representative images of HEK293 (A) and S2R+ (B) cells stably expressing GFP-tagged α-arrestins.

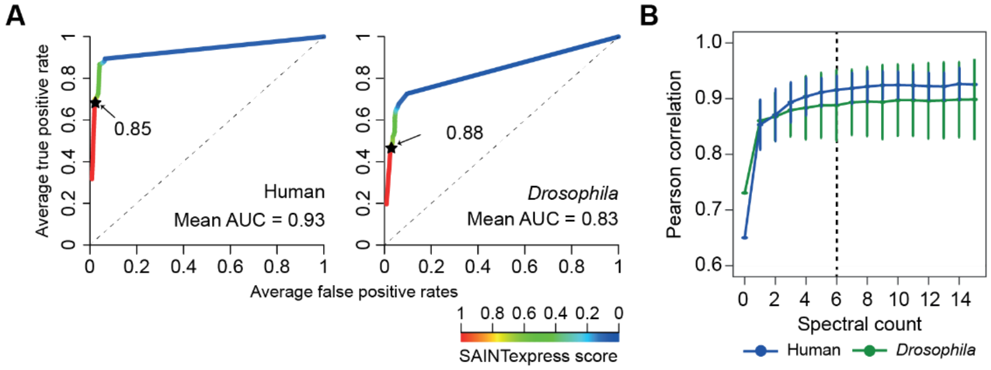


Figure I-2. Filtering of AP/MS data to generate high-confidence PPI sets of α-arrestins from human and *Drosophila*

(**A**) ROC curves of SAINTexpress scores along with AUC values. The arrows point to the cutoff scores used in subsequent studies in human (left) and *Drosophila* (right). **(B)** Average Pearson correlation coefficients of $\log_2$ spectral counts between replicates of AP/MS of each α-arrestin at varying cutoffs are shown (mean ± standard deviation(sd)). The cutoff used in this study, 6, is shown as a dashed line.

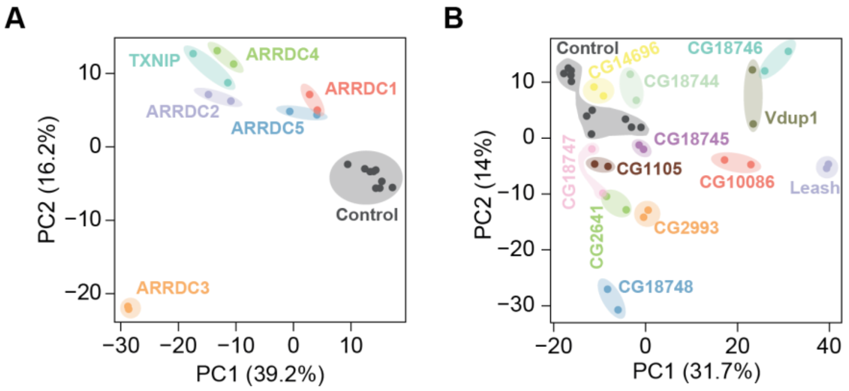


Figure I-4. Spectral count profile of filtered AP/MS data are highly reproducible
PCA plots based on log₂ spectral counts of high-confidence PPIs for human **(A)** and *Drosophila* **(B)** are shown.

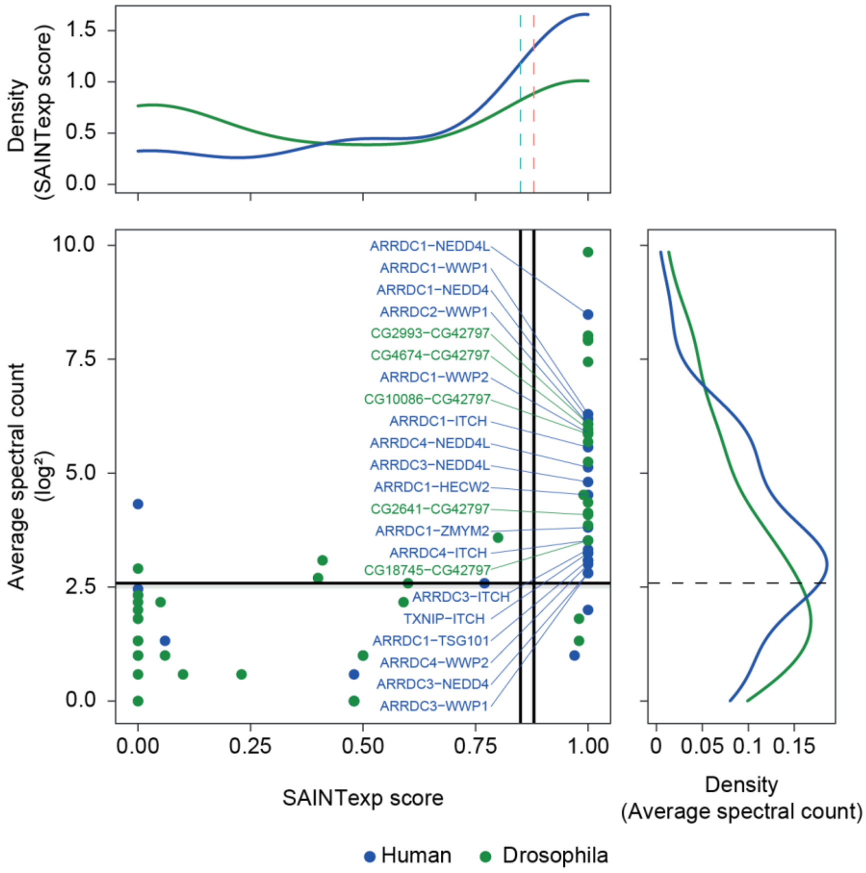


Figure I-5. Substantial part of previously reported PPIs involving α-arrestins are detected by our AP/MS approach

SAINTexpress scores and average spectral counts ($\log_2$) of the positive are shown and density plots for each axis are also plotted. The positive PPIs that are included in the filtered set are selectively labeled.

I-３.２ Short-linear motifs and protein domains enriched in α-arrestins and their interactomes

To validify our high-confidence PPIs, we sought to analyze known short-linear motifs in α-arrestins, which are commonly 3-15 stretches of amino acids that are known to participate in interactions with other protein domains (Dinkel et al., 2015). Utilizing the known affinities between short linear motifs in α-arrestins and protein domains in interactomes from eukaryotic linear motif (ELM) database (Dinkel et al., 2015), we evaluated whether our high-confidence PPIs could be explained by the known affinities between them. The fractions of our high-confidence PPIs (green, Figure I-6A), supported by the known affinities were significantly greater than those of all raw PPIs (red, Figure I-6A top) in both species (P < 9.37 X 10^{-11} for human and P < 0.0012 for *Drosophila*, one-sided Fisher's exact test, Figure I-6 top). One of the most well-known short-linear motifs in α-arrestin is PPxY, which is reported to bind with high affinity to the WW domain found in various proteins, including ubiquitin ligases (Macias et al., 1996). Our analysis revealed the specific enrichment of WW domain-containing proteins in the interactomes of α-arrestins with at least one PPxY motif but not in that of the human α-arrestin (ARRDC5) without a PPxY motif (Figure I-6A, bottom-left). The interactomes of five out of the eight *Drosophila* α-arrestins with a PPxY motif were enriched for WW domain-containing proteins, but there was no such enrichment for any of the *Drosophila* α-arrestins without a PPxY motif (Figure I-6A, bottom-right). In conclusion, a considerable portion of the high-confidence PPIs identified in this study can be evident by known affinities between short-linear motifs and protein domains.

Next, we conducted enrichment analyses of proteins domains among interactome of each arrestin to investigate known and novel protein domains commonly or specifically interacting with α-arrestins (Figure I-6B). The most prominent interacting domains in both species were the Homologous to E6AP C-terminus (HECT), WW, and C2 domains (Figure

I-6B). HECT and C3 domains are well known to be embedded in the E3 ubiquitin ligases such as NEDD4, HECW2, and ITCH along with WW domains (Weber, Polo, & Maspero, 2019) and as we observed strong preference of WW domains to PPxY containing proteins (Figure I-6A), these domains were significantly enriched in binding proteins of α-arrestins with PPxY motif in human and *Drosophila* (FDR < 0.033 ~ 1.23 X 10^{-11} for human; FDR < 0.045 ~ 4.10 X 10^{-6} for *Drosophila*, Figure I-6B). Other common protein domains involved in the protein degradation process, such as proteasome domains, were also significantly enriched in the interactomes (of ARRDC4 in human and Leash in *Drosophila*) in both species (FDR < 6.41 X 10^{-4} for human and FDR < 1.30 X 10^{-5} for *Drosophila,* Figure I-6B). Interestingly, some α-arrestins (ARRDC3 in human and Vdup1, Leash, and CG18746 in *Drosophila*) appeared to interact in common with RNA binding domains, such as DEAD box, helicase, WD40, and RNA recognition motif, but others did not. In addition, the cargo and motor protein domains IBN_N (FDR < 0.0076 for human and FDR < 2.50 X 10^{-4} ~ 2.11 X 10^{-6} for *Drosophila*) and myosin_head (FDR < 0.033 for human and FDR < 2.11 X 10^{-6} for *Drosophila*) also interacted with several α-arrestins in common (ARRDC4 in human and CG1105, CG18745, and CG18748 in *Drosophila,* Figure I-6B). These enriched domains explain the conserved interactomes associated with RNA splicing and protein transport in both species. In addition, human α-arrestins seem to interact with human-specific domains, such as PDZ, Rho-GEF, MCM, laminin, zinc finger, and BAG6 domains, providing an expanded interactome of human α-arrestins (Figure I-6B, domains in black), indicating the presence of both conserved and specific protein domains interacting with α-arrestins.

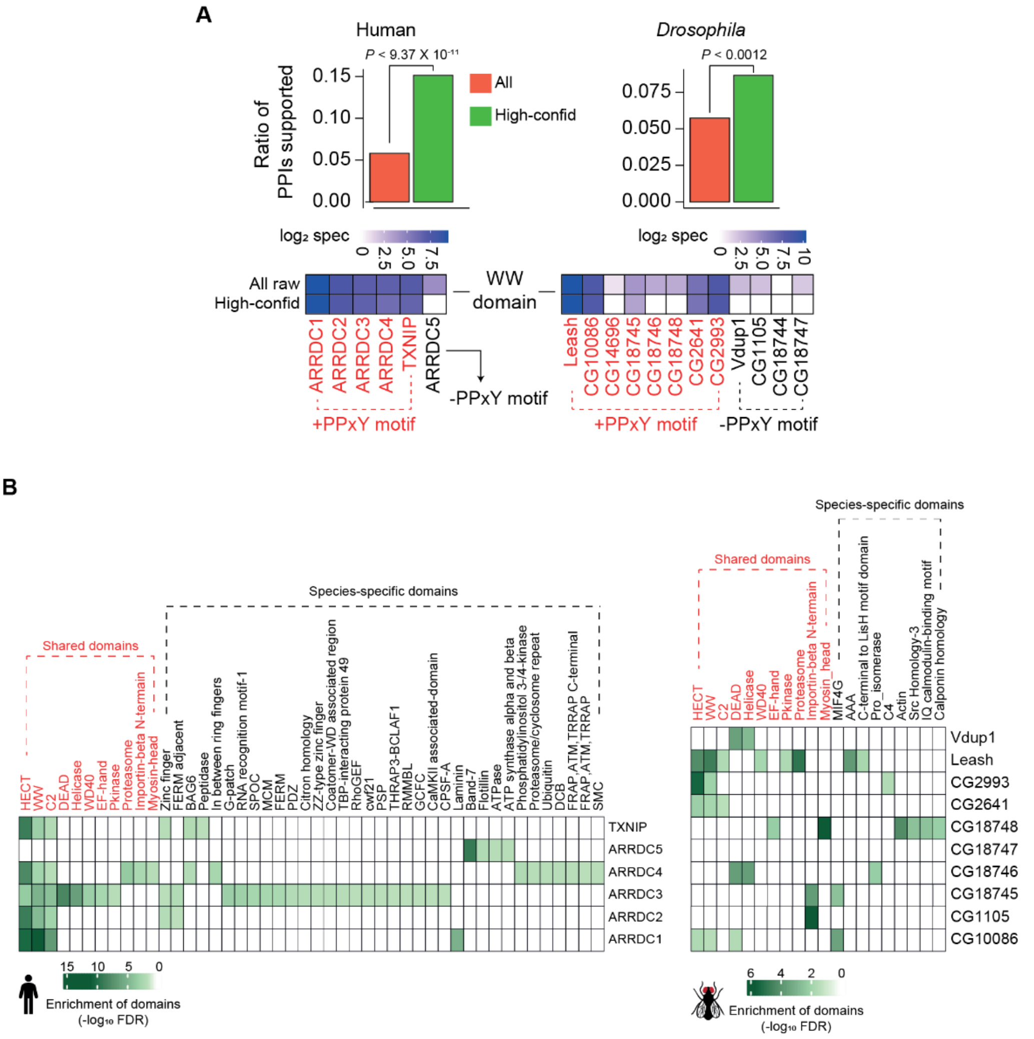


Figure I-6. Extensive landscape of protein domains associated with PPIs of α-arrestins

(A) (Top) The fraction of high-confidence and all raw (unfiltered) PPIs that are supported by known affinities between short linear motifs and protein domains in human (left) and *Drosophila* (right). One-sided, Fisher's exact test was performed to test the significance. (Bottom) The sum of $\log_2$ spectral counts ($\log_2$ spec) of interacting proteins with WW domains observed in the high confidence and all raw PPIs are visualized in the heatmap. **(B)** Protein domains enriched in each α-arrestin interactome for human (top) and *Drosophila* (bottom) are shown. The significance of the enrichment test ($-\log_{10}$ FDR) is indicated in shades of green, as depicted in the legend. SPOC, spen paralogue and orthologue C-terminal; MCM, minichromosome maintenance protein complex; FDRM, F for 4.1 protein, E for ezrin, R for radixin and M for moesin; TBP, TATA binding protein; GEF, guanine nucleotide exchange factor; THRAP3, thyroid hormone receptor-associated protein 3; BCLAF1, Bcl-2-associated transcription factor1; RMMBL, RNA metabolising metallo beta lactamase; CaMKII, C-terminus of the Calcium/calmodulin dependent protein kinases II; CPSF, cleavage and polyadenylation specificity factor; DCB, dimerization and cyclophilin-binding domain; FRAP, FKBP12-rapamycin complex-associated protein; ATM, ataxia telangiectasia mutant; THRAP, transformation/transcription domain associated proteins; MIF4G, middle domain of eukaryotic initiation factor 4G; AAA, ATPase family associated with various cellular activities; C4, C-terminal tandem repeated domain in type 4 procollagen; SMC, structural maintenance of chromosomes.

I-３.３ Expanded functional signatures of α-arrestin interactomes

Because the functions of α-arrestins can be inferred based on their binding partners, the prey proteins were grouped based on their interactions with α-arrestins, which revealed specialized functions of the respective α-arrestins with some redundancy as well as both known and novel functions (Figure I-7). The analysis of protein class enrichment by the PANTHER classification system (Thomas et al., 2003) revealed previously reported functions, such as 'Ubiquitin ligase' (FDR < 0.0019 and 5.01 X 10^{-7} for human; Benjamini-Hochberg correction) and 'Protease' (FDR < 1.93 X 10^{-6} for human and 5.02 X 10^{-6} for *Drosophila*) (Dores et al., 2015; Y. Kwon et al., 2013; Nabhan et al., 2012; Puca, Chastagner, Meas-Yedid, Israel, & Brou, 2013; Rauch & Martin-Serrano, 2011; Shea et al., 2012; J. Xiao et al., 2018). In fact, the known binding partners, NEDD4, WWP2, WWP1, and ITCH in human and CG42797, Su(dx), Nedd4, Yki, Smurf, and HERC2 in *Drosophila*, are related to ubiquitin ligases and protein degradation (Figure I-5; Figure I-7). In addition, novel biological functions of α-arrestins were uncovered. For instance, in human, prey proteins interacting with ARRDC3 displayed enrichment of 'RNA splicing factor and helicase' functions as well as 'GTPase-activating proteins', those of ARRDC4 were enriched with 'Apolipoprotein', and those of ARRDC5 with 'ATP synthase' (Figure I-7, up). Motor protein, protease, ubiquitin ligase, RNA splicing factor, and helicase were functions that were also enriched in *Drosophila* prey proteins (Figure I-7, bottom). Among them, the motor protein and RNA splicing, and helicase functions seemed to be novel conserved functions between human and *Drosophila*. The functional compositions of the interacting proteins summarized the common or highly specialized functions of α-arrestins well (Figure I-7, right panel). For example, in human, proteins that interacted with TXNIP, ARRDC2, and ARRDC4 showed similar ubiquitination and protease-related functions, whereas ARRDC3 and ARRDC5 displayed unique interactomes associated with other functions. For *Drosophila*, the interactomes of the [Vdup1, CG10086 and CG18744], [CG18748 and

CG18747], and [CG1105 and CG14696] α-arrestin subsets each exhibited similar functional compositions, but the Leash interactome showed a distinct enrichment of ubiquitination-related and protease functions. Taken together, these results suggest that the resulting high-confidence PPIs of α-arrestins expanded the functional interactome maps of α-arrestins in both human and *Drosophila*.

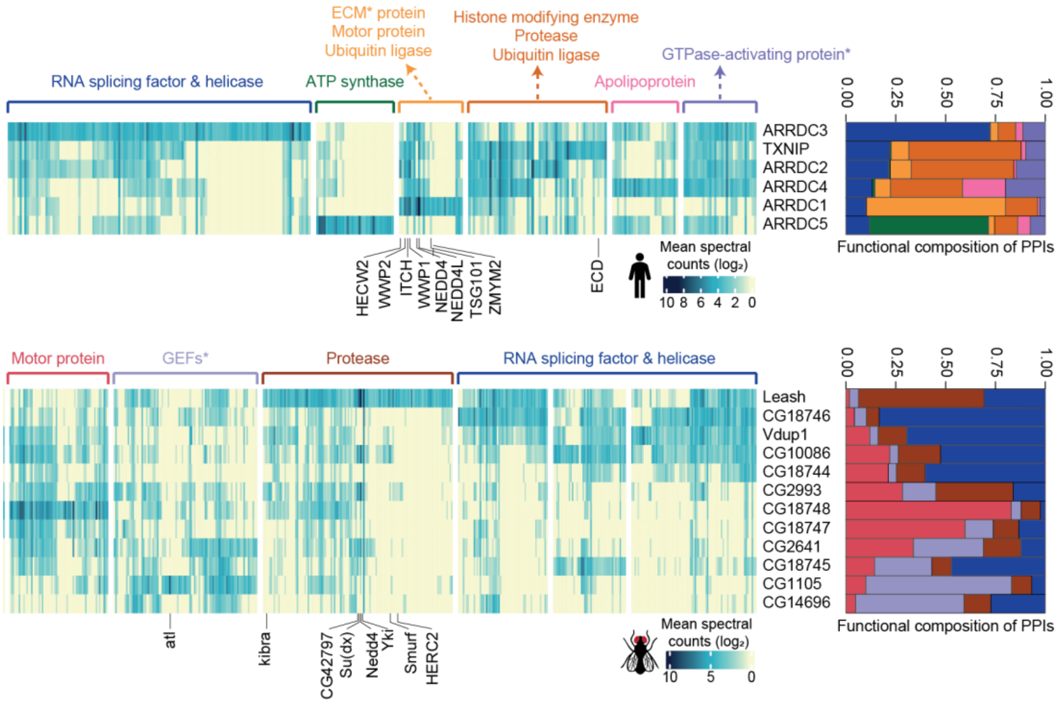


Figure I-7. Expanded functional signatures associated with PPIs of α-arrestins in both human and *Drosophila*

The α-arrestins and interacting prey proteins were hierarchically clustered based on the $\log_2$ mean spectral counts and summarized for human (top) and *Drosophila* (bottom) in the heatmaps. The functionally enriched protein groups of preys are indicated at the top. Previously reported proteins interacting with α-arrestins are labeled at the bottom. On the right, the functional composition of prey groups is summarized with the sum of $\log_2$ mean spectral counts of each prey group, which are colored to correspond with the labels on the left.

I-３.４ Subcellular localizations of α-arrestin interactomes

Cellular localizations of proteins often provide valuable information of their functions and activity, but only a small number of α-arrestins are known for their preferential subcellular localization. We thus examined the subcellular localizations of the interacting proteins using the cellular component feature in Gene Ontology (GO) using DAVID (Huang da, Sherman, & Lempicki, 2009a, 2009b) (Figure I-8). Prey proteins (246 for human and 245 for *Drosophila*) that were localized in at least one cellular compartment were examined. We found that prey proteins of ARRDC5 were preferentially localized in the endoplasmic reticulum and at the plasma membrane (PM) but were less often localized in the nucleus, compared to those of other human α-arrestins (Figure I-8, left). Similarly, the prey proteins of ARRDC1 and 4 were less often localized in the nucleus, instead being preferentially localized in the cytoplasm (ARRDC4) or extracellular space (ARRDC1), in agreement with previous reports (Nabhan et al., 2012; Q. Y. Wang et al., 2018). TXNIP seemed to preferentially interact with prey proteins in cytoplasm and nucleus (Figure I-8, left), consistent with a previous report (S. K. Kim, Choe, & Park, 2019; Saxena, Chen, & Shalev, 2010). ARRDC3, which was suggested to be localized in cytoplasm in previous study (Nabhan et al., 2010), appeared to interact with proteins preferentially localized in nucleus in addition to the ones in cytoplasm, implying novel functions of ARRDC3 in the nucleus. In *Drosophila*, the localization of interacting proteins is often uncharacterized compared to human, but a preference for a localization for part of the interactomes can be observed (Figure I-8, right). Some of them are preferentially localized at the PM (CG18747), mitochondria (CG14696), peroxisome (CG14696), lysosome (CG2641), or cytoskeleton (CG18748), compared to others. However, interactomes of Leash, Vdup1, CG2641, CG18745, CG18746, and CG10086 are preferentially localized in the nucleus. Taken together, these data about the preferential localizations of interacting proteins provide evidence about the functions and activity of α-arrestins in cells.

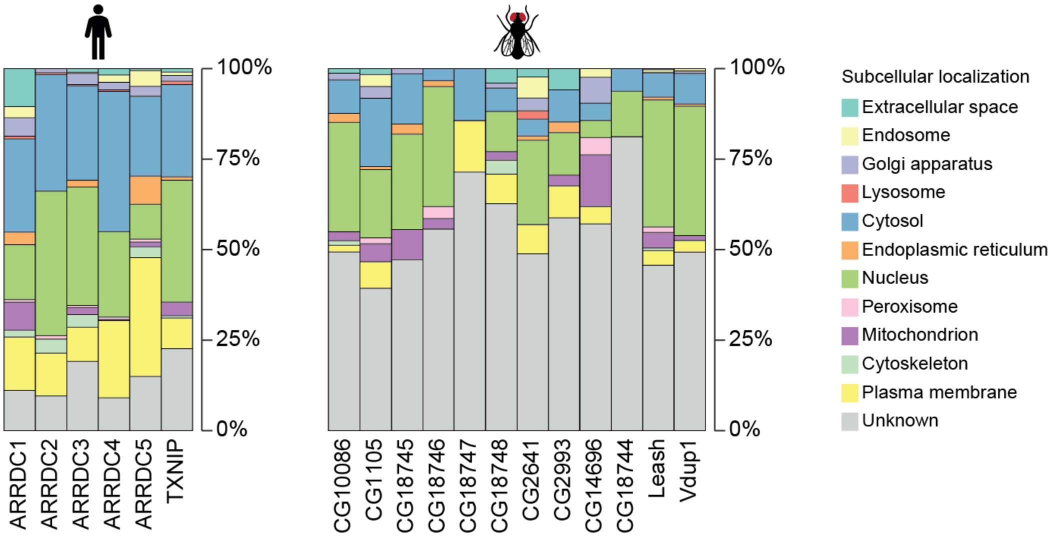


Figure I-8. Subcellular localizations of α-arrestins interactomes

Subcellular localizations of prey proteins of each α-arrestin for human (left) and *Drosophila* (right).

I-３.５ Functional complexes in α-arrestin interactomes

The fact that protein functions are often realized in complexes (Hartwell, Hopfield, Leibler, & Murray, 1999) urged us to search for functional complexes that extensively interact with α-arrestins. For this analysis, protein complexes that are significantly connected with each α-arrestin were examined using the COMPlex Enrichment Analysis Tool (COMPLEAT) (Vinayagam et al., 2013), resulting in the detection of 99 and 18 protein complexes for human and *Drosophila*, respectively. The complexes were iteratively combined with cellular components from GO based on the overlap coefficients (Vijaymeena & Kavitha, 2016). The significance of the resulting combined complexes was then tested with the connectivity to each α-arrestin using the interquartile means (IQMs) of SAINTexpress scores compared to those from 1000 random cohorts ($P < 0.05$). This approach showed that 44 clustered complexes comprising 324 protein subunits were significantly interacting with six human α-arrestins (Figure I-9) and 21 clustered complexes comprising 192 subunits were significantly interacting with *Drosophila* α-arrestins (Figure I-10).

The two largest complexes found to interact with α-arrestins were related to protein degradation (proteasome and ubiquitin-dependent proteolysis) and RNA splicing and processing in both species (Figure I-9; Figure I-10). ARRDC1, 2, and 4 and TXNIP in human and Leash and CG2993 in *Drosophila* were found to interact with protein degradation complexes. CG2993 and CG18747 appeared to bind to a putative complex comprising NEDD4 family interacting protein 2, recently reported to be a mediator of ubiquitin ligase (Trimpert et al., 2017). On the other hand, ARRDC2, 3, and 4 in human and Leash, CG18746, Vdup1, CG10086, and CG18744 in *Drosophila* were found to interact with RNA splicing and processing complexes. Although the above-mentioned α-arrestins interacted in common with the two complexes described above, they were also found to bind to distinct complexes. For instance, TXNIP specifically binds to transcriptional and histone deacetylase (HDAC) complexes, ARRDC1 to axon guidance, endosomal sorting, and laminin complexes, ARRDC2 to the Set1C/COMPASS complex, ARRDC3 to transcription elongation factors and spindle assembly checkpoint and cell polarity complexes, and ARRDC4 to clathrin-coated pit and BAT3 complexes in human. In *Drosophila*, Leash specifically binds to AP-2 adaptor and WASH complexes and CG18746 to the UTP B complex. In addition to the two largest complexes and their associated α-arrestins, ARRDC5 in human and CG2641, CG1105, CG14696, and CG18745 in *Drosophila* interact in common with protein transport and localization complexes. ARRDC5 is specifically associated with V-type ATPase and vacuolar protein sorting complexes in human. CG18748 and CG18747 are associated with motor protein complexes including actin, myosin, and microtubule-associated complexes in *Drosophila*. Taken together, the results from this analysis provide a glimpse of underexplored roles for α-arrestins in diverse cellular processes.

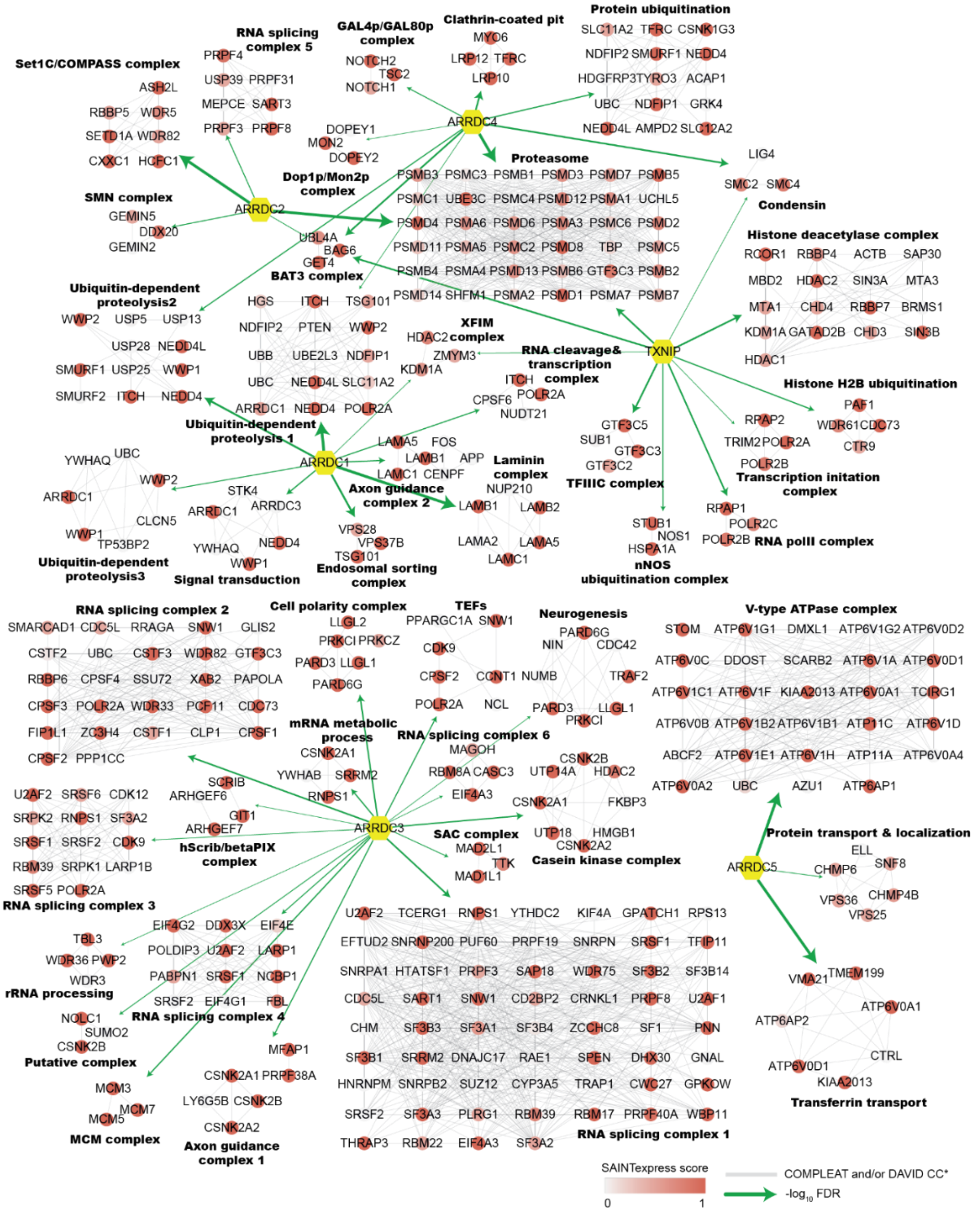


Figure I-9. Network of α-arrestins and their interacting protein complexes in human

Network of α-arrestins and the functional protein complexes that significantly interact with them in human. α-arrestins are colored yellow and prey proteins in protein complexes are colored according to the SAINTexpress scores of the PPIs. The gray edges indicate that evidence supporting the complex was provided by COMPLEAT and/or GO cellular components. The thickness of the green arrows indicates the strength of the interaction between α-arrestins and the indicated protein complexes, which was estimated with -log$_{10}$ FDR of complex association scores. COMPASS, complex proteins associated with Set1; SMN, survivor of motor neurons; TFIIIC, transcription factor III C; RNA polII, RNA polymerase II; MCM, minichromosome maintenance protein complex; SAC, spindle assembly checkpoint.

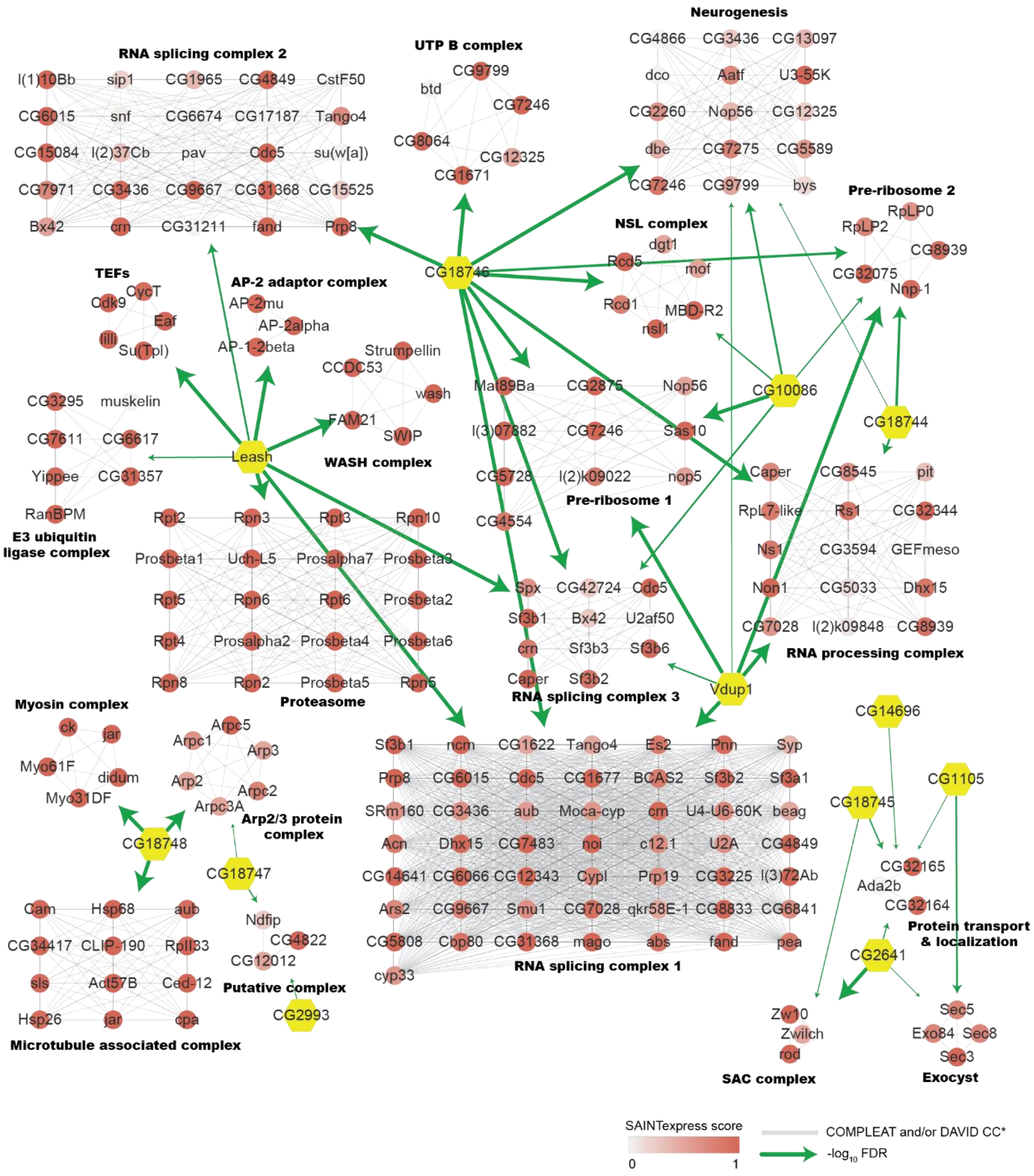


Figure I-10. Network of α-arrestins and their interacting protein complexes in *Drosophila*

Network of α-arrestins and the functional protein complexes that significantly interact with them in *Drosophila*, plotted as in Figure 9. NSL, non-specific lethal; WASH, Wiskott-Aldrich syndrome protein and scar homolog; Arp2/3, actin related protein 2/3. TEF, transcription elongation factor.

I-３.６ Conserved interactomes of α-arrestins

Given that α-arrestins are widely conserved in metazoans (DeWire, Ahn, Lefkowitz, & Shenoy, 2007), we sought to exploit the evolutionally conserved interactomes of human and *Drosophila* α-arrestins. For this analysis, we searched for orthologous relationships in the α-arrestin interactomes using the DRSC Integrative Ortholog Prediction Tool (DIOPT) (Hu et al., 2011). Among high-confidence prey proteins, 68 in human and 64 in *Drosophila* were reciprocally predicted to have ortholog relationships, defining 58 orthologous prey groups (DIOPT score ≥ 2). α-arrestins were then hierarchically clustered based on the $\log_2$-transforemd mean spectral counts of these orthologous interactome, defining seven groups of α-arrestins. Orthologous prey proteins were grouped according to their shared biological function, defining nine functional groups and others of diverse functions (Figure I-11). The resulting clusters revealed PPIs that were functionally conserved. For instance, ARRDC3 in human and CG18746 in *Drosophila* actively interact with proteins in RNA binding and splicing groups. Leash in *Drosophila* appeared to interact with proteins in similar functional groups as ARRDC3 but, like ARRDC1, it also extensively interacts with members of ubiquitin-dependent proteolysis groups. In addition, ARRDC4 interacts with proteins in the motor protein and trafficking group, similar to CG18748 in *Drosophila*, and binds to proteins in the ubiquitin-dependent proteolysis group, similar to TXNIP. Similarly, CG10086 and Vdup1, CG14696 and ARRDC5, and CG2993 and ARRDC2 appeared to have conserved interactomes between human and *Drosophila.*

The most prominent functional modules shared across both species were the ubiquitin-dependent proteolysis, endosomal trafficking, and small GTPase binding modules, which are in agreement with the well-described functions of α-arrestins in membrane receptor degradation through ubiquitination and vesicle trafficking (Dores et al., 2015; Nabhan et al., 2012; Puca et al., 2013; J. Xiao et al., 2018) (Figure I-11). In contrast, the functional modules involving cyclin and cyclin-dependent kinase, casein kinase complex, and laminin seemed to be conserved between relatively specific sets of α-arrestins, whereas those related to motor proteins and RNA binding and splicing were more generally conserved. Taken together, the comparative analyses led us to identification of detailed, orthologous interactome maps of α-arrestins, which extend beyond the limited insights provided by sequence-based comparative analysis alone (Figure I-12). Conserved roles of α-arrestins in both established and previously uncharacterized signaling pathways expand our understanding of the diverse roles of α-arrestins in cellular signaling.

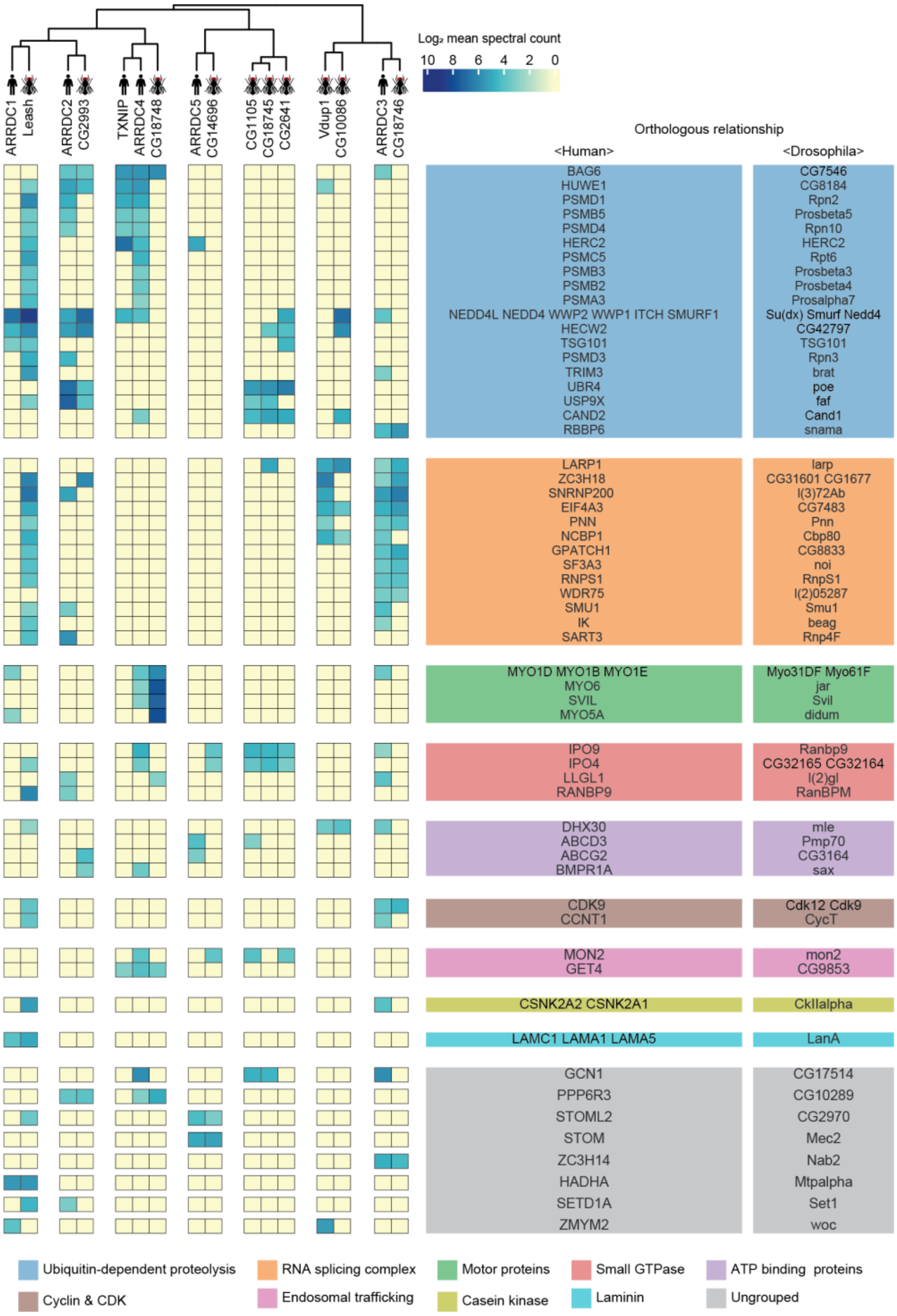


Figure I-11. A substantial fraction of α-arrestin-PPIs are conserved across species

Human and *Drosophila* α-arrestins are hierarchically clustered based on $\log_2$-transformed mean spectral counts of their orthologous interactome. They are then manually grouped according to shared biological functions and assigned distinct colors. The names of orthologous proteins that interact with α-arrestins are displayed on the right side of the heatmap.

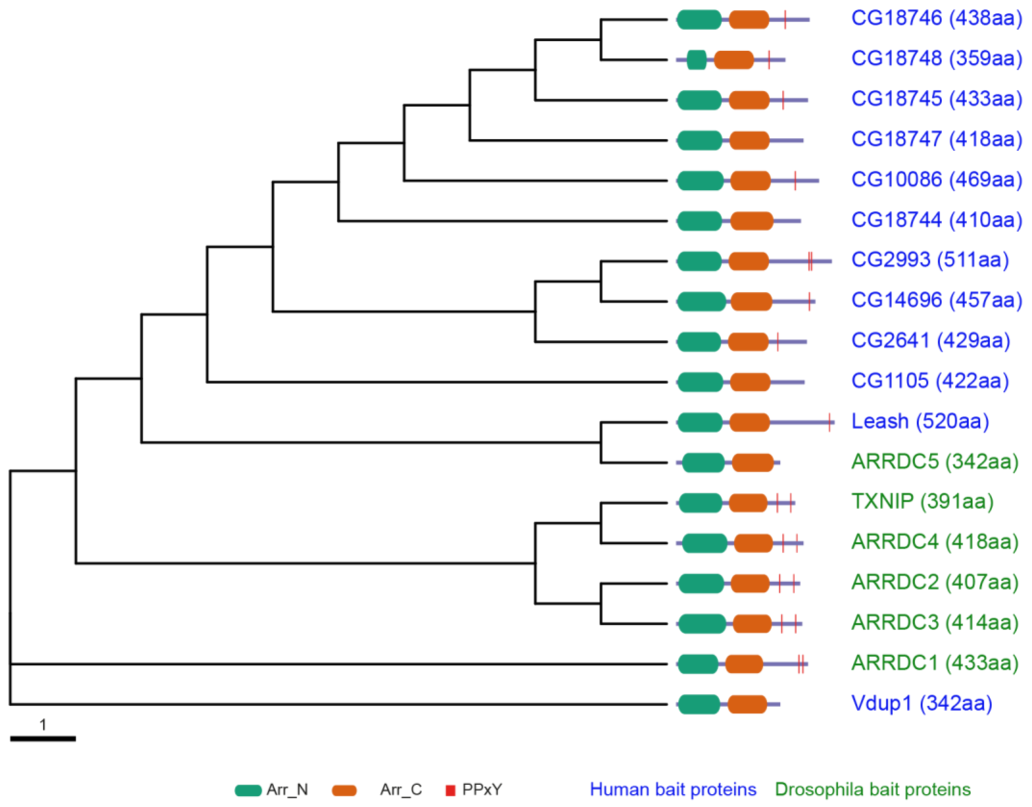


Figure I-12. Phylogenetic tree showing relationships between α-arrestins from human and *Drosophila*

I-３.７ Accessible chromatin regions and gene expression profiling upon TXNIP depletion in HeLa cells

TXNIP is one of the most well-studied α-arrestins. Previous studies reported that TXNIP interacts with transcriptional repressors, such as FAZF, PLZF, and HDAC1 or HDAC3, to exert antitumor activity (S. H. Han et al., 2003) or repress NF-kB activation (H. J. Kwon et al., 2010). However, although such studies provided information about interactions with a few transcriptional repressors, they barely provided a systematic view of the roles of TXNIP in controlling the chromatin landscape and gene expression. In that sense, our PPI analysis first revealed that TXNIP extensively binds to chromatin remodeling complexes, such as the HDAC and histone H2B ubiquitination complexes, as well as to transcriptional complexes, such as the RNA polymerase II and transcription factor IIIc complexes (Figure I-9). Such PPIs indicate that TXNIP could control transcriptional and epigenetic regulators. To examine how the global epigenetic landscape is remodeled by TXNIP, we knocked down its expression in HeLa cells with a small interfering RNA (siTXNIP) and confirmed a decrease at both the RNA and protein levels (Figure I-13A and B). We then produced two biological replicates of ATAC- and RNA-seq experiments in HeLa cells with TXNIP depletion (Table I-1) to detect differentially accessible chromatin regions (dACRs) and differentially expressed genes (DEGs) (Figure I-14A). The replicated samples were well grouped by the siTXNIP condition in principle component spaces (Figure I-14B and C). The normalized ATAC-seq signal and the RNA level of expressed genes clearly showed the enrichment of open chromatin signals around the transcription start sites (TSSs) of genes that are actively transcribed (Figure I-15).

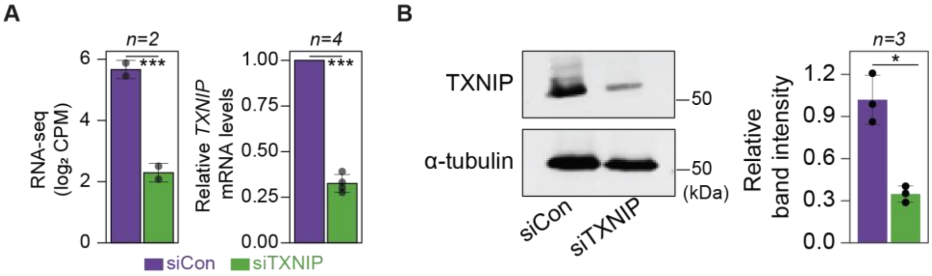


Figure I-14. Confirmation of TXNIP knockdown in HeLa cells in both RNA and protein levels

(A-B) HeLa cells were treated with either siRNA against TXNIP (siTXNIP) or negative control (siCon) for 48 hours (hr) and analyzed of changes in the mRNA **(A)** and protein levels **(B)** of TXNIP. **(A)** Expression levels of RNAs were quantified by RNA-seq (left, log2 counts per million mapped reads (CPM), see "Processing of RNA-seq data" in "Materials and Methods") and RT-qPCR (right, relative levels of TXNIP in siTXNIP compared to siCon condition, see "Quantitative Reverse-transcription PCR" in Supplementary Information). **(B)** Protein levels were first visualized by western blot analysis of lysates from HeLa Cells and band intensities of three independent experiments were quantified (right). **(A-B)** Gray dots depict actual values of each experiment and bar plots indicate mean ± standard deviation (sd). ***FDR < 0.001 (test of differential expression by edgeR (Robinson, McCarthy, & Smyth, 2010), see "Processing of RNA-seq data" in "Materials and Methods") for RNA-seq. **P* < 0.05, *** *P* < 0.001 (two-sided paired Student T test) for RT-qPCR and western blots.

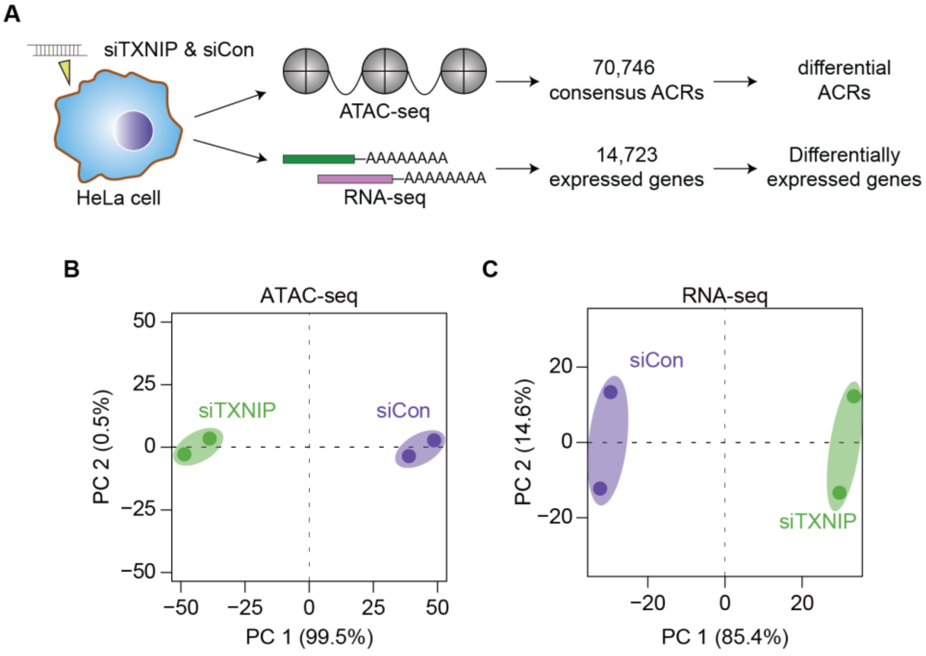


Figure I-13. ATAC- and RNA-seq data of HeLa cells are clearly distinguished between WT and TXNIP depleted conditions

(A) A schematic workflow for detecting dACRs and DEGs using ATAC- and RNA-seq analyses, respectively. **(B and C)** PCA plots of ATAC- **(B)** and RNA-seq **(C)** results based on batch-corrected $\log_2$ counts and CPM, respectively. Numbers in parentheses are percentages of explained variance for the corresponding PCs.

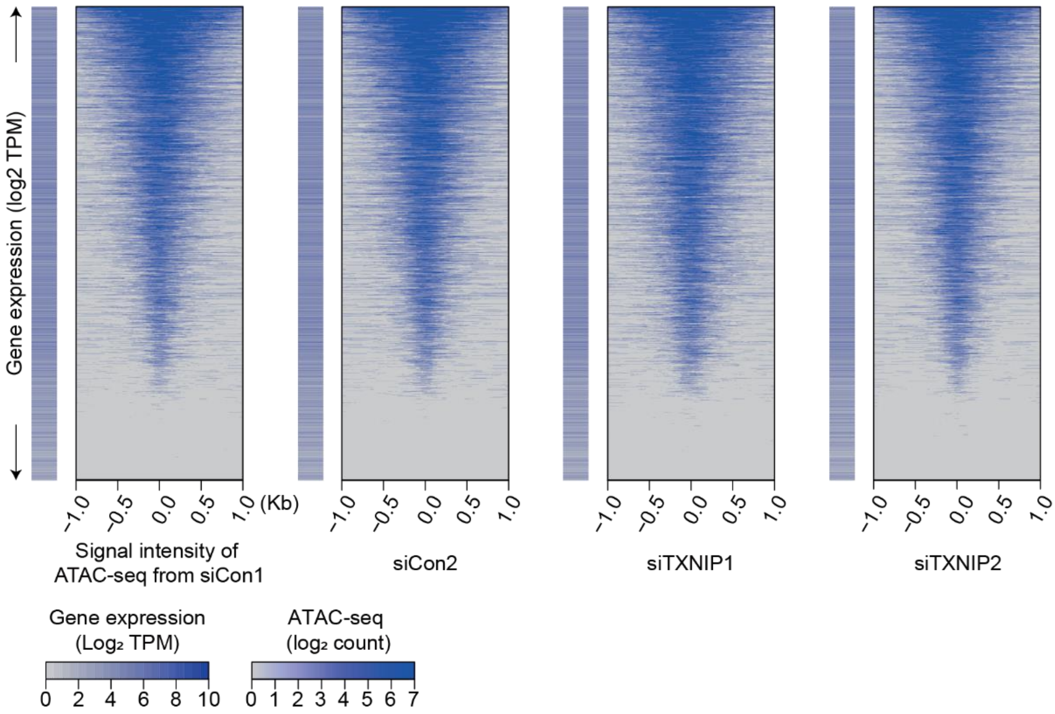


Figure I-15. Open chromatin regions are enriched in promoters of actively transcribed genes Heatmaps of ATAC-seq read counts (read counts have been transformed into a $\log_2$ function and corrected for batch effects) in regions surrounding TSSs along with $\log_2$ (RNA level in siTXNIP-treated cells/RNA level in siCon-treated cells) for genes having the corresponding TSS are plotted

	ATAC-seq			RNA-seq	
Sample	Properly paired reads (%)	Filtered/dedup reads*	Called peaks	Filtered reads (%)	Alignable reads
siNegative 1	117,217,586 (97.8%)	17,929,628	74,373	41,756,384 (99.8%)	37,548,784
siNegative 2	203,055,772 (97.7%)	31,497,080	141,799	41,900,786 (99.4%)	36,139,515
siTXNIP1	123,045,656 (97.8%)	20,050,776	69,431	41,729,984 (99,8%)	37,185,131
siTXNIP2	179,673,798 (98%)	25,159,908	125,301	39,503,312 (99.5%)	33,418,535

Table I-1. Summary of ATAC- and RNA-seq read counts before and after processing**.** For ATAC-seq, the number of properly paired reads, filtered/deduplicated reads, and identified narrow peaks are summarized. For RNA-seq, the number of filtered and alignable reads are summarized. *Filtered/dedup reads, filtered/deduplicated reads.

I-３.８ Chromatin accessibility is globally decreased upon TXNIP depletion

We detected 70,746 high-confidence accessible chromatin regions (ACRs) across all samples, most of which were located in gene bodies (38.74%), followed by intergenic regions (32.03%) and promoter regions (29.23%, Figure I-16). TXNIP knockdown appeared to induce a global decrease in chromatin accessibility in many genomic regions including promoters (Figure I-17A and B). Of the high-confidence ACRs, 7.38% were dACRs under TXNIP depletion; most dACRs showed reduced chromatin accessibility under this condition. dACRs(-) were preferentially localized in gene bodies, whereas dACRs(+) were more often observed in promoter regions (Figure I-17C).

The global chromatin changes induced by TXNIP knockdown could impact gene expression at corresponding loci. In fact, our gene expression analysis showed that 956 genes were downregulated, and 295 genes were upregulated by TXNIP knockdown compared to the control (Figure I-18A), suggesting that the global decrease in chromatin accessibility induced by TXNIP depletion would mediate the repression of gene expression. To confirm this phenomenon, we first selected sets of differentially (“Down” and “Up” in Figure I-18B and C) and non-differentially expressed genes (“None” in Figure I-18B and C) with at least one detectable ACR in promoter or gene body. Next, the cumulative distribution function (CDF) of accessibility changes demonstrated that the genes with a decreased RNA level (“Down”) showed significantly reduced chromatin accessibilities at promoters compared to those with no changes in the RNA level (“None”) (Figure I-18B: $P < 5.81 \text{ X } 10^{-28}$ for max changes, Figure I-18C: $P < 3.76 \text{ X } 10^{-32}$ for mean changes, Kolmogorov-Smirnov (KS) test). In contrast, genes with increased RNA expression (“Up”) exhibited no changes in chromatin accessibility at the promoter (Figure I-18B: $P < 0.68$ for max changes, Figure I-18C: $P < 0.49$ for mean changes, KS test), indicating that chromatin opening at promoters is necessary but not sufficient to induce gene expression. ACRs located in gene bodies also showed a similar trend: genes with a decreased RNA level (“Down”) showing decreased chromatin accessibility upon TXNIP depletion (Figure I-19A: *P* < 0.002 for max changes, Figure I-19B: *P* < 7.68 X 10^{-7} for mean changes, KS test), suggesting that TXNIP is likely to be a negative regulator of chromatin repressors that induce heterochromatin formation. We then used GO analysis(Raudvere et al., 2019) to examine the biological functions of genes that exhibited decreased chromatin accessibility at their promoter and decreased RNA expression upon TXNIP knockdown. In general, genes associated with developmental process, signaling receptor binding, cell adhesion and migration, immune response and extracellular matrix constituents appeared to be repressed upon TXNIP depletion (Figure I-20).

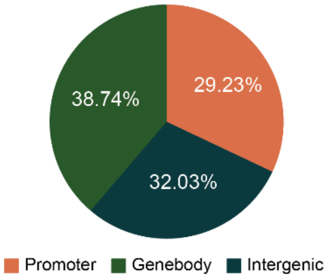


Figure I-16. Genomic distribution of ACRs

Genomic locations of 70,746 consensus ACRs identified from ATAC-seq analysis.

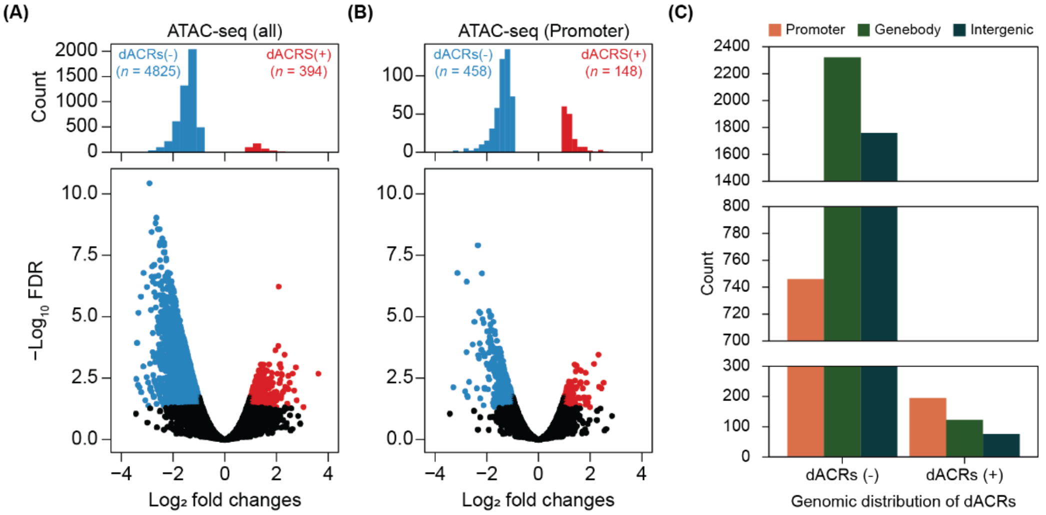


Figure I-17. Global decrease in chromatin accessibility is induced upon TXNIP depletion in HeLa cells

Volcano plots of differential chromatin accessibility for all ACRs **(A)** and those associated with promoters **(B)**. (A-B) Blue dots denote "dACRs(-)", which are differential accessible chromatin regions that exhibit significantly decreased chromatin accessibility in siTXNIP-treated cells (FDR ≤ 0.05, $\log_2$(siTXNIP / siCon) ≤ -1); red dots denote "dACRs(+)", which are differential accessible chromatin regions that exhibit significantly increased chromatin accessibility in siTXNIP-treated cells (FDR ≤ 0.05, $\log_2$(siTXNIP / siCon) ≥ 1). Black dots denote data points with no significant changes. **(C)** Genomic locations of 4,825 dACRs(-) and 394 dACRs(+) are depicted.

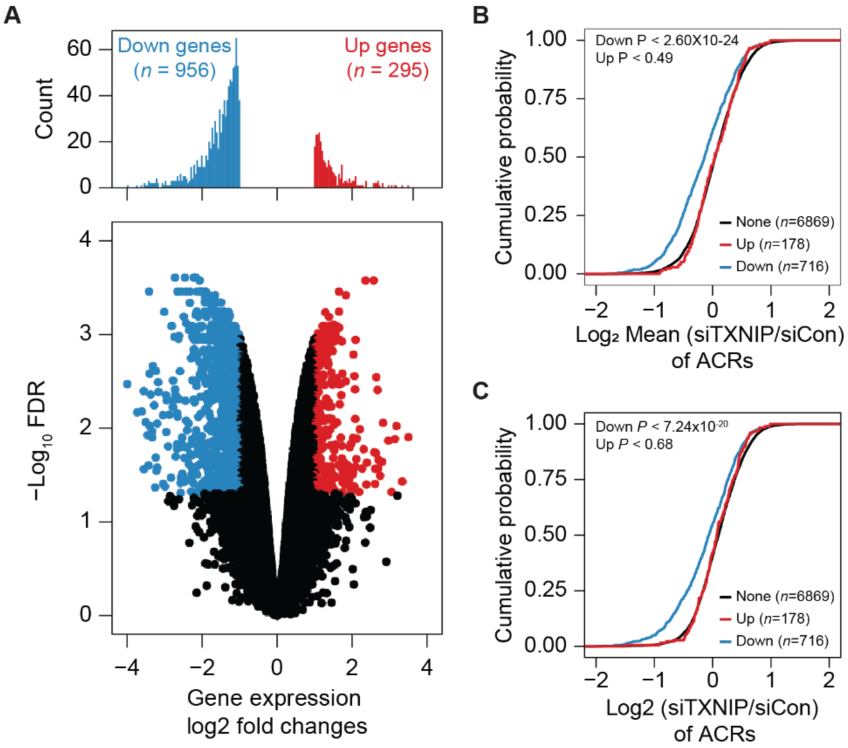


Figure I-18. Global decrease in chromatin accessibility in gene promoters upon TXNIP depletion is significantly associated with repression of gene expression

(A) Volcano plots of differential gene expression. Blue dots denote “Down” genes, which are significantly down-regulated genes in siTXNIP-treated cells (FDR ≤ 0.05, $\log_2$(siTXNIP / siCon) ≤ -1); red dots denote “Up” genes, which are significantly up-regulated genes in siTXNIP-treated cells (FDR ≤ 0.05, $\log_2$(siTXNIP / siCon) ≥ 1). Black dots denote data points with no significant changes. **(B)** Changes in chromatin accessibility of ACRs located in the promoter region of genes were plotted as CDFs. Genes were categorized into three groups based on changes in RNA levels (“Up”, “Down” as in **(A)** and “None” indicating genes with -0.5 ≤ $\log_2$(siTXNIP / siCon) ≤ 0.5. The number of genes in each group are shown in parentheses and *P* values in the left upper corner were calculated by one-sided KS test. **(C)** Cumulative distribution function (CDF) of mean changes in accessibility of all ACRs located in gene promoters. The genes were categorized into three groups (“None”, “Down”, and “Up”) as explained in **(B)**. *P* values on the left upper corner were calculated with the one-sided Kolmogorov-Smirnov (KS) test comparing “Down” or “Up” groups to the “None” group.

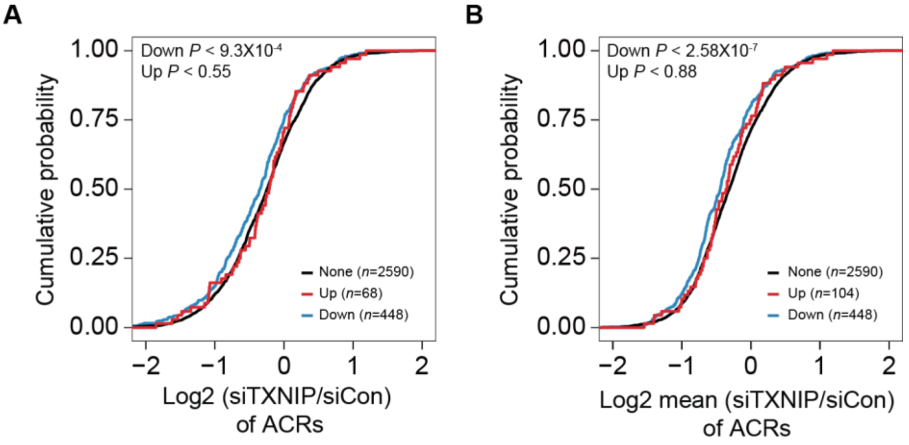


Figure I-19. Genes that exhibited decreased chromatin accessibility at their promoter and decreased RNA expression upon TXNIP depletion are associated various signaling pathways

Top 10 GO terms (biological process and molecular function) enriched in genes that exhibited decreased chromatin accessibility at their promoter and decreased RNA expression upon TXNIP knockdown.

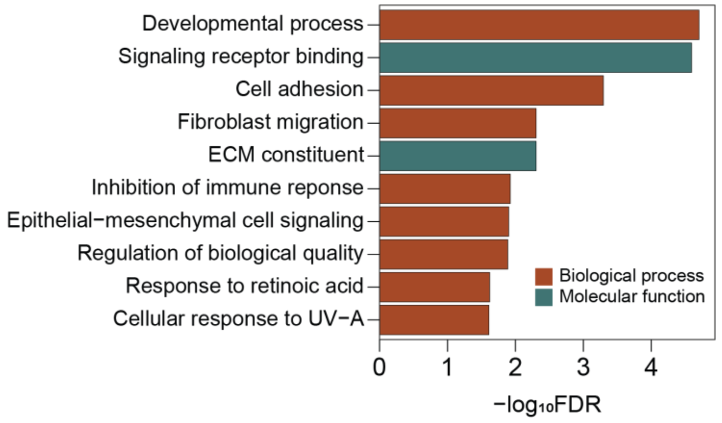


Figure I-20. Global decrease in chromatin accessibility in gene bodies upon TXNIP depletion is also significantly associated with repression of gene expression

CDF of changes in accessibility of ACRs located in gene bodies. Changes in accessibility of ACRs whose intensity is highest among all ACRs located in gene bodies are depicted on **(A)** and mean changes in accessibility of all ACRs located in gene bodies are depicted on the **(B)** right. *P* values on the upper left corners are calculated in the same manner as in Figure 18 B-C.

I-３.９ TXNIP represses the recruitment of HDAC2 to target loci

Given that TXNIP knockdown led to a global reduction in chromatin accessibility with decreased transcription, we focused on identifying the potential role of the epigenetic silencer HDAC2, one of the strong binding partners of TXNIP in the AP/MS analysis, in mediating the TXNIP-dependent epigenetic and transcriptional modulation. Consistent with the AP/MS data, immunoprecipitation (IP) experiments showed that the two proteins indeed interact with each other. Furthermore, TXNIP knockdown reduced the amount of TXNIP-interacting HDAC2 protein but did not affect the HDAC2 expression level (Figure I-21). To find out how the TXNIP-HDAC2 interaction impacts the epigenetic and transcriptional reprogramming of target loci, we first checked whether the TXNIP-HDAC2 interaction causes cytosolic retention of HDAC2 to inhibit nuclear HDAC2-mediated global histone deacetylation. However, both the expression level and subcellular localization of HDAC2 were unaffected by a reduction in TXNIP, as confirmed by Western blot analysis using cytoplasmic and nuclear fractions (Figure I-22A) as well as by an immunofluorescence assay (Figure I-22B), indicating that TXNIP might modulate HDAC2 activity in a different way.

We next asked if the transcriptional suppression of TXNIP-target genes was mediated by changes in HDAC2 recruitment to and histone acetylation of chromatin. To address this question, genes that were significantly downregulated by TXNIP knockdown and that contained at least one dACR in the promoter were selected by the following additional criteria: 1) the RNA level in normal HeLa cells is $\geq$ 10 TPM and 2) the total ATAC-seq read count at the promoter in siTXNIP-treated HeLa cells is reduced $\geq$ 1.5-fold compared to that in normal cells. Among the four TXNIP-target genes selected by the above-mentioned criteria, the expression levels of CD22 and L1CAM were significantly reduced ($P < 0.05$, Student’s T test, Figure I-23). The two genes were further examined to determine whether the levels of HDAC2-binding signal and histone acetylation in their promoter regions were changed upon TXNIP knockdown (Figure Ⅰ-24). We observed that RNA- and ATAC-seq coverages in exonic and promoter region of CD22 and L1CAM genes were clearly reduced upon TXNIP depletion (Figure Ⅰ-24 top) and an analysis of chromatin immunoprecipitation (ChIP) signals for HDAC2 and histone H3 acetylation at each dACR(-) detected in the L1CAM and CD22 promoters revealed that TXNIP knockdown increased the recruitment of HDAC2 to TXNIP-target loci, accompanied by decreased histone H3 acetylation (Figure Ⅰ-24 bottom). Therefore, these results suggest that the TXNIP interaction with HDAC2 inhibits the chromatin occupancy of HDAC2 and subsequently reduces histone deacetylation to facilitate global chromatin accessibility.

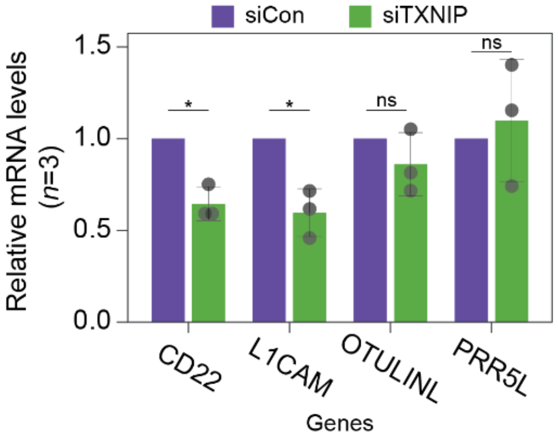


Figure I-22. Experimental validation of target genes repressed upon TXNIP depletion

RT-qPCR results of four target genes whose RNA expression and chromatin accessibility in their promoters, quantified using high-throughput sequencing data, were observed to be strongly repressed in HeLa cell. Data are presented as the mean ± sd, n=3). Gray dots depict actual values of each experiment. *P < 0.05, ns: not significant (two-sided paired Student T test).

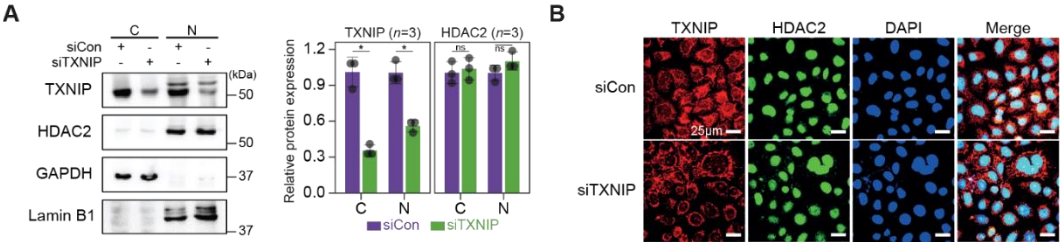


Figure I-21. TXNIP depletion does not affect the protein level or subcellular localization of HDAC2

(A) Nuclear and cytoplasmic fractions of HeLa cells were analyzed with Western blots following transfection with siCon or siTXNIP for 48 hr (left). Lamin B1 and GAPDH were used as nuclear and cytoplasmic markers, respectively. Western blot results from three independent experiments for TXNIP and HDAC2 were quantified as in Figure 4B. C, cytoplasm; N, nucleus. **(B)** Representative immunofluorescence images of TXNIP and HDAC2 after HeLa cells were transfected with either siCon or siTXNIP for 48 hr (magnification ×600); TXNIP (red), HDAC2 (green), and DAPI (blue).

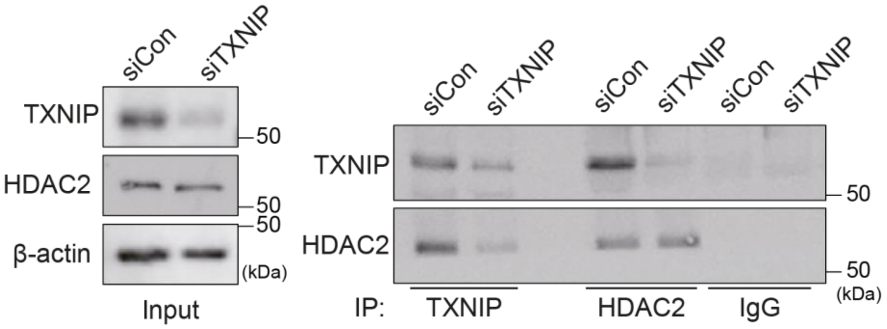


Figure I-23. Validation of TXNIP interaction with HDAC2 through co-IP

Analysis of co-IP between the TXNIP and HDAC2 proteins. Lysates from HeLa cells that had been treated with either siCon or siTXNIP for 48 hr were subjected to IP and immunoblotting with antibodies recognizing TXNIP and HDAC2, with IgG used as the negative control.

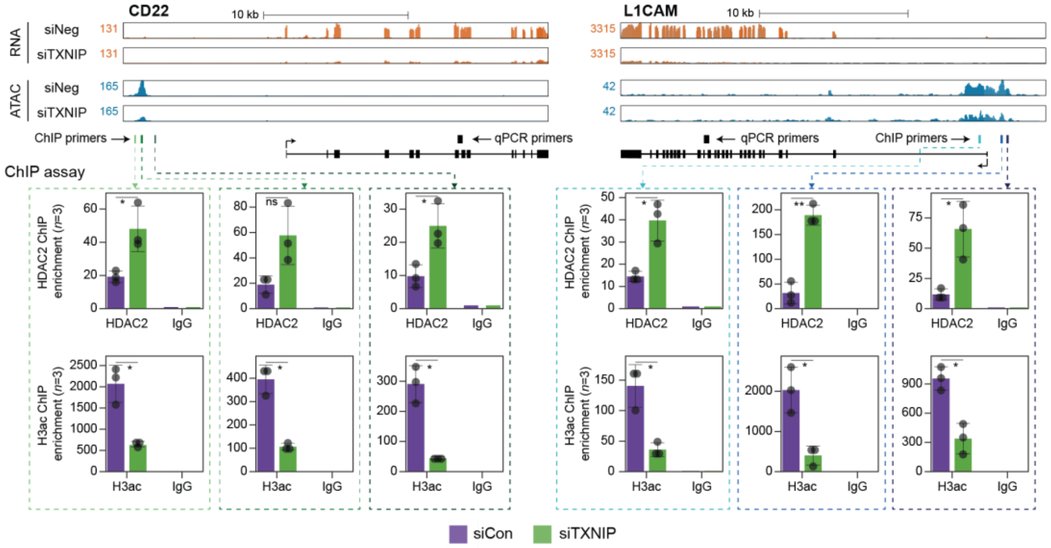


Figure I-24. TXNIP directly represses the recruitment of HDAC2 to target loci

Genomic regions showing RNA expression and chromatin accessibility at CD22 and L1CAM gene loci (top). Through the ChIP-qPCR analysis, the fold enrichment of HDAC2 and histone H3 acetylation (H3ac) at the CD22 and L1CAM promoter regions in HeLa cells treated with either siCon or siTXNIP for 48 hr were quantified (bottom). Data are presented as the mean ± sd (*n*=3, biological replicates). Gray dots depict actual values of each experiment. **P* < 0.05, ***P* < 0.01, ns : not significant (two-sided paired Student T test).

Ⅰ-３.１０　ARRDC5 plays a role in osteoclast differentiation and function

Given that various subunits of the V-type ATPase interact with ARRDC5, we speculated that ARRDC5 might be involved in the function of this complex (FigureⅠ-25). V-type ATPase plays an important role in the differentiation and function of osteoclasts, which are multinucleated cells responsible for bone resorption in mammals (Feng et al., 2009; Qin et al., 2012). Therefore, we hypothesized that ARRDC5 might be also important for osteoclast differentiation and function. To determine whether ARRDC5 affects osteoclast function, we prepared osteoclasts by infecting bone marrow-derived macrophages (BMMs) with lentivirus expressing either GFP-GFP or GFP-ARRDC5 and differentiating the cells into mature osteoclasts. After five days of differentiation, ectopic expression of GFP-ARRDC5 had significantly increased the total number of tartrate resistant acid phosphatase (TRAP)-positive multinucleated cells compared to GFP-GFP overexpression (Figure Ⅰ-26A). In particular, the number of TRAP-positive osteoclasts with a diameter larger than 200 μm was significantly increased by GFP-ARRDC5 overexpression (Figure Ⅰ-26A), suggesting that ARRDC5 expression increased osteoclast differentiation. Additionally, the area of resorption pits produced by GFP-ARRDC5-expressing osteoclasts in a bone resorption pit assay was approximately 4-fold greater than that of GFP-GFP expressing osteoclasts (Figure Ⅰ-26B). These results imply that the ectopic expression of ARRDC5 promotes osteoclast differentiation and bone resorption activity.

The V-type ATPase is localized at the osteoclast PM (Toyomura et al., 2003) and its localization is disrupted by bafilomycin A1, which is shown to attenuate transport of the V-type ATPase to the membrane (Matsumoto & Nakanishi-Matsui, 2019). We analyzed changes in V-type ATPase localization in GFP-GFP and GFP-ARRDC5 overexpressing osteoclasts. GFP signals were detected at the cell cortex when GFP-ARRDC5 was overexpressed, indicating that ARRDC5 might also localized to the osteoclast PM (Figure Ⅰ-27). In addition, we detected more V-type ATPase signals at the cell cortex in the GFP-

ARRDC5 overexpressing osteoclasts, and ARRDC5 and V-type ATPase were co-localized at the osteoclast membrane (Figure I-27). Notably, bafilomycin A1 treatment reduced not only the V-type ATPase signals detected at the cortex but also the GFP-ARRDC5 signals (Figure I-27). These results indicate that ARRDC5 might control the membrane localization of the V-type during osteoclast differentiation and function.

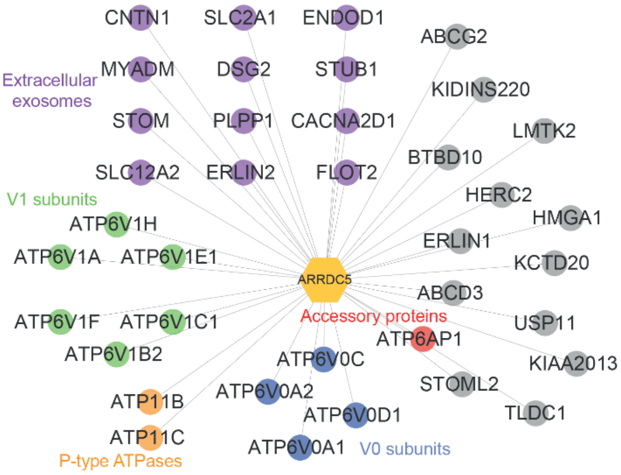


Figure I-25. Interaction of ARRDC5 with the ATPases and extracellular exosome related proteins

The human ARRDC5-centric PPI network. V-type and P-type ATPases, their related components, and extracellular exosomes are labeled and colored. Other interacting proteins are indicated with gray circles.

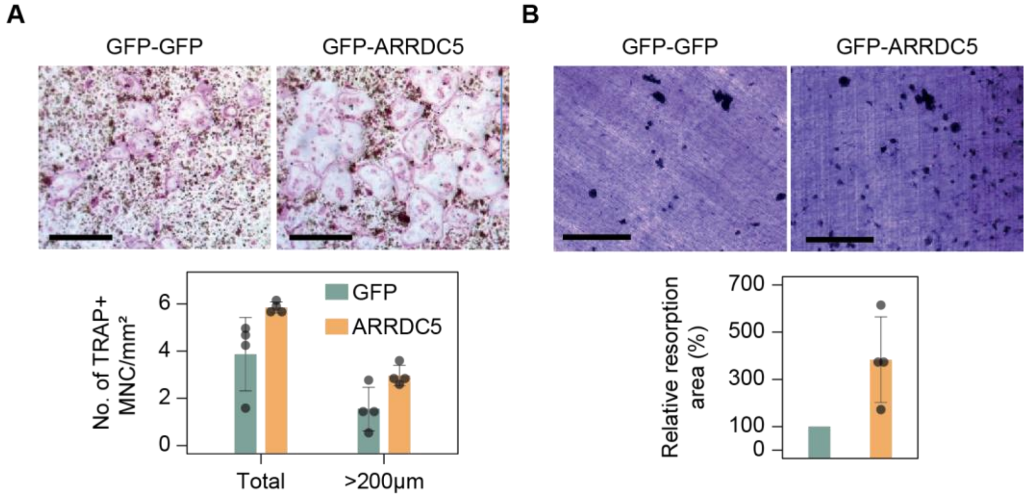


Figure I-26. Ectopic expression of ARRDC5 promotes osteoclast differentiation and bone resorption activity

(A) TRAP staining of osteoclasts. Cell differentiation was visualized with TRAP staining of GFP-GFP or GFP-ARRDC5 overexpressing osteoclasts (scale bar = 500 μm). TRAP-positive multinucleated cells (TRAP+ MNC) were quantified as the total number of cells and the number of cells whose diameters were greater than 200 μm. * *P* < 0.05. **(B)** Resorption pit formation on dentin slices. Cell activity was determined by measuring the level of resorption pit formation in GFP-GFP or GFP-Arrdc5 overexpressing osteoclasts (scale bar = 200 μm). Resorption pits were quantified as the percentage of resorbed bone area per the total dentin disc area using ImageJ software. The resorption area is relative to that in dentin discs seeded with GFP-GFP overexpressing osteoclasts, which was set to 100%. ** *P* < 0.01.

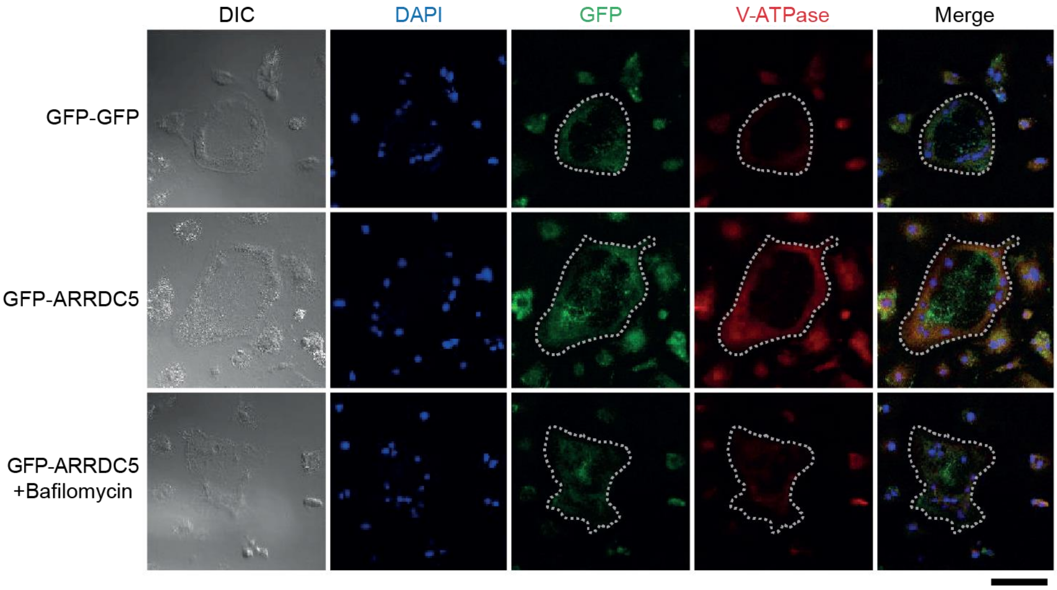


Figure I-27. ARRDC5 might control the membrane localization of the V-type during osteoclast differentiation and function

Localization of Arrdc5 and the V-type ATPase in osteoclasts. The V-type ATPase was visualized with immunofluorescence (red), GFP-GFP and GFP-ARRDC5 were visualized with GFP fluorescence (green), and nuclei were visualized with DAPI (blue). Representative fluorescence images are shown. Dashed lines were used to outline representative osteoclasts (scale bar = 100 μm).

I-３.１１ Alternative RNA splicing induced upon perturbation of ARRDC3 gene expression

To assess association of ARRDC3 in regulation of RNA alternative splicing (AS), we knocked down ARRDC3 in HeLa cells using the same procedure as for TXNIP. Illumina RNA-seq data were produced for WT and ARRDC3 depleted condition of HeLa cells and RNA AS events were analyzed by Whippet (Sterne-Weiler, Weatheritt, Best, Ha, & Blencowe, 2018). For the comparison, we performed an identical analysis on RNA-seq data from TXNIP- perturbated conditions and publicly available RNA-seq(W. Xiao et al., 2016) generated in HeLa cells under splicing factors-perturbated conditions, resulting in delta (Δ) percent spliced in (PSI) between normal and knock down conditions. PSI is a metric used to quantify AS events in RNA-seq data through measuring the proportion of mRNA transcripts that include a specific exon or splicing junction relative to the total number of transcripts generated from the same gene (Figure I-28A). It ranges from 0 to 1 and 0 indicates that the exon or splice junction is completely excluded from all transcripts and 1 indicates that exon or splicing is completely included in all transcripts. Among the analyzed AS events, alternative last exons (AL) and core exons (CE) were notably affected under ARRDC3-depleted conditions. Specifically, ARRDC3 knockdown induced inclusion of exons with proximal 3' ends, suggesting a role for ARRDC3 in regulating the selection of alternative last exons. In addition, the number of significant AS events under ARRDC3-depleted condition was comparable to those observed under splicing factor-depleted condition, with the count of AL events being the highest among all the conditions analyzed (Figure I-28B). This trend could not be observed in TXNIP-depleted condition, suggesting unique regulatory axis on RNA AS by ARRDC3.

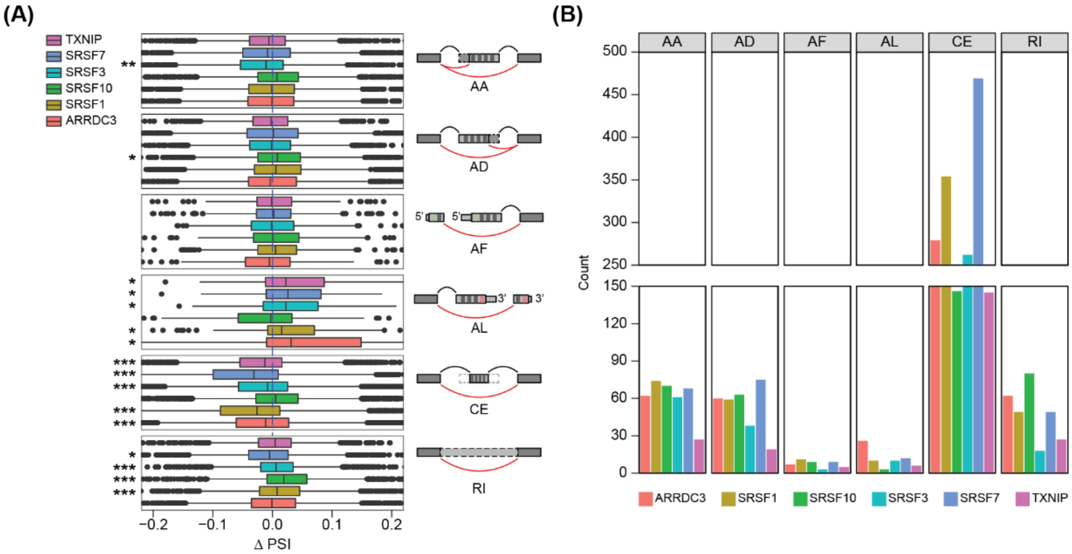


Figure I-28. RNA alternative splicing landscape upon depletion of ARRDC3, TXNIP and splicing factors

(A) Boxplot showing Δ PSI between normal and conditions in which α-arrestins or splicing factor was knocked down. Six types of AS events were analyzed here: AA, alternative acceptor splice site ; AD, alternative donor splice site; AF, alternative first exon; AL, alternative last exon; CE, core exon; RI, retained intro. **(B)** The number of significant AS events under knock condition of ARRDC3, TXNIP or splicing factors. Significant AS events are defined as follows: |Δ PSI| $\geq \mathbf{0.2}$ and Whippet probability $\geq$ 0.9.

Ⅰ-３.１２ Vdup1 affect hematopoiesis of *Drosophila* larvae

Next, Vdup1, which is another *Drosophila* α-arrestins related with RNA splicing complex (Figure Ⅰ-10), was inspected. Using the scRNA-seq data of lymph gland in *Drosophila* larvae from previous study (Cho et al., 2020) , we first quantified gene expression levels of Vdup1 in specific cell types. Vdup1 was highly expressed prohemocyte 1(PH 1) cell type, which was reported to be most naïve subcluster of PH cell types and shown to express Notch and Delta gene in high levels (Cho et al., 2020) (Figure Ⅰ-29A).

To identify function of Vdup1 in the larval hematopoietic organ, lymph gland, we generated Vdup1 mutant using the CRISPR/Cas9 method. We used two different gRNAs that target different regions of Vdup1 DNA. One of the gRNA targets first exon and the other gRNA targets second exon (Figure Ⅰ-29B). To examine the effect of Vdup1 mutation on blood cell differentiation, we used antibodies against Pxn or NimC1 for plasmatocyte and Hnt for crystal cells to check the differentiation phenotype in the mutant larvae, respectively. Additionally, we validated the expression of *STAT::edGFP* in the lymph gland which is known to be expressed in the PH1 population in the lymph gland (Cho et al., 2020). As a result, we found that *Vdup1* mutant larvae had smaller lymph glands than the wild-type larvae (Figure Ⅰ-29C and D; Figure Ⅰ-30A). Furthermore, the Vdup1 mutant larvae did not show any PH1 marker expression (*STAT::edGFP*) in the lymph gland (Figure Ⅰ-29C and D). However, the proportion of plasmatocytes is found to be increased compared to the wild-type (Figure Ⅰ-29C and D; Figure Ⅰ-30B and C), while the number of crystal cells remained unchanged (Figure Ⅰ-29C and D; Figure Ⅰ-30D). These results suggest that Vdup1 specifically regulate PH1 cell type, which was identified to be a precursor of prohemocytes reminiscent of mammalian hematopoietic stem cells in the previous study (Cho et al., 2020). Regulatory axis involving Vdup1 is yet to be discovered and further studies are required if this regulatory axis involves interaction of Vdup1 with RNA splicing complex, thus perturbating and regulating RNA AS of key transcripts. associated with PH1 cell types.

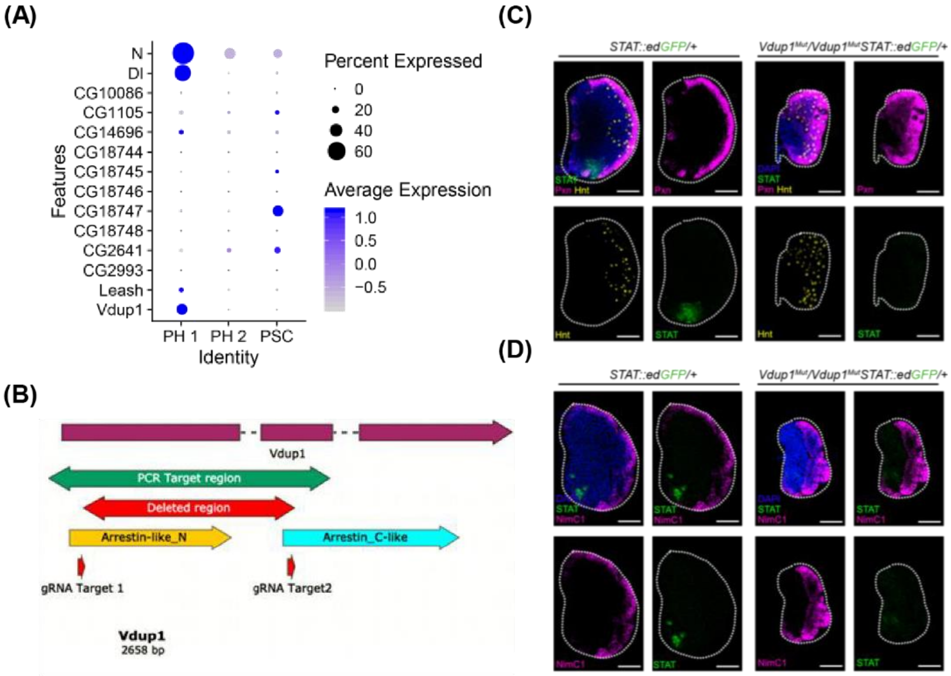


Figure I-30. Phenotypes of Vdup1 mutants in the *Drosophila* hematopoietic organ, lymph gland.

(A) Expression of Vdup1 in the *Drosophila* hematopoietic organ, lymph gland, based on the previous single cell RNA sequencing data (Cho et al., 2020). Vdup1 is expressed in the earliest prohemocyte population (PH1) that expresses Notch (N) and Delta (Dl). (**B)** Schematic representation of gRNA targeting regions of Vdup1 DNA. One of the gRNAs (Target 1; red, one side arrow) targets first exon and the other gRNA (Target 2; red one side arrow) targets second exon of Vdup1 DNA, respectively. Red two side arrow represents deleted region of Vdup1 mutant. Green two side arrow represents PCR target region that used for vdup1 mutant confirmation. Deletion of Vdup1 was validated by Sanger sequencing method. (**C-D)** Phenotype of Vdup1 homozygote mutant in the lymph gland. Compared to wild type lymph gland (*STAT::edGFP/+*), Vdup1 homozygote mutant (*Vdup1^{Mut}/Vdup1^{Mut}; STATedGFP/+*) shows small size of the lymph gland (DAPI; blue), loss of PH1+ cells (STAT; green), and increased differentiating plasmatocyte phenotype (Pxn; magenta). However, crystal cell (Hnt; yellow) does not changed (**C**). Increased plasmatocyte phenotype also confirmed by mature plasmatocyte marker NimC1 (Magenta) **(D)**. White dotted line demarcates lymph gland primary lobe. White scale bar: 40μm

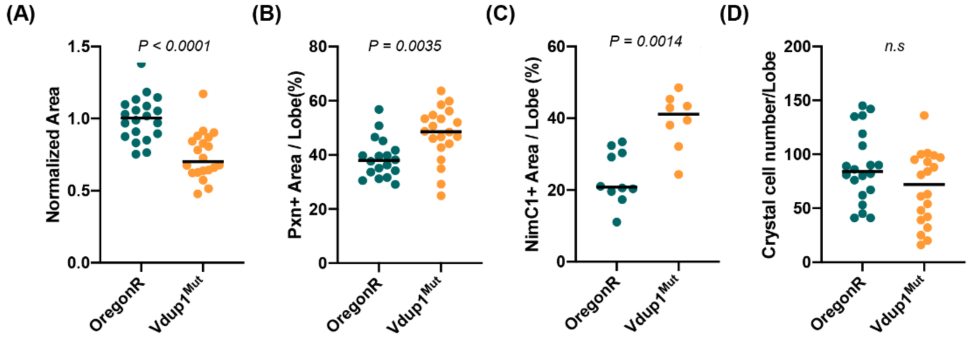


Figure I-29. Quantification of Figure I-29C and D.

(A) Normalized area of lymph gland primary lobe in both wild type (*Oregon R*) and *Vdup1* mutant. (**B)** Quantification of Pxn+ plasmatocyte area in both wild type (*Oregon R*) and *Vdup1* mutant. (**C)** Quantification of NimC1+ mature plasmatocyte area in both wild type (*Oregon R*) and *Vdup1* mutant. (**D)** Quantification of Hnt+ crystal cells in both wild type (*Oregon R*) and *Vdup1* mutant. *P*-value is annotated in the top of. each graph and "n.s" represent not significant. Mann-Whitney test was performed. Bar in the graph represents average.

I- 4 Discussion

We constructed high-confidence interactomes of α-arrestins from human and *Drosophila*, comprising 307 and 467 interacting proteins, respectively. The resulting interactomes greatly expanded previously known PPIs involving α-arrestins and the majority of interactomes were first reported in this study, which needs to be validated experimentally (Tian, Kang, & Benovic, 2014; Zbieralski & Wawrzycka, 2022). However, some known PPIs were missed in our interactomes due to low spectral counts and SAINTexpress scores, probably resulting from different cellular contexts, experimental conditions, or other factors (Figure I-5).

Integrative map of protein complexes that interact with α-arrestins (Figure I-9; Figure I-10) hint towards many aspects of α-arrestins's biology that remain uncharacterized. For example, role of α-arrestins in the regulation of β2AR in human remained controversial. One study proposed that α-arrestins might act coordinately with β-arrestins at the early step of endocytosis, promoting ubiquitination, internalization, endosomal sorting and lysosomal degradation of activated GPCRs (Shea et al., 2012). The another study, however, proposed different hypothesis suggesting that α-arrestins might act as secondary adaptor localized at endosomes to mediate endosomal sorting of cargo molecules (S. O. Han et al., 2013). Among the protein complexes that interact with α-arrestins, we identified those related with clathrin-coated pit in human (Figure I-9) and AP-2 adaptor complex in *Drosophila* (Figure I-10). They are multimeric proteins to induce internalization of cargo molecules to mediate clathrin-mediated endocytosis, which suggests involvement of α-arrestins in early step of endocytosis.

Among the interacting proteins, 58 orthologous interacting groups were observed to be conserved between human and *Drosophila,* suggesting conserved roles *of* α-arrestins between two species (Figure I-11). Among conserved proteins, proteins known to interact with human α-arrestins, such as NEDD4, WWP2, WWP1, and ITCH, were identified along with its orthologs in *Drosophila*, which are Su(dx), Nedd4, and Smurf, implying that regulatory pathway of ubiquitination-dependent proteolysis by α-arrestins is also present in invertebrate species. Besides the known conserved functions, the novel conserved functions of α-arrestins interactomes were also identified, such as RNA splicing (Figure I-7; Figure I-11). Because our protocol did not include treatment with RNase before the AP/MS, it is possible that RNA binding proteins could co-precipitate with other proteins that directly bind to α-arrestins through RNAs, and thus could be indirect binding partners. Nevertheless, other RNA binding proteins except for RNA splicing and processing factors were not enriched in our interactomes, indicating that this possibility may be not the case. Supporting this notion, we identified that perturbating ARRDC3 significantly altered specific types of RNA AS events, including the inclusion of alternative last exon and core exons, and affected a number of transcripts comparable to those under splicing factor-perturbated conditions. Besides ARRDC3, we also examined Vdup1, a *Drosophila* α-arrestin shown to interact with RNA splicing complex, and discovered that perturbating Vdup1 affected specific cell type, PH1, which resembles mammalian hematopoietic stem cells. Therefore, it might be of interest to explore how α-arrestins in both human and *Drosophila* are linked to RNA processing and subsequently regulate key signaling pathways and cellular compositions in future.

Some protein complexes and functional modules were found to be involved in specific cellular processes discovered in only human, suggesting that some functional roles of α-arrestins have diverged through evolution. As examples of specific cellular functions of α-arrestins, we explored the biological relevance of two interacting protein complexes: 1) the interaction between TXNIP and chromatin remodelers and 2) the interaction between ARRDC5 and the V-type ATPase complex. Given that TXNIP interacts with chromatin remodelers, such as the HDAC, we speculated that chromatin structures could be affected by the interactions. Although we showed that siTXNIP treatment directed a global decrease in chromatin accessibilities and gene expression by inhibiting the binding of HDAC2 to targets, histones themselves could be also controlled by the interaction between TXNIP and the H2B ubiquitination complex. An impact of TXNIP on histone ubiquitination could strengthen the negative regulation of target loci by siTXNIP treatment. In addition, TXNIP interacts with the proteasome, which induces the degradation of binding partners (Figure I-9). However, we observed that the cellular level and localization of HDAC2 were not affected by TXNIP reduction (Figure I-22), meaning that the proteasome seems not to be involved in TXNIP's influence on HDAC2; rather, TXNIP directly hinders HDAC2 recruitment to target loci.

Because the V-type ATPase plays a key role in osteoclast differentiation and physiology (Feng et al., 2009; Qin et al., 2012), we investigated a possible role for the ARRDC5-V-type ATPase interaction in this cell type. We showed that the ectopic expression of ARRDC5 increased both the differentiation of osteoclasts into their mature form and their bone reabsorption activity. Additionally, ARRDC5 co-localized with the V-type ATPase at the PM (Figure I-27). Thus, further characterization of ARRDC5 and its interactome in osteoclasts might clarify how ARRDC5 regulates the V-type ATPase to play a role in osteoclast differentiation and function. With the results, the discovery of new binding partners and their functions of TXNIP and ARRDC5 will facilitate the further investigations to explore the novel PPIs of α-arrestins.

Given the plethora of PPIs uncovered in this study, we also anticipate that our study could provide insight into many disease models. In fact, despite a limited knowledge of their biology, α-arrestins have already been linked to a range of cellular processes and several major health disorders, such as diabetes (Batista et al., 2020; Wondafrash et al., 2020), cardiovascular diseases (Domingues, Jolibois, Marquet de Rouge, & Nivet-Antoine, 2021), neurological disorders (Tsubaki, Tooyama, & Walker, 2020), and tumor progression (Y. Chen et al., 2020; Mohankumar et al., 2015; Oka et al., 2006), making them potential therapeutic targets. In addition, we summarized RNA and protein expression levels of α-arrestins in human tissues based on information from the Human protein atlas (Uhlen et al., 2015) (Figure I-31A). Except for ARRDC5, α-arrestins appear to be ubiquitously expressed across human tissues at the RNA level. In several of these tissues, protein expressions have also been confirmed, make them as promising targets for future studies aimed at elucidating biological functions and mechanisms involving α-arrestins. We could also find evidence of association of α-arrestins with a few cancer types, also making them as promising target for future studies of α-arrestins as therapeutic targets (Figure I-31B). In summary, using high-throughput AP/MS data, we have successfully identified and characterized comprehensive PPI networks involving α-arrestins in human and *Drosophila*. Using experimental approaches and computational analysis of other high-throughput multi-omics data, we have validated human-specific and conserved interactome and its' related biological functions involving α-arrestins (Figure I-32). For the community, we provide comprehensive α-arrestin interactome maps on our website (human: http://big.hanyang.ac.kr/alphaArrestin_Human and *Drosophila*: http://big.hanyang.ac.kr/alphaArrestin_Fly). Researchers can search and download their interactomes of interest as well as access information on potential cellular functions associated with these interactomes.

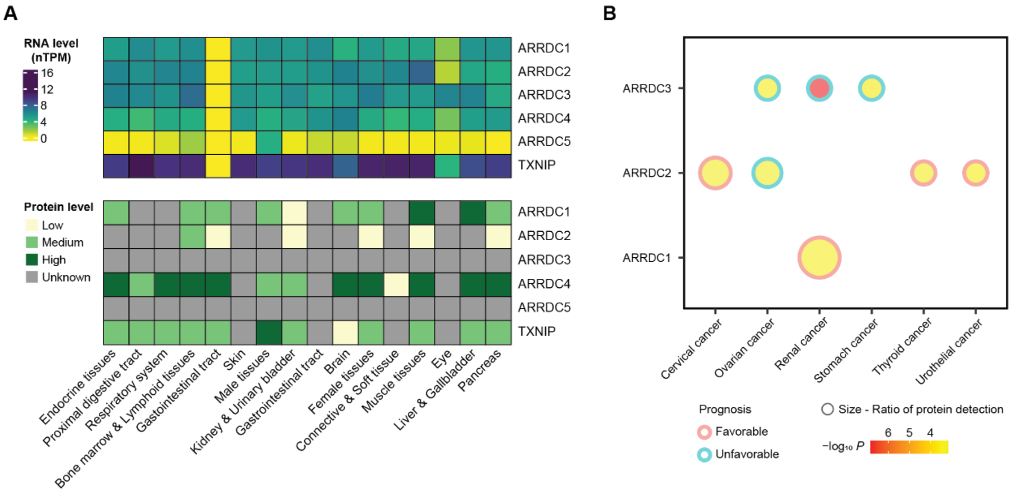


Figure I-32. RNA and protein expression levels of α-arrestins in normal tissues and prognosis of α-arrestins in cancer

(A) Consensus transcript expression levels (top) and protein levels (bottom) in normal tissue are depicted. nTPM is transcript per million values that were normalized by Trimmed mean of M values. Protein level was measured based on immunohistochemical data manually scored with regard to staining intensity and fraction of stained cells. **(B)** Prognostic summary of α-arrestins in cancers. Only the significant ones (P < 0.05) are depicted in here. All expression values and prognostic summary were derived from Human protein atlas (Uhlen et al., 2015).

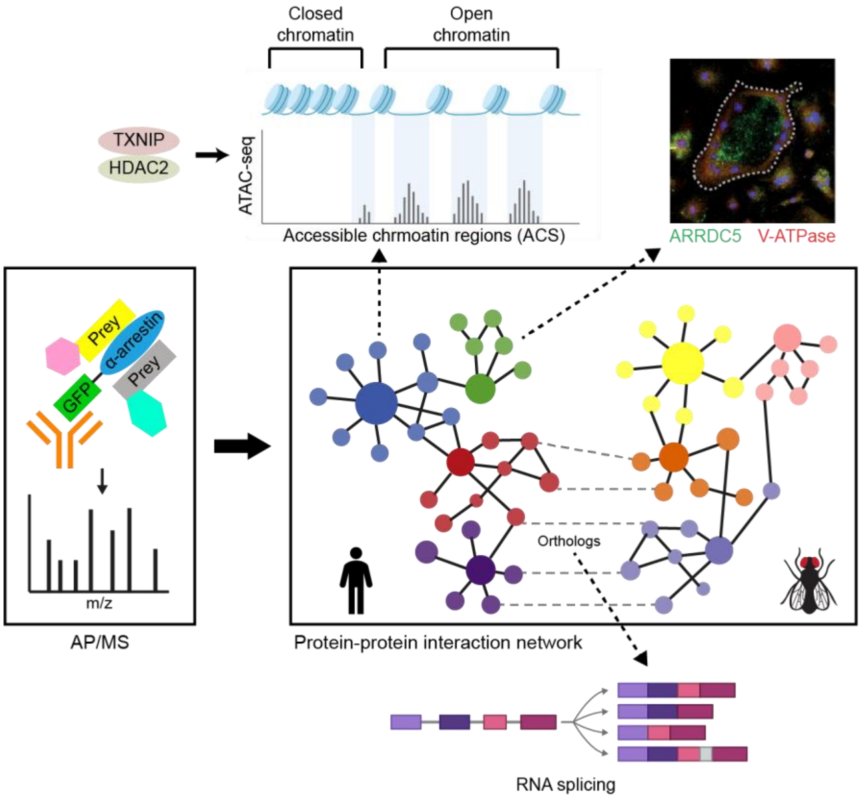


Figure I-31. Summary of chapter I: comparative Interactome Analysis of α-arrestin families in Human and *Drosophila*

I-５ Materials and Methods

I-５.１ Experimental procedures

I-５.１.１ Generating *Drosophila* α-arrestin-GFP fusion DNA constructs

To create Drosophila ARRDC entry clones, we gathered cDNA sequences of twelve *Drosophila* α-arrestins : CG2993 (#2276, Drosophila Genomics Resource Center, DGRC, Bloomington, IN, USA), CG18744 (#1388606, DGRC), CG18745 (#12871, DGRC), CG18746 #9217, DGRC), CG18747 (#1635366, DGRC), CG18748 (#1387253, DGRC), CG2641 (#1649402, DGRC), CG10086 (#8816, DGRC), CG14696 (#1644977, DGRC), CG1105 (#4234, DGRC), Vdup1 (#1649326, DGRC), and Leash (Y. Kwon et al., 2013). We then subcloned each cDNA sequence of *Drosophila* α-arrestins into pCR8 entry clone vector using pCR8/GW/TOPO TA cloning kit (#K250020, Thermo Fisher Scientific, Waltham MA, USA), by following manufacturer's protocol. To generate plasmids with suitable system for protein expression in *Drosophila* cell culture, we then subcloned these α-arrestins-containing-pCR8 plasmids into pMK33-Gateway-GFP destination vector (Y. Kwon et al., 2013; Kyriakakis, Tipping, Abed, & Veraksa, 2008) using Gateway LR Clonase II enzyme mix (#11791020, Thermo Fisher Scientific), where coding sequences of α-arrestins are inserted before GFP sequence. Final constructs were validated by GENEWIZ Sanger Sequencing.

I-５.１.２ Establishing *Drosophila* α-arrestin-GFP stably expressing cell lines

S2R+ cells were maintained in Schneider's *Drosophila* Medium (#21720024, Thermo Fisher Scientific) supplemented with 10% heat inactivated FBS (#16140071, Thermo Fisher Scientific) and 1% Penicillin Streptomycin (#15070063, Thermo Fisher Scientific) at 24°C. To establish α-arrestin-GFP stably expressing *Drosophila* cell lines,

0.4×10^6 S2R+ cells were seeded in 6-well plates and were transfected with 1 µg of each pMK33-ARRDC-GFP construct using Effectene transfection reagent (#301425, Qiagen, Venlo, Netherlands). pMK33 plasmid is a copper-induced protein expression vector, which carries Hygromycin B-antibiotic-resistant gene. Therefore, we selected for α-arrestin-GFP stable cell lines by maintaining cells in Schneider's *Drosophila* Medium supplemented with 200 µM Hygromycin B (#40-005, Fisher Scientific). The stable cells were transferred into T25 cm^2 flasks to repopulate. To induce the expression of α-arrestin-GFP fusion proteins, we exposed the stable cells to 500 µM $CuSO_4$ (#C8027, Sigma Aldrich, Burlington, MA, USA) to the media. We confirmed the GFP-tagged α-arrestin protein expressions using fluorescence microscopy.

I-5.1.3 Synthesizing human α-arrestin coding sequence

Due to the lack of commercially available stock, we utilized GENEWIZ (South Plainfield, NJ, USA) gene synthesis service to synthesize human ARRDC5 coding sequence (NM_001080523).

I-5.1.4 Generating mammalian GFP- α-arrestin fusion DNA constructs

To create human α-arrestin entry clones, we subcloned ARRDC3 (#38317, Addgene, Watertown, MA, USA) and ARRDC5 (GENEWIZ) into pCR8 entry clone vector using pCR8/GW/TOPO TA cloning kit (#K250020, (Thermo Fisher Scientific), by following manufacturer's protocol. ARRDC1 (BC032346, GeneBank), ARRDC2 (BC022516, GeneBank), ARRDC4 (BC070100, GeneBank), and TXNIP (BC093702, GeneBank) were cloned into pCR8. To generate plasmids with suitable system for protein expression in mammalian cell culture, we then subcloned these α-arrestin s-containing-pCR8 plasmids into pHAGE-GFP-Gateway destination vector (gift from Dr. Chanhee Kang at Seoul National Univesity) using Gateway LR Clonase II enzyme mix (#11791020, Thermo Fisher

Scientific), where coding sequences of α-arrestin are inserted after GFP sequence. Final constructs were validated by GENEWIZ Sanger Sequencing.

I-5.1.5 Establishing mammalian GFP- α-arrestin stably expressing cell lines

We produced GFP-α-arrestins lentiviral particles by seeding 5 x10^6 HEK293T cells in 10 cm^2 dish with Dulbecco's Modified Eagle Medium (#11965118, Thermo Fisher Scientific) supplemented with 25 mM HEPES, 10% heat-inactivated fetal bovine serum (#16140071, Thermo Fisher Scientific), and 1% Penicillin Streptomycin (#15070063, Thermo Fisher Scientific) at 37°C in humidified air with 5% CO_2. Approximately after 16-24 hours (hr), at 90% cell confluency, we changed the cell media to Opti-MEM medium (#31985070, Thermo Fisher Scientific) and transfected the cells with 10 µg pHAGE-GFP-α-arrestin construct, 10 µg lentivirus packaging plasmid (pCMV-dR8.91), and 10 µg virus envelope plasmid (VSV-G) using PEIPro DNA transfection reagent (#115010, VWR, Radnor, PA, USA). GFP-α-arrestins lentiviral particles were harvested 40 hr-post transfections. To establish GFP-α-arrestins stably expressing mammalian cell lines, HEK293 cells were seeded in 10 cm^2 dish with Dulbecco's Modified Eagle Medium (#11965118, Thermo Fisher Scientific) supplemented with 25 mM HEPES, 10% heat-inactivated fetal bovine serum (#16140071, Thermo Fisher Scientific) and 1% Penicillin Streptomycin (#15070063, Thermo Fisher Scientific) at 37°C in humidified air with 5% CO_2. At 90% cell confluency, cells were infected with pHAGE-GFP-ARRDC lentivirus particle, and stable cells were selected by maintaining cells in media supplemented with1.5 µg/mL puromycin (#BP2956100, Thermo Fisher Scientific). We confirmed the GFP-tagged α-arrestin protein expressions using fluorescence microscopy.

I-5.1.6 Immunofluorescence imaging of human α-arrestins

Stably α-arrestin-GFP expressing HEK293 cells were cultured in a 12 well-plate with pre-sterilized round glass coverslips in each well. Cells on coverslip were fixed in 4% paraformaldehyde (PFA) (RT15710, Electron Microscopy Sciences, Hatfiled, PA, USA) diluted in PBS for 30 min and then washed three times with PBST (PBS supplemented with 0.2% Triton X-100) with 5 min intervals. To label the nucleus, samples were stained with DAPI (1:5000; D9542, Sigma Aldrich) in PBST supplemented with 1% BSA (A7906, Sigma Aldrich) for 1 hr at room temperature. Stained cells samples were washed three times with PBST and preserved in Vectashield (H-1000, Vector Laboratories, Burlingame, CA, USA). Fluorescence images were acquired using an Olympus FV1200 confocal microscope with 40X oil objective lens and 2X zoom factor. NIH ImageJ software was used for further adjustment and assembly of the acquired images.

I-５.１.７ Affinity purification of *Drosophila* and human GFP-tagged α-arrestin complexes

We seeded each of the *Drosophila* α-arrestin-GFP stable cells in six T-75 cm^2 flasks (2.1x 10^6 cells per flask) and α-arrestin-GFP expression was induced for 48 hr with 500 μM $CuSO_4$. Meanwhile, we seeded each of the human GFP-α-arrestin stable cells in eight T-75 cm^2 flasks and grown for 48 hr before collection. The cells were harvested by spinning down cells at 1,000g for 5 minutes (min) and washed once with cold PBS. We lysed the cells by resuspending cells in lysis buffer (10 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.5 mM EDTA, 1.5 mM $MgCl_2$, 5% glycerol, 0.5% NP-40, 25 mM NaF, 1mM DTT, and 1x HALT protease and phosphatase inhibitor (#PI78442, Thermo Fisher Scientific)) and incubating them for 40 min. The lysate was separated from the insoluble fraction by centrifugation at 20,000 g for 15 min at 4℃. To capture the α-arrestins and their native interactors, each α-arrestin-containing lysate was incubated with GFP-nanobody-conjugated to Dynabeads M-270 Epoxy magnetic beads (#14301, Thermo Fisher Scientific). The supernatant was separated from the beads using a magnetic rack, and the beads were washed five times with lysis buffer. The protein complexes were eluted from the beads by adding 200 mM glycine pH 2.5 and the pH was neutralized with Tris base pH 10.4. Purified α-arrestin proteins were confirmed by running a fraction of the eluted proteins on SDS-PAGE/Coomassie gel.

I-5.1.8 Protein sample preparation for mass spectrometry

To digest protein samples into peptides for mass spectrometry analysis, we precipitated the eluted proteins by adding trichloroacetic acid (#T0699, Sigma Aldrich) to 20% final concentration, followed by spinning down samples at maximum speed for 30 min at 4℃. The precipitates were washed with 10% trichloroacetic acid solution and three additional times with Acetone (#A929, Thermo Fisher Scientific), and left to dry in room temperature. Protein precipitations were digested with Trypsin (Promega, #V5113) diluted in Digestion buffer (100 mM Ammonium Bicarbonate and 10% Acetonitrile) in 1:40 ratio. Resulting peptides were purified using ZipTip Pipet tips (#ZTC18M096, Thermo Fisher Scientific).

I-5.1.9 LC/MS-MS analysis

We used cells stably expressing GFP and wild-type HEK293 or S2R+ cells alone as control baits. AP/MS experiments for all *Drosophila* and human α-arrestin baits were performed in two biological replicates, with the exception of human ARRDC3 baits (two technical replicates). Samples were resuspended in Mass Spectrometry buffer (5% Formic Acid and 5% Acetonitrile) and were analyzed on an Liquid Chromatography Orbitrap Fusion Lumos Tribrid Mass Spectrometer (#IQLAAEGAAPFADBMBHQ, Thermo Fisher Scientific) equipped with a nano-Acquity UPLC system and an in-house developed nano spray ionization source. Peptides were separated using a linear gradient, from 5-30% solvent B (LC-MS grade 0.1% formic acid (#A117, Thermo Fisher Scientific) and acetonitrile) in a 130 min period at a flow rate of 300 nL/min. The column temperature was maintained at a constant 5^{o}C during all experiments. Peptides were detected using a data dependent method. Survey scans of peptide precursors were performed in the Orbitrap mass analyzer from 380 to 1500 m/z at 120K resolution (at 200 m/z) with a 5 x 105 ion count target and a maximum injection time of 50 milliseconds (ms). The instrument was set to run in top speed mode with 3 seconds (sec) cycles for the survey and the MS/MS scans.

I-5.1.10 TXNIP knockdown in HeLa cells

HeLa cells (CCL-2; ATCC, Manassas, VA, USA) were cultured in complete DMEM supplemented with 10% FBS and 1% penicillin-streptomycin. Cells were cultured in an incubator at 37°C in humidified air containing 5% CO2. For siRNA-induced knockdown of TXNIP in HeLa cells, the following siRNA duplex was synthesized (Bioneer, Daejeon, South Korea): sense: 5’-GUCAGUCACUCUCAGCCAUdTdT -3', anti-sense: 5'-AUGGCUGAGAGUGACUGACdTdT-3'. Random sequence siRNAs (AccuTarget Negative control siRNA; Bioneer), which are non-targeting siRNAs that have low sequence homology with all humans, mouse, and rat genes, were used as negative controls (siCon). 100 nM of each siRNA was transfected into 105 HeLa cells using Lipofectamine RNAiMAX (#13778075, Invitrogen, Carlsbad, CA, USA; Thermo Fisher Scientific) according to the manufacturer's instructions. Transfected cells were harvested after 48 hr for RNA-seq and ATAC-seq (two biological replicates for each sequencing data).

I-5.1.11 RNA sequencing

For RNA-seq, total RNA was extracted using TRIzol (#15596018, Invitrogen; Thermo Fisher Scientific) according to the manufacturer's protocol. Total RNA concentration was calculated by Quant-IT RiboGreen (#R11490, Invitrogen; Thermo Fisher Scientific). To assess the integrity of the total RNA, samples are run on the TapeStation RNA screentape (#5067-5576, Agilent Technologies, Santa Clara, CA, USA). Only high-quality RNA preparations, with RNA integrity number greater than 7.0, were used for RNA library construction. A library was independently prepared with 1ug of total RNA for each sample by Illumina TruSeq Stranded mRNA Sample Prep Kit (#RS-122-2101, Illumina, Inc., San Diego, CA, USA). The first step in the workflow involves purifying the poly-A containing mRNA molecules using poly-T-attached magnetic beads. Following purification, the mRNA is fragmented into small pieces using divalent cations under elevated temperature. The cleaved RNA fragments are copied into first strand cDNA using SuperScript II reverse transcriptase (#18064014, Invitrogen, Thermo Fisher Scientific) and random primers. This is followed by second strand cDNA synthesis using DNA Polymerase I, RNase H and dUTP. These cDNA fragments then go through an end repair process, the addition of a single 'A' base, and then ligation of the adapters. The products are then purified and enriched with PCR to create the final cDNA library. The libraries were quantified using KAPA Library Quantification kits (#KK4854, KAPA BIOSYSTEMS, Wilmington, MA, USA) for Illumina Sequencing platforms according to the qPCR Quantification Protocol Guide and qualified using the TapeStation D1000 ScreenTape (#5067-5582, Agilent Technologies). Indexed libraries were then submitted to an Illumina NovaSeq 6000 (Illumina, Inc.) as the paired-end (2×100 bp) sequencing. Both library preparation and sequencing were performed by the Macrogen (Macrogen, Inc., Seoul, South Korea).

I-５.１.１２ ATAC sequencing

100,000 cells were prepared using LUNA-FL™ Automated Fluorescence Cell Counter (#L20001, logos biosystems, Gyeonggi-do, South Korea). Cells were lysed using cold lysis buffer, which consist of Nuclease-free water (#10977023, Invitrogen; Thermo Fisher Scientific), IGEPAL CA-630 (#I8896, Sigma Aldrich), 1M Trizma HCl(PH7.4) (#T2194, Sigma Aldrich), 5M NaCl (#59222C, Sigma Aldrich), and 1M MgCl2 (#M1028, Sigma Aldrich). The nuclei concentration was determined using Countess II Automated Cell Counter (#AMQAX1000, Thermo Fisher Scientific) and nuclei morphology was examined using microscopy. Immediately after lysis, resuspend nuclei (50,000 cells) were put in transposition reaction mix 50 μl, which consist of TED1 2.5μl and TD 17.5 μl (#20034197, Illumina, Inc.), nuclease free water 15 μl, and the nuclei resuspension (50,000 nuclei, 15 μl). The transposition reaction was incubated for 30 min at 37°C. Immediately following transposition, the products were purified using a MinElute PCR purification Kit (#28004, Qiagen). Next, transposed DNA fragments were amplified using Nextera DNA Flex kit (#20018704, Illumina, Inc.). To reduce GC and size bias in PCR, the appropriate number of cycles was determined as follows: qPCR side reaction was run, the additional number of cycles needed were calculated, liner Rn versus cycle was plotted and the cycle number that corresponds to 1/4 of maximum fluorescent intensity was determined. The remaining PCR reaction was run to the cycle number determined. Amplified library was purified and then quantified using KAPA library quantification kit (#07960255001, Roche, Basel, Switzerland) and Bioanalyzer (Agilent technologies). The resulting libraries were sequenced using HiSeq X Ten (Illumina, Inc.). Both library preparation and sequencing were performed by the Macrogen (Macrogen, Inc).

I-５.１.１３ Immunoblotting and co-immunoprecipitation Assays

Cells were lysed in radioimmunoprecipitation assay (RIPA) buffer supplemented with protease inhibitor. For immunoblotting, the cell lysates were separated by 10% SDS-polyacrylamide gel electrophoresis (PAGE) and transferred to nitrocellulose membranes. After blocking membranes with 5% skim milk in Tris buffered Saline containing 0.1% Tween-20 (TBS-T) for 2 hours (hr) at room temperature, the nitrocellulose membranes were incubated with appropriate primary antibodies overnight at 4°C and subsequently reacted with horseradish peroxidase (HRP)-conjugated secondary antibodies for 1 hr at room temperature. Bands were visualized using an enhanced chemiluminescence (ECL) detection system, West-Q Pico ECL Solution (W3652-02, GenDEPOT, Katy, TX, USA). For quantification of immunoblot results, the densities of target protein bands were analyzed with Image J.

For immunoprecipitation, the cell lysates (2 mg) were incubated with appropriate antibodies (1 µg) overnight at 4°C and precipitated with TrueBlot Anti-Rabbit Ig IP agarose beads (Rockland, Philadelphia, PA) for 2 hr at 4°C. The immunocomplexes were washed with chilled PBS three times and heated with 3x sample loading buffer containing ß-mercaptoethanol. The samples were separated by 6-8 % SDS-polyacrylamide gel electrophoresis (PAGE) and immunoblot was performed as described above.

The following antibodies were used for immunoblotting and co-immunoprecipitation assays: anti-TXNIP (#14715), anti-HDAC2 (#57156) and anti-alpha Tubulin (#3873) were obtained from Cell Signaling Technology (Beverly, MA); anti-H3ac (39139) was obtained from Active Motif (Carlsbad, CA); anti-ß-actin (GTX629630) was obtained from GeneTex; normal anti-rabbit IgG (sc-2027) was obtained from Santa Cruz Biotechnology (Dallas, TX); TrueBlot anti-rabbit IgG HRP (18-8816-31) was obtained from Rockland (Philadelphia, PA).

I-５.１.１４ Quantitative Reverse-transcription polymerase chain reaction (PCR)

Total RNA was isolated using TRIzol reagent (#15596018, Invitrogen, Carlsbad, CA, USA; Thermo Fisher Scientific) and subjected to reverse transcription PCR (RT-PCR) with ReverTra Ace qPCR RT kit (#FSQ-101, Toyobo, Osaka, Japan) or GoScript RT-PCR system (#A5001, Promega, Madison, WI, USA) according to the manufacturer’s instructions. The mRNA expression levels of target genes were quantified using the CFX Opus 96 (Biorad, Hercules, CA) or Applied Biosystems QuantStudio 1 (Applied Biosystems, Foster city, CA) real-time PCR. AccuPower 2X GreenStar™ qPCR Master Mix (#K6251, Bioneer, Daejeon, Republic of Korea) or SYBR Green Realtime PCR Master Mix (#QPK-201, Toyobo, Osaka, Japan) were applied according to the manufacturer’s protocols. The data normalized by GAPDH or alpha-tubulin mRNA levels and calculated using the ΔΔCt method (Hellemans, Mortier, De Paepe, Speleman, & Vandesompele, 2007). The primers used for qRT-PCR analysis are summarized in Table I-2.

Primer name	Forward/reverse	Sequence	Application
alpha-tubulin	Forward	CTGGACCGCATCTCTGTGTACT	RT-qPCR
	Reverse	GCCAAAAGGACCTGAGCGAACA	
TXNIP	Forward	GCTCCTCCCTGCTATATGGAT	
	Reverse	AGTATAAGTCGGTGGTGGCAT	
CD22	Forward	GCGCAGCTTGTAATAGTTGGTGC	
	Reverse	CACATTGGAGGCTGACCGAGTT	
L1CAM	Forward	TCGCCCTATGTCCACTACACCT	
	Reverse	ATCCACAGGGTTCTTCTCTGGG	
CD22	Forward	GCGCAGCTTGTAATAGTTGGTGC	
	Reverse	CACATTGGAGGCTGACCGAGTT-	
OTULINL	Forward	GTGTGGAGGCAGAGGTTGAT	
	Reverse	ATGCCGCCAAAATAGCTCCT	
PRR5L	Forward	GCGGCTGTTGAAGAGTGAAC	
	Reverse	AGCCAGAACCTCAATGCGAT	
SDC3	Forward	CTCCTGGACAATGCCATCGACT	
	Reverse	TGAGCAGTGTGACCAAGAAGGC	
GAPDH	Forward	ATCACCATCTTCCAGGAGCGA	
	Reverse	CCTTCTCCATGGTGGTGAAGAC	
CD22 #1	Forward	CGCTGGAGAAGTGAGTTCGG	ChIP-qPCR
	Reverse	TCCCTGCCTCCACTGATAGC	
CD22 #2	Forward	GACGCTGAGATGAGGGTTGG	
	Reverse	TGACTCAGGAGGTTGGCAGA	
CD22 #3	Forward	TCCCCACTCTTCTCGCTCTC	
	Reverse	ATTTGCGAGGTTGAGGTTGTC	
L1CAM #1	Forward	CAGCTCAGTGCCTCATGGAA	
	Reverse	GAGACTGCTTCCAGAGTGGG	
CD22 #2	Forward	GGAATGCTTCACTGGGCAAC	
	Reverse	GGGGTAAGAATTCCGGAGCC	
CD22 #3	Forward	CGTGTCTGAGAAAGGAAGCCA	
	Reverse	CGGCTTATCCCGATCTACCC	

Table I-2. List of primer sequences used in this study.

I-5.1.15 Immunofluorescence of HDAC2 and TXNIP

HeLa cells were cultured in 6-well plates with cover slips in each well (1.5 x10^4 cells/well). After cells were incubated overnight in Opti-MEM, TXNIP knockdown was induced by transfection of siRNA at a concentration of 100 nM. Following 48 hr of transfection, the cells were washed twice with PBS and then fixed with 100% ice-cold methanol for 10 min at -20˚C. After rinsing three with PBSTw (PBS containing 0.1% Tween 20), the cells were blocked with 3% BSA in PBS and incubated for 45 min at room temperature. Next, cells were incubated with the primary antibody for 150 min followed by the secondary antibody for 60 min in the dark. For co-staining with a second primary antibody, the blocking step followed by the primary and secondary antibody incubation steps were repeated. All of the antibodies were diluted in antibody dilution buffer (1% BSA in PBS). Information of the antibodies are listed in “antibody” section in STAR Method. The cover slips were rinsed three times with PBSTw and then mounted with VECTASHIELD Antifade Mounting Medium containing DAPI (Vector Laboratories, Newark, CA, USA) according to the manufacturer's instructions. The fluorescence was visualized with a Nikon C2 Si-plus confocal microscope.

I-5.1.16 Nuclear–cytoplasmic fractionation

Prior to transfection, HeLa cells were seeded in 100 mm cell culture dishes containing Opti-MEM medium and incubated overnight (reaching a confluency of approximately 30%-40%). The cells were then transfected with siTXNIP. Cells were harvested after 48 hr of transfection and fractionated according to the manufacturer’s instructions using NE-PER Nuclear and Cytoplasmic Extraction Reagents (#78833, Thermo Fisher Scientific). Protease inhibitor cocktail (P8340; Sigma Aldrich) was added as a supplement to the lysis buffer and the protein concentration was measured using a Pierce BCA Protein Assay Kit (#23225, Thermo Fisher Scientific).

I-５.１.１７ Chromatin Immunoprecipitation (ChIP) Assay

Cells were crosslinked with 1% formaldehyde at 37°C or room temperature for 15 min and the reaction was stopped by the addition of 0.125M glycine. ChIP was then performed using a ChIP-IT High Sensitivity kit (#53040, Active Motif, Carlsbad, CA, USA) according to the manufacturer's instructions. Enrichment of the ChIP signal was detected by quantitative real-time PCR (qPCR). The data of each biological replicate were normalized with negative control IgG signals and enrichment values were calculated using the ΔΔCt method (Hellemans et al., 2007). The following antibodies were used: TXNIP (14715, Cell Signaling Technology, Beverly, MA), HDAC2 (57156, Cell Signaling Technology), H3ac antibody (39139; Active Motif, Carlsbad, CA), and normal rabbit IgG antibodies were used. The primers used for ChIP-qPCR are summarized in Table I-2.

I-５.１.１８ Osteoclast differentiation and collection of lentiviruses for ARRDC5 expression

BMMs were cultured as previously described (S. Y. Kim et al., 2019). Briefly, bone marrow was obtained from mouse femurs and tibias at 8 weeks of age, and BMMs were isolated from the bone marrow using Histopaque (1077; Sigma Aldrich). BMMs were seeded at a density of 1.2 × 10^5 cells/well into 24-well culture plates and incubated in α-MEM (SH30265.01; Hyclone, Rockford, IL, USA) containing 20 ng/mL macrophage colony-stimulating factor (M-CSF) (300-25; PeproTech, Cranbury, NJ, USA). To induce osteoclast differentiation, BMMs were treated for 24 hr with lentiviral-containing medium that also contained M-CSF, after which the medium was changed to α-MEM containing 20 ng/ml M-CSF and 20 ng/ml RANKL (462-TEC; R&D Systems, Minneapolis, MN, USA). The differentiation medium was changed every 24 hr during the 5-day differentiation period.

To obtain the media containing lentivirus, HEK293 cells were cultured in DMEM containing 4.5 g/L glucose (SH30243.01; Hyclone) supplemented with 10% FBS (SH30084.03; Hyclone) and 1% penicillin-streptomycin. After seeding cells at a density of

1 × 10^5 cells/well into 6-well culture plates, the cells were incubated with lentivirus co-transfected media for 16 hr. Lentivirus co-transfected media was prepared according to the manufacturer's instructions using the CRISPR & MISSION® Lentiviral Packaging Mix (SHP002; Sigma Aldrich) and the lentiviral transfer vector, pHAGE-GFP-GFP or pHAGE-GFP-ARRDC5. After the incubation, the medium was replaced with fresh α-MEM medium supplemented with 10% FBS and 1% penicillin-streptomycin. The medium was collected twice (after 24 and 48 hr), designated as lentiviral-containing medium, and stored in a deep freezer until used to infect BMMs.

I-５.１.１９ TRAP staining and bone resorption pit assay

Osteoclast differentiation and activity were determined by TRAP staining and a bone resorption pit assay, respectively. TRAP staining was performed using a TRAP staining kit (PMC-AK04F-COS; Cosmo Bio Co., LTD., Tokyo, Japan) following the manufacturer's instructions. TRAP-positive multinucleated cells with more than three nuclei were counted under a microscope using ImageJ software (NIH, Bethesda, MD, USA). The bone resorption pit assay was performed using dentin discs (IDS AE-8050; Immunodiagnostic Systems, Tyne & Wear, UK). Cells were differentiated to osteoclasts on the discs over a 4-day period, after which the discs were stained with 1% toluidine blue solution and the resorption pit area was quantified using ImageJ software.

I-５.１.２０ Immunofluorescence staining of the V-type ATPase and visualization with GFP-ARRDC5

To inhibit V-type ATPase transport to the membrane (Matsumoto et al., 2019), osteoclasts on the fifth day of differentiation were incubated with 100 nM bafilomycin A1 (19-148; Sigma Aldrich) for 3 hr. Then, immunofluorescence staining was performed to visualize the localization of the V-type ATPase in bafilomycin A1-treated and untreated cells. The cells were fixed using a 4% paraformaldehyde solution (PC2031-100; Biosesang, Gyeonggi-do, Korea) and permeabilized using 0.05% Triton X-100 at room temperature for 5 min. The cells were incubated with anti-V-type ATPase antibody (SAB1402125-100UG; Sigma Aldrich) at room temperature for 1 hr, and then stained with the Alexa Fluor 594-conjugated anti-mouse antibody (A-21044; Invitrogen) at room temperature for 30 min. Finally, cells were mounted using Antifade Mountant with DAPI (P36962; Invitrogen). Fluorescence images were observed under a ZEISS confocal microscope (LSM5; Carl Zeiss, Jena, Germany).

I-５.１.２１ Fly husbandry

Following flies were used for this studies: Oregon R (BDSC 5), STAT::edGFP (N.Perrimon), Vdup1Mut (In this study).

I-５.１.２２ Generation of Vdup1 mutant fly

To generate Vdup1 mutant flies, we crossed female, nos-Cas9 (BDSC 54591) flies with male flies that has an expression of two different Vdup1 gRNAs (VDRC 341810). All F1 generation flies were single outed and crossed with w1118 flies. At F2 generation, flies that has a deletion of Vdup1 gene were validated by general genomic DNA PCR method and mutated flies were further validated by sanger sequencing methods.

I-５.１.２３ Immunohistochemistry

Wondering 3rd instar larvae’s lymph gland were dissected in the PBS and fixed in 3.7% formaldehyde solution. After the 30 minutes fix, samples were washed with 0.4% PBS Triton-X solution for 10 minutes, three times. Before the primary antibody incorporation, samples were blocked by 1% BSA solution for 30 minutes and following primary antibodies that targets Pxn (1:1000), Hnt (1:10), NimC1 (1:100) were used for the study. Samples with primary antibody were kept in 4’C for overnight. After the primary antibody incorporation, samples were washed with 0.4% PBS Triton-X solution for 10 minutes, three times and treated with secondary antibodies (1:250) for 3 hours. After the seconday antibody incorporation, samples were washed with 0.4% PBS Triton-X solution for 10 minutes, three times and kept in the Vectashield until the mounting on the slide glass. Samples were visualized with Nikon C2Si-plus confocal microscope and analyzed by ImageJ software.

I-５.２　Computational and statistical analysis

I-５.２.１　Database searching and analysis of mass spectrometry data

MS/MS spectra were queried using the Comet search engine (Eng, Jahan, & Hoopmann, 2013) to search for corresponding proteins in Flybase (Gramates et al., 2017) and Uniprot (The UniProt, 2017). Common contaminant protein sequences from the Common Repository of Adventitious Proteins (cRAP) Database (ftp://ftp.thegpm.org/fasta/cRAP) were used to filter contaminating sequences. Searching was done with following parameters: tryptic digest, internal decoy peptides, the number of missed cleavages=2, precursor tolerance allowing for isotope offsets=20 ppm, a 1.00 fragment bin tolerance, static modification of 57.02 on cysteine, and variable modification of 16.00 on methionine. The acetylation, phosphorylation, and ubiquitination searches add variable modifications of 42.01 on lysine, 79.97 on serine/threonine/tyrosine, and 114.04 on lysine, respectively. The search results were then processed through the Trans-Proteomic Pipeline suite of tools version 4.8.0 (Keller, Eng, Zhang, Li, & Aebersold, 2005) where the PeptideProphet tool (Keller, Nesvizhskii, Kolker, & Aebersold, 2002) was applied to calculate the probability that each search result is correct and the ProteinProphet tool (Nesvizhskii, Keller, Kolker, & Aebersold, 2003) was applied to infer protein identifications and their probabilities.

I-５.２.２　Functional annotations and multiple sequence alignment of α-arrestin sequences

The sequences of twelve *Drosophila* and six human α-arrestins were retrieved from the Uniprot database (UniProt Consortium, 2018). Domains and motifs including the PPxY motif were annotated based on sequences from Pfam version 31.0 (El-Gebali et al., 2019) and the eukaryotic linear motif (ELM) database (Dinkel et al., 2015). The sequences were subjected to the multiple-sequence alignment tool T-COFFEE (Notredame, Higgins, & Heringa, 2000) using default parameters. The output of T-COFFEE was applied to RAxML (version 8.2.11) (Stamatakis, 2014) to generate a consensus phylogenetic tree with 1,000 rapid bootstrapping using "-m PROTGAMMAWAGF" as the parameter (https://cme.h-its.org/exelixis/resource/download/NewManual.pdf).

I-５.２.３ Identification of high-confidence bait-prey PPIs

A. *SAINTexpress analysis:* To identify high-confidence bait-prey PPIs, spectral counts of AP/MS data from S2R+ and HEK293 cells were subjected to the SAINTexpress algorithm (version 3.6.1) (Teo et al., 2014), which calculates the probability of authenticity for each bait-prey PPI. The program outputs the SAINTexpress scores and the Bayesian false discovery rates (BFDR) based on the spectral count distribution of true and false PPI sets. Before calculating the scores, bait-to-bait self-interactions were removed manually. SAINTexpress was run with the "-R 2" parameter, which specifies the number of replicates, and the "-L 3" parameter, which specifies the number of representative negative control experiments to be considered.

B. *PPI validation datasets:* To evaluate the performance of the PPI prediction based on the SAINTexpress score, validation datasets including positive and negative PPIs were precompiled as described in previous studies (Y. Kwon et al., 2013; Vinayagam et al., 2016). Briefly, the positive PPIs were initially collected by searching for known PPIs involving α-arrestins from STRING version 10.5 (Szklarczyk et al., 2015), GeneMANIA version 3.4.1 (Warde-Farley et al., 2010), Bioplex (Huttlin et al., 2015), and DpiM (Guruharsha et al., 2011). For human, additional positive PPIs were curated from the literature (Colland et al., 2004; Dotimas et al., 2016; Nabhan et al., 2012; Nishinaka et al., 2004; Puca & Brou, 2014; Wu et al., 2013). After these steps, 30 PPIs (21 preys) for human and 46 PPIs (17 preys) for *Drosophila* were considered as positive PPIs. Proteins manually curated from the Contaminant Repository for Affinity Purification (CRAPome) (Mellacheruvu et al., 2013) were compared to those detected in our negative controls and only those that were detected in both were considered as were negative PPIs. As a result of these steps, 1,372 PPIs (268 preys) for human and 1,246 PPIs (122 preys) for *Drosophila* were compiled as negative PPIs.

C. *Construction of high-confidence PPI networks:* The performance of SAINTexpress was evaluated using the positive and negative PPIs. Because there is an imbalance between positive and negative PPIs, 1000 random cohorts of negative PPIs number-matched with that of positive PPIs were generated. The average true positive and false positive rates were plotted as ROC curves over different SAINTexpress scores as a cutoff, and AUC values were calculated using the ROCR R package (version 1.0-11, https://cran.r-project.org/web/packages/ROCR). Based on these results, we chose an optimal cutoff for high-confidence PPIs with a BFDR of 0.01, where the false positive rates were less than 3% (~1.8 % for human and ~2.7% for *Drosophila*) in both species, and the true positive rates were substantially higher (~66.7 % for human and ~45.7% for *Drosophila*). The cutoffs correspond to SAINTexpress scores of 0.85 and 0.88 for human and *Drosophila*, respectively.

I-５.２.４ Checking the reproducibility of spectral counts among replicates

If multiple proteins isoforms were detected, they were collapsed into a single gene. To avoid the divide-by-zero error, spectral counts of “0” were converted to a minimum non-zero value, “0.01”. To examine the integrity and quality of spectral counts from the AP/MS, the average correlation coefficients (Pearson) of spectral counts from α-arrestins were calculated and plotted. At each cutoff of spectra counts from 1 to 15, only the PPIs with spectral counts that were the same or higher than the cutoff for all replicates were kept and used to calculate correlation coefficients between replicates. The resulting coefficients from the α-arrestin interactomes were then averaged and plotted. At the cutoff of 6 spectral counts, saturation of average correlation coefficients was observed and chosen as an optimal cutoff to filter the PPIs. Principal component analysis (PCA) of the filtered PPIs was conducted based on spectral counts (with a pseudo count 1 added) transformed into a $\log_2$ using the factoextra R package (version 1.0.7).

I-５.２.５ Hierarchical clustering of high-confidence PPIs

Hierarchical clustering based on $\log_2$ spectral counts (pseudo count 1 added) of high-confidence PPIs was conducted using the Pearson correlation as the clustering distance and Ward's method as the clustering method. Heatmaps were visualized through the ComplexHeatmap R package (version 2.6.2) (Gu, Eils, & Schlesner, 2016). Six clusters were identified for each species based on the results of hierarchical clustering; the PANTHER protein class overrepresentation test was performed for the proteins in each cluster (Thomas et al., 2003). False discovery rates (FDRs, Fisher's exact test) of indicated protein classes were ≤ 0.05 for all classes except for "GTPase-activating protein" in human (FDR < 0.133) and "GEFs" in *Drosophila* (FDR < 0.109), respectively. Interacting prey proteins from the positive PPIs were selectively labeled.

I-５.２.６ Domain and motif analysis of bait and prey proteins

For human and *Drosophila*, respectively, 53 and 65 short linear motifs in α-arrestins were annotated using the ELM database (Dinkel et al., 2015), and 423 and 546 protein domains in prey proteins were annotated using the Uniprot database (UniProt Consortium, 2018). To test for enrichment of protein domains, we implemented the Expression Analysis Systematic Explorer (EASE) score (Hosack, Dennis, Sherman, Lane, & Lempicki, 2003), which is calculated by subtracting one gene within the query domain and conducting a one-sided Fisher's exact test. Protein domains enriched in the interactomes of each α-arrestin (Benjamini-Hochberg FDR ≤ 0.05) were plotted using the ComplexHeatmap R package (version 2.6.2). Next, to see how reliable our filtered PPIs were, we utilized information about known affinities between domains and short linear motifs from the ELM database (Dinkel et al., 2015). Because the arrestin_N (Pfam ID : PF00339) and arrestin_C (Pfam Id : PF02752) domains in α-arrestins do not have known interactions with any of the short linear motifs in the ELM database (Dinkel et al., 2015), only the interactions between the short linear motifs in α-arrestins and protein domains in the interactome (prey proteins) were considered in this analysis. We found that 59 out of the 390 human PPIs and 64 out of the 740 *Drosophila* PPIs were supported by such known affinities. One-sided Fisher's exact test was used to test the significance of the enrichment of the supported PPIs in the filtered PPI sets versus those in the unfiltered PPI sets (Figure I-6A).

I-５.２.７ Subcellular localizations of bait and prey proteins

To search for annotated subcellular localizations of the proteins in the α-arrestin interactomes, we first obtained annotation files of cellular components (Gene Ontology (GO) : CC) for human and *Drosophila* from the Gene Ontology Consortium (Ashburner et al., 2000). From the annotations, we only utilized GO terms for 11 subcellular localizations (name of subcellular localization – GO term ID: Cytosol – GO:0005829; Plasma membrane – GO:0005886; Nucleus – GO:0005634; Mitochondrion – GO:0005739; Endoplasmic reticulum – GO:0005783; Golgi apparatus – GO:0005794; Cytoskeleton – GO:0005856; Peroxisome – GO:0005777; Lysosome – GO:0005764; Endosome – GO:0005768; Extracellular space – GO:0005615). If a protein was annotated to be localized in multiple locations, a weighted value (1/the number of multiple localizations) was assigned to each location. Finally, the relative frequencies of the subcellular localizations associated with the interacting proteins in the filtered PPIs were plotted for each α-arrestin (Figure I-8).

I-５.２.８ Identification of protein complexes associated with α-arrestins

To examine protein complexes significantly enriched in the α-arrestin interactomes, we collected known protein complexes from two databases: COMPLEAT (Vinayagam et al., 2013), which is a comprehensive resource of protein complexes built from information in the literature and predicted by orthologous relationships of proteins across species (human, *Drosophila*, and yeast), and the DAVID GO analysis of cellular components (Huang da et al., 2009a, 2009b) (Benjamini-Hochberg FDR ≤ 0.05), from which bulk cellular compartments such as the nucleus, cytosol, and so on were excluded. From the COMPLEAT database, we evaluated the association of the resulting protein complexes with each α-arrestin by the complex association score, which is the IQM of SAINTexpress scores (Equation 1)

$$Complex\ association\ score\ (IQM) = \frac{\sum_{i=Q1}^{Q3} SAINTexpress\ score_i}{(Q3-Q1)+1} \quad \text{[Equation 1]}$$

, where the first quartile is *Q1* = $\frac{N}{4}+1$, the third quartile is *Q3* = $\frac{3N}{4}$, and *N* is the total number of preys in the complex. The significance of the complex association score was estimated by comparing the score to the null distribution of the scores calculated from 1,000 random complexes of input proteins. The significance was tested through the online COMPLEAT tool, and protein complexes with *P* < 0.05 were selected for further analysis. Next, we iteratively combined (clustered) the pairs of protein complexes from any two databases (COMPLEAT and GO analysis of cellular components) that showed the highest overlap coefficients, $Overlap(X,Y)$ (Equation 2) (Vijaymeena & Kavitha, 2016), until there was no pair of complexes whose coefficients were higher than 0.5.

$$Overlap(X,Y) = \frac{|X \cap Y|}{(|X|,|Y|)} \qquad \text{[Equation 2]}$$

From the clustered set of complexes, we manually removed those with fewer than three subunits or two PPIs. Subunits in the complexes that have no connection among themselves were also removed. Lastly, the significance of associations of the resulting complexes with each α-arrestin were tested in the same manner as done in COMPLEAT using complex association score. The resulting P values were corrected by the Benjamini-Hochberg procedure and only interactions with statistical significance (FDR < 0.05) were visualized with Cytoscape v3.5.1 (Shannon et al., 2003) (Figure I-9; Figure I-10).

I-５.２.９ Orthologous networks of α-arrestin interactomes

DIOPT (version 7.1) was used to search for orthologs of all prey proteins and only those with a DIOPT score ≥ 2 were selected for the identification of orthologous PPIs between *Drosophila* and human. Next, the orthologs were tested for the enrichment of GO biological process and molecular functions and Kyoto Encyclopedia of Genes and Genomes pathway using the DAVID (Huang da et al., 2009a, 2009b). In addition, manual curation of individual genes was performed through the Uniprot database (UniProt Consortium, 2018). The orthologs were manually grouped into functional modules based on the results and α-arrestins were modularized into seven groups based on hierarchical clustering of $\log_2$-transformed mean spectral counts using the correlation distance and the Ward linkage method. The heatmap was plotted using the pheatmap R package (version 1.0.12).

I-５.２.１０ Processing of RNA-seq data

For quality checks and read trimming, RNA-seq data were processed by FastQC (version 0.11.8) (Andrews, 2010) and sickle (version 1.33) (Joshi NA, 2011) with default parameters. After the trimming, the reads were aligned to human transcriptomes

(GENCODE version 29, GRCH38/hg38) (Frankish et al., 2019) using STAR (version 2.5.3a_modified) (Dobin et al., 2013) with default parameters and read counts were determined using RSEM (version 1.3.1) (B. Li & Dewey, 2011). The DEG analysis was performed using the edgeR R package (version 3.32.1) (Robinson et al., 2010). Batch information was added as confounding variables to adjust for batch effects.

I-５.２.１１ Processing of ATAC-seq data

Each ATAC-seq dataset was processed using the ENCODE ATAC-seq pipeline implemented with Caper (https://github.com/ENCODE-DCC/atac-seq-pipeline) (Jin Lee, 2016). Briefly, reads were mapped to the human reference genome (GRCH38/hg38) using Bowtie2 (version 2.3.4.3), and unmapped reads, duplicates, and those mapped to the mitochondrial genome were removed. Peaks were called by MACS2 (Zhang et al., 2008) and optimal peaks that were reproducible across pseudo replicates were used in the downstream analysis. The numbers of processed reads and peaks are summarized in Table I-1. Plots of ATAC-seq signals around the TSSs of expressed genes were generated by the R genomation package (version 1.22.0) (Akalin, Franke, Vlahovicek, Mason, & Schubeler, 2015). The batch effects of the signals were corrected by the removeBatchEffect function from the limma R package (version 3.46.0) (Ritchie et al., 2015). Of the broad and narrow peaks resulting from the ENCODE ATAC-seq pipeline, the latter were used as an input to obtain consensus ACRs using the diffBind R package (version 3.0.15) (Ross-Innes et al., 2012). The dACRs were detected using the edgeR R package (version 3.32.1) (Robinson et al., 2010). In total, 70,746 ACRs and 5,219 dACRs were detected in HeLa. The genomic positions of the ACRs were annotated through the ChIPseeker R package (version 1.26.2) (Yu, Wang, & He, 2015). If the ACRs spanned more than one genomic region, their positions were assigned based on the following priority: promoters ＞ 5’ untranslated regions (UTRs) ＞ 3’UTRs ＞ other exons ＞ introns ＞

downstream ＞ intergenic regions. The promoter of a gene was defined as the region 5 kb upstream and 500 bp downstream of the TSS.

I-５.２.１２ PCA of ATAC- and RNA-seq data

For ATAC-seq, normalized read counts derived from the diffBind R package (version 3.0.15) (Ross-Innes et al., 2012) were transformed into a $\log_2$ function. Batch effect corrections were done using the limma R package (version 3.46.0) (Ritchie et al., 2015). For RNA-seq, counts per million mapped reads (CPM) were also processed in the same manner. For PCA, 2,000 features with the highest variance across samples were extracted and utilized. Plots of principal components 1 and 2 were generated by the factoextra R package (version 1.0.7).

I-５.２.１３ Functional signatures of repressed genes upon TXNIP depletion

Genes that exhibited decreased chromatin accessibility at their promoter and decreased RNA expression upon TXNIP knockdown were selected based on the following criteria: 1. $\log_2$ (RNA level in siTXNIP-treated cells/RNA level in siCon-treated cells) (hereafter, siTXNIP/siCon) $\leq$-1; 2. $\log_2$ (siTXNIP/siCon) of ACRs in the promoter region $\leq$-1 (If there are multiple ACRs in the promoter region, the one with the highest ATAC-seq signal was selected) or $\log_2$ mean (siTXNIP/siCon) of all ACRs in the promoter region $\leq$-1. Enrichment analysis of the GO terms in the gene set was performed by g:Profiler (Raudvere et al., 2019). Top 10 enriched terms from the biological process and molecular functions categories were plotted (Figure I-20).

Chapter II

Hybrid transcriptome analysis of *Drosophila* larvae under immune responses

II-1 Abstract

Drosophila immune system is principally comprised of myeloid-like immune cells, known as hemocytes, and their progenitor cells, prohemocytes. Hemocytes, divided into plasmatocytes, crystal cells and lamellocytes, play a crucial role in immune defense mechanisms, such as combating wasp infestations. Previous research has cataloged cellular subtypes present during development and immune challenges in early-stage Drosophila larvae using single-cell RNA-seq data. Remarkably, the population size of lamellocytes, typically negligible under basal conditions, dramatically increase in response to wasp infestation. This increase is accompanied by a specific and pronounced expression of certain non-coding RNAs. To further investigate novel non-coding RNAs that could potentially influencing lamellocyte development, we employed both Illumina short- and Nanopore long-read sequencing to constructed integrative, hybrid transcriptomes. Our updated gene models, generated from this hybrid approach, led to the discovery of novel non-coding RNAs distinctly expressed in lamellocytes and related lineages, as inferred from single-cell RNA-seq. Currently, we are examining the functional roles of known and novel lncRNAs in the development of lamellocytes. Furthermore, we are investigating a potential global shift in alternative splicing and isoform usages in infested conditions, and we plan to analyze dynamics of expression levels of the affected genes in bulk and single-cell level. Finally, through our long-read RNA-seq data, we were able to identify fusion genes, some of which were highly prevalent across tissues and multiple time points of *Drosophila* larvae. Experimental validation and functional exploration are underway. In summary, we have devised a pipeline to construct hybrid transcriptomes, discovered novel lncRNA markers in lamellocyte populations surging in response to immune challenges and explored global alternative splicing and isoform usage and fusion genes with Nanopore long-read RNA-seq data in *Drosophila* larvae.

II- 2 Introduction

In *Drosophila* melanogaster, which is a well-established model organism, blood cells known as hemocytes play a crucial role in the immune response. Recent advances in single-cell RNA sequencing (scRNA-seq) have allowed researchers to characterize blood cell lineages and identify novel cell types and markers in Drosophila (Cattenoz et al., 2020; Cho et al., 2020; Fu, Huang, Zhang, van de Leemput, & Han, 2020; Girard et al., 2021; Leitao et al., 2020; Tattikota et al., 2020). Briefly, plasmatocytes, crystal cells, lamellocytes, adipohemocytes, primocytes, and fat body-like cells have been identified and characterized in *Drosophila* hemocytes: Plamatocytes are known to function as phagocytes, crystal cells are known to function in melanization upon wound healing process, and lamellocytes are known to be active and involved in encapsulation upon parasitic wasp infestations (Hultmark & Ando, 2022). Among them, lamellocytes, which are one of the rarest cell type under normal condition, were shown to be dramatically increased in their numbers during wasp infestation (Markus, Kurucz, Rus, & Ando, 2005; Rizki & Rizki, 1992). In the preceding research (Tattikota et al., 2020), subtypes of lamellcytes have been characterized and many known and novel lamellcytes were identified in mature lamellocyte subtypes. Among the catalogue of this marker genes, a few lncRNAs were strongly expressed in these mature lamellcytes, implying novel regulatory axis in development and differentiation of lamellocytes in immune response against parasitic wasp infestation.

Emerging evidence has demonstrated the involvement of lncRNAs in the regulation of various biological processes in Drosophila, including development, behaviour, sex determination, and dosage compensation. stress responses, and aging (K. Q. Li et al., 2019). In immune responses, however, functional roles of lncRNAs are currently limited in *Drosophila*. Based on the appearance of lncRNA markers in lamellocytes (Tattikota et al., 2020), identifying and characterization of both known and novel lncRNAs in the cell type will shed lights into non-coding RNA landscape and its contribution to immune responses in *Drosophila* larvae.

In this study, we sought to elucidate the non-coding RNA landscape in Drosophila larve during immune response against parasitic wasp infestation by *Leptopilina boulardi.* We generated both Nanopore cDNA sequencing (long-read) and Illumina short-read sequencing data from seven different conditions in *Drosophila* larvae, including tissues, which are lymph gland and circulating blood, and time points associated with immune responses in infested *Drosophila* larvae. Long-read RNA sequencing technologies, such as Nanopore sequencing have revolutionized transcriptome assembly by enabling the identification of full-length transcripts, including those with complex splicing patterns and repetitive regions (Byrne et al., 2017; Workman et al., 2019). It also reduced ambiguity in isoform identification of genes with multiple isoforms, which could lead to accurate quantification of complex transcriptome. However, long-read sequencing data also have limitations, including higher sequencing error rates and lower throughput compared to short-read sequencing (Byrne et al., 2017; Workman et al., 2019). To overcome the limitations of each data type and generate more accurate transcriptome in *Drosophila* larvae, we developed a comprehensive pipeline that combines both types of sequencing data. Accuracy of transcript structures have been validated through diverse measures and we have complied a number of novel lncRNAs and alternatively polyadenylated isoforms that have never been reported in *Drosophila*.

By leveraging the extensive gene annotation model assembled from the hybrid sequencing approach and scRNA-seq data from two previous studies (Cho et al., 2020; Tattikota et al., 2020), we have identified novel lncRNAs specifically expressed in distinct cell types. We focused on lncRNAs expressed in lamellocytes and are currently working to validate the expression and biological functions of these lncRNAs in specific cell types. Our findings will provide valuable insights into the non-coding RNA landscape in *Drosophila* immune response against parasitic wasps and their biological importance.

II-３ Results

II-３.１ Hybrid sequencing approach to decipher transcriptome of wasp infested Drosophila larvae

To investigate transcriptome landscape of lymph gland and circulating blood cells upon parasitic wasp infestation, we performed Illumina short-read RNA-seq and third generation Nanopore long-read RNA-seq on seven conditions of *Drosophila* larvae. Drosophila larvae were infested at 72 hours (hr) after egg laying (AEL) with the wasps of the species *Leptopilina boulardi* and harvested at 24- and 48-hour post infestation (hPI), which correspond to 96 and 120 hr AEL. Wild-type (WT) larvae were harvested at the same time points and at each time point of both WT and infested Drosophila larvae, circulating blood and lymph gland cells were collected, except for 120 hr AEL 48hPI in which lymph gland dissociate at 48 hPI of parasitic wasps. In total, seven samples of *Drosophila* larvae were collected and subjected to Nanopore sequencing (Figure II-1A). Sequencing reads from two different platforms were processed independently except for two steps in which sequencing errors of Nanopore reads were corrected based on kmers from Illumina RNA-seq read and exon-junction positions defined by Nanopore reads were corrected and updated based on those identified from Illumina RNA-seq reads (Figure II-1B). These steps helped in ameliorating the quality of error prone Nanopore sequencing reads. Transcriptomes assembled from multiple Nanopore sequencing samples were merged using in-house script based on following criteria: 1. For single-exon transcripts, those showing exonic overlap were merged into single, long transcript. 2. For multi-exon transcripts that share same intron structure and whose 5’end differ by same or less than 100 nucleotides (nt) or 3’end differ by same or less than 15 nt were merged into single, longest transcript. After that, transcriptomes assembled from each sequencing platform were compared and classified into tier 1 and 2 based on their structural similarity. For final transcriptome assembly, only Nanopore tier1, Nanopore tier2, and IIlumina tier1 transcriptomes were used and henceforth, they will be collectively referred to as the “hybrid transcriptome” throughout this manuscript (Figure II-1B). Number of sequencing reads and length (N50) of Nanopore cDNA reads are summarized in Figure II-2. In summary, combining long- and short-read RNA-seq data, we were able to generate a hybrid transcriptome that leverages the strengths of each sequencing method to overcome their respective limitations.

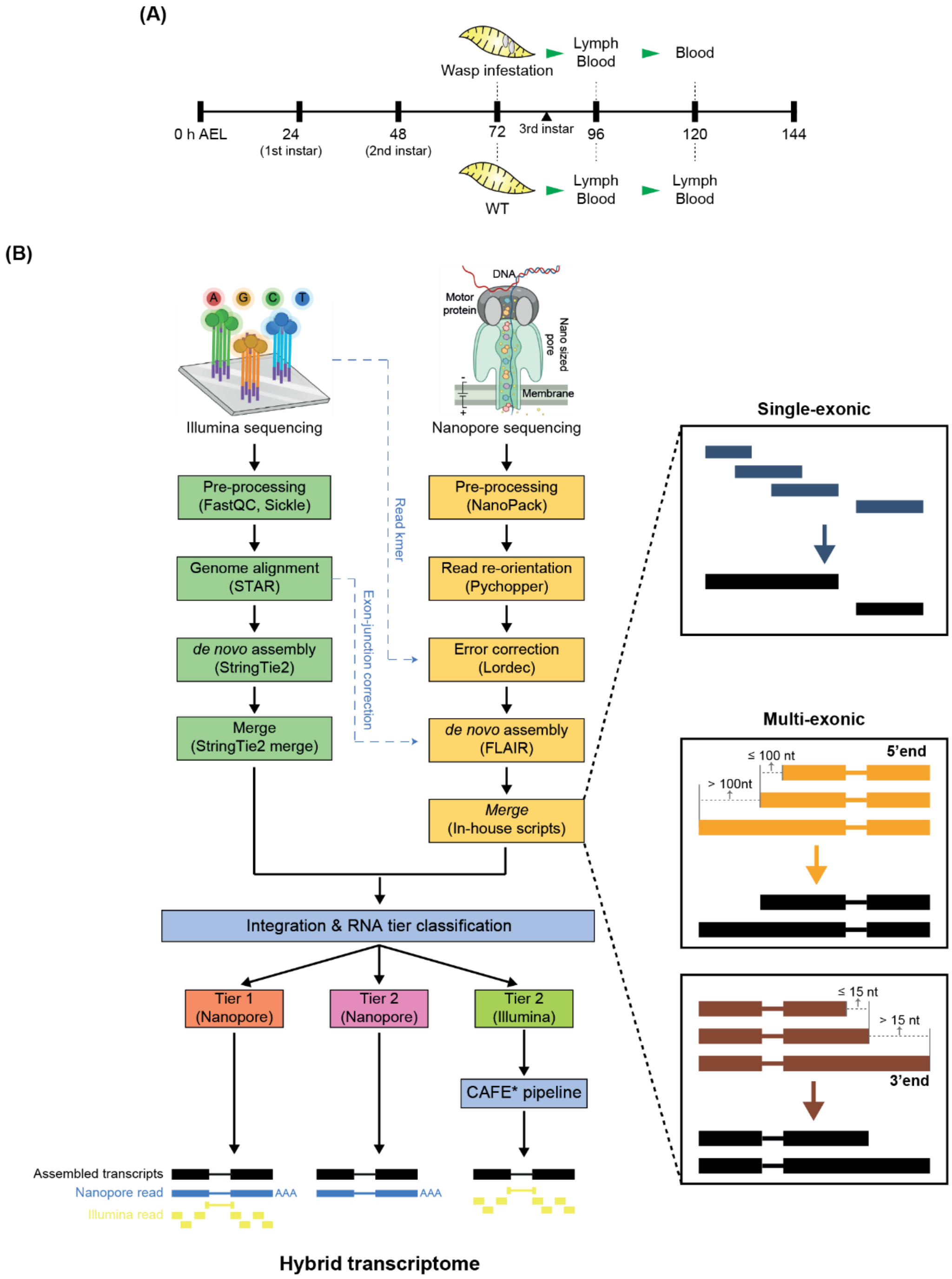
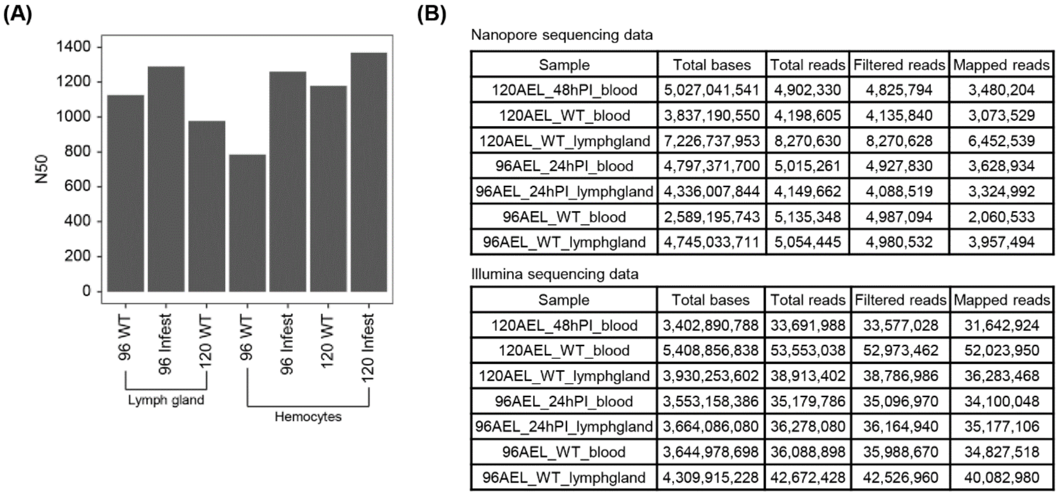


Figure II-1. Hybrid high-throughput sequencing of circulating hemocytes and lymph gland in wasp infested *Drosophila* larvae.

(A) Schematic of developmental stages of *Drosophila* larvae. Wasp infestation was induced at 72 AEL, and circulating hemocytes and lymph glands were collected at 96 after egg laying (AEL) and 120AEL of wild-type (WT) and wasp infested D*rosophila* larvae except for lymph gland at 120 hr AEL 48hPI. **(B)** Schematic of hybrid sequencing approach to construct hybrid transcriptome.



Nanopore sequencing data

Sample	Total bases	Total reads	Filtered reads	Mapped reads
120AEL_48hPI_blood	5,027,041,541	4,902,330	4,825,794	3,480,204
120AEL_WT_blood	3,837,190,550	4,198,605	4,135,840	3,073,529
120AEL_WT_lymphgland	7,226,737,953	8,270,630	8,270,628	6,452,539
96AEL_24hPI_blood	4,797,371,700	5,015,261	4,927,830	3,628,934
96AEL_24hPI_lymphgland	4,336,007,844	4,149,662	4,088,519	3,324,992
96AEL_WT_blood	2,589,195,743	5,135,348	4,987,094	2,060,533
96AEL_WT_lymphgland	4,745,033,711	5,054,445	4,980,532	3,957,494

Illumina sequencing data

Sample	Total bases	Total reads	Filtered reads	Mapped reads
120AEL_48hPI_blood	3,402,890,788	33,691,988	33,577,028	31,642,924
120AEL_WT_blood	5,408,856,838	53,553,038	52,973,462	52,023,950
120AEL_WT_lymphgland	3,930,253,602	38,913,402	38,786,986	36,283,468
96AEL_24hPI_blood	3,553,158,386	35,179,786	35,096,970	34,100,048
96AEL_24hPI_lymphgland	3,664,086,080	36,278,080	36,164,940	35,177,106
96AEL_WT_blood	3,644,978,698	36,088,898	35,988,670	34,827,518
96AEL_WT_lymphgland	4,309,915,228	42,672,428	42,526,960	40,082,980

Figure II-2. A substantial quantity of high-quality, high-throughput sequencing data were generated for the hybrid transcriptome assembly methodology

(A) N50 of Nanopore sequencing reads from each sample. **(B)** Statistics of Nanopore (top) and Illumina sequencing reads during computational processing steps. For Nanopore sequencing data, filtered reads represent those whose Phred-scaled quality score is same or above 7 and for Illumina data, filtered reads represent those trimmed by Sickle tool (Joshi NA, 2011)**.**

II-３.２ Third generation Nanopore sequencing provides superior coverage of the entire gene body compared to traditional short-read sequencing data, particularly in the 3'end regions

Next, we compared coverage of two different sequencing platforms across gene body. In all samples, Nanopore sequencing reads were more evenly distributed across entire gene body, particularly at the 3'end region (Figure II-3). Coverage of Illumina sequencing data was shown to drop rapidly in both end of gene body, exhibiting inherent sequencing bias in short-read sequencing data. To assess if Nanopore sequencing data captures authentic 3'end of poly adenylated transcripts, we analyzed base compositions at the 3'end of transcriptome in reference gene annotation, Berkeley Drosophila Genome Project (BDGP) 6.22, and those that were classified into different tiers (Figure II-4). Whether it is tier 1 or 2, transcripts assembled using Nanopore sequencing data showed very similar base compositions to those in reference gene annotation, which correspond to A-rich segment, polyadenylation signal (PAS) and U-rich motif that are typically found in 3'end of mRNA and had also been reported to be enriched in 3p-seq data (Jan, Friedman, Ruby, & Bartel, 2011). In contrast, 3'end of transcripts assembled from Illumina sequencing data exhibited relatively low enrichment of these motifs. Exon counts, exon lengths, intron lengths and transcript lengths of reference gene annotation and assembled transcriptomes are summarized in Figure II-5.

Next, the hybrid transcriptomes that consist of Nanopore tier1 (n=10,634), Illumina tier2 (n=9,173), and Nanopore tier 2 (n=7,551) transcripts (Figure II-6A) were compared against BDGP6.22 reference gene annotation to find what types of reference genes were identified through our pipeline (Figure II-6B). As expected, protein coding genes were most frequently identified followed by novel RNAs, lncRNAs and so on. The observation that the number of novel RNAs exceeds the detected lncRNAs suggest that three remains an undiscovered repertoire of RNAs, although some of these may be fragmented or artifactual

RNAs. In conclusion, Nanopore sequencing data exhibit superior performance in capturing entire gene body, particularly 3’end region of poly adenylated transcripts.

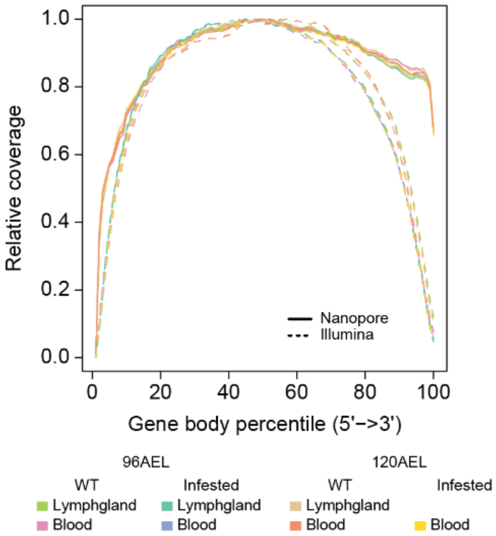


Figure II-3. Nanopore sequencing data provides a more uniform coverage across the entire gene body, especially at the 3'end.

Relative coverage of sequencing data from Illumina (dotted line) and Nanopore (line) platforms across the entire gene body (0 in x-axis indicate 5' end and 1 in x-axis indicate 3'end).

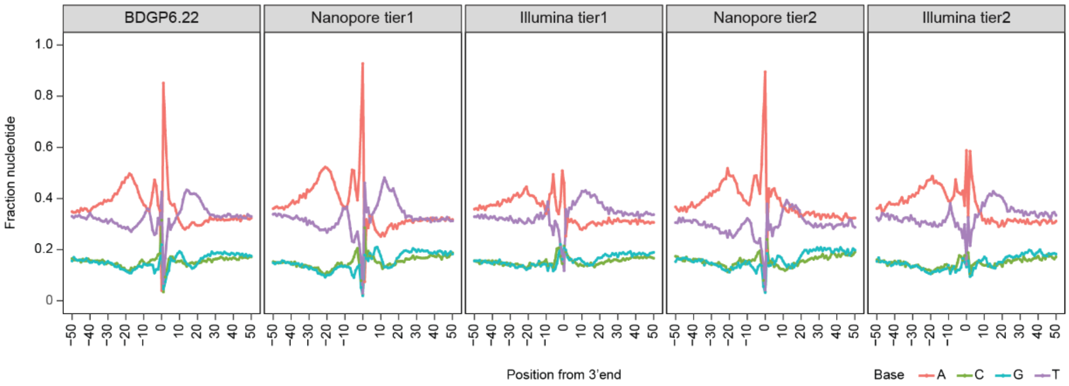


Figure II-4. Nanopore sequencing data ensures identification of accurate 3'end of poly adenylated mRNA

Nucleotide sequence composition at 3'end regions of RNAs from BDGP6.22 gene annotation and assembled transcriptome classified into different tiers based on supporting evidence from Nanopore and Illumina sequencing data.

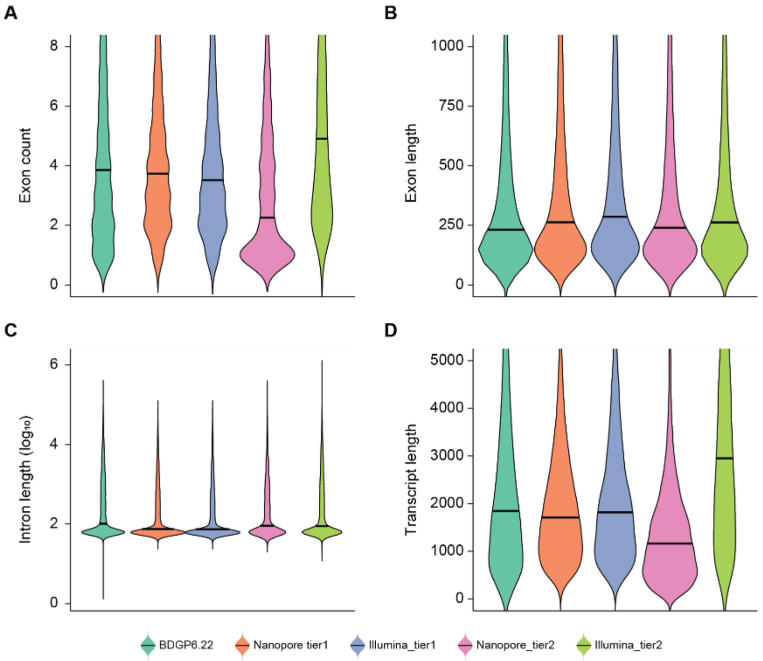


Figure II-5. Characteristics of transcriptomes assembled using short- and long-read RNA sequencing data

Distribution of exon counts **(A)**, exon length **(B)**, intron length **(C)**, and transcript length **(D)** of transcripts annotated in BDGP6.22 and assembled transcriptome classified into different tiers based on supporting evidence from Nanopore and Illumina sequencing data.

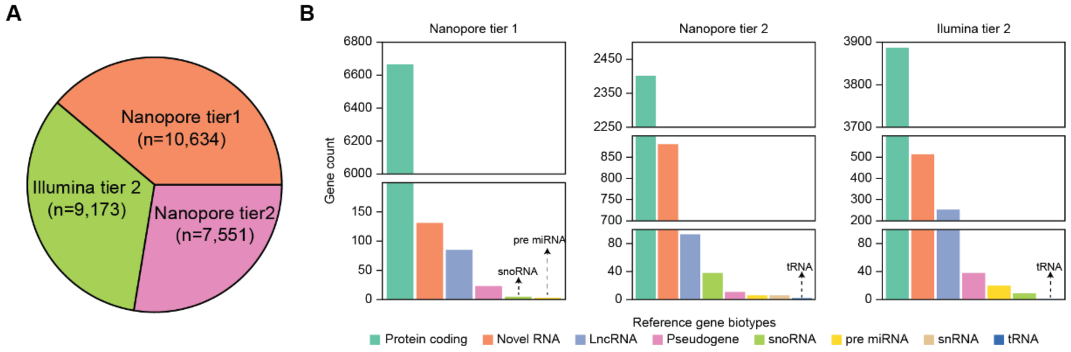


Figure II-6. Hybrid transcriptome detected genes of various biotypes

(A) Number of transcripts assembled and classified into Nanopore tier1, Nanopore tier2, and Illumina tier2. **(B)** Number of genes of various biotypes that match or overlap the transcripts from each tier.

II-３.３ Identification of novel lncRNAs and alternatively polyadenylated (APA) isoforms

To identify novel RNAs and APA isoforms expressed in lymph gland and circulating hemocytes of WT and wasp-infested *Drosophila* larvae, we developed new pipeline as shown in Figure II-7. Briefly, the hybrid transcriptome was filtered based on expression levels in Illumina and Nanopore sequencing data. The filtered transcriptome was compared to the BDPG 6.22 reference gene annotation. RNAs originating from novel loci, as well as those overlapping with known lncRNAs but displaying distinct transcript structures, were extracted. Coding potential assessment of these RNAs were performed by in silico prediction, resulting in 393 and 65 novel lncRNAs originating from novel and known lncRNA loci, respectively, along with ambiguous (One tool predicted that RNA is coding while the other predicted that RNA is non-coding) and putative protein coding RNAs. Among the assembled RNAs that were identified to be originating from known genes, those with 3' end at least 15 nt distant from those of reference transcripts and expressed above the thresholds in a minimum of one tissue under specific conditions (see "Materials and Methods" for details) were selected and defined as novel APA isoforms. The identified novel RNAs and APA isoforms were integrated with BDGP6.22 reference gene annotation for comprehensive analysis of transcriptome dynamics in *Drosophila* larvae under immune responses against parasitoid wasp eggs.

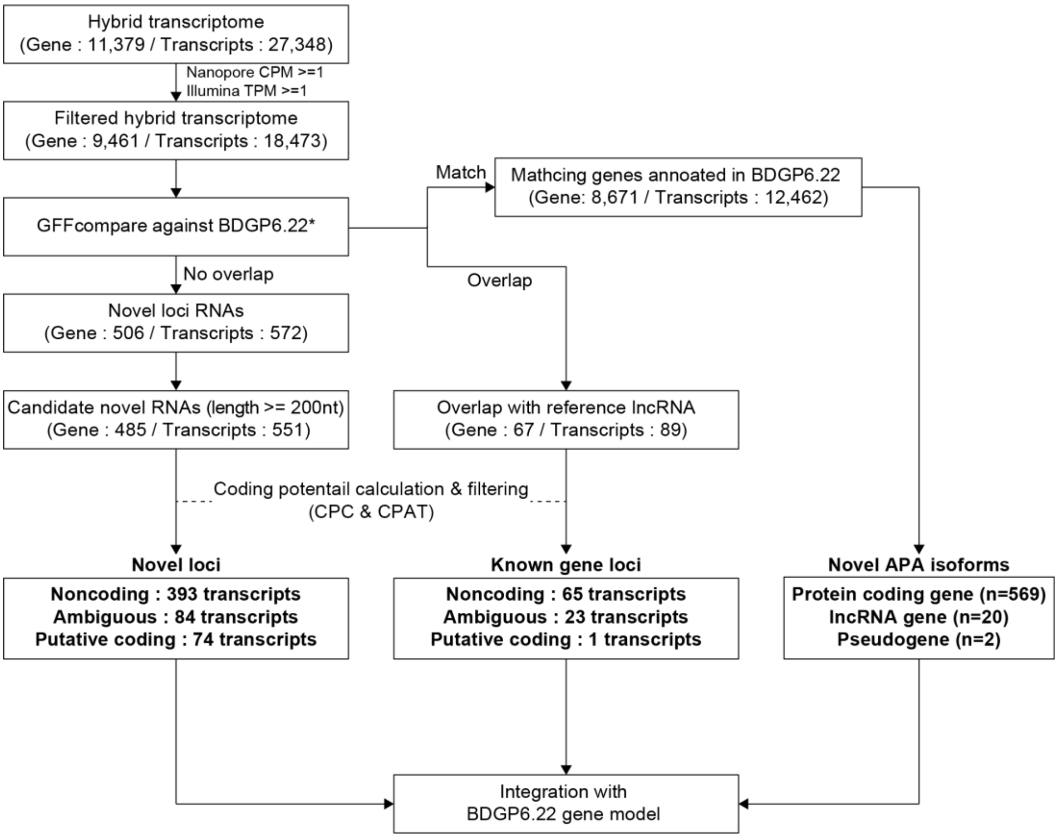

Hybrid transcriptome
(Gene : 11,379 / Transcripts : 27,348)
Nanopore CPM >=1
Illumina TPM >=1
Filtered hybrid transcriptome
(Gene : 9,461 / Transcripts : 18,473)
GFFcompare against BDGP6.22*
Match
Mathcing genes annoated in BDGP6.22
(Gene: 8,671 / Transcripts : 12,462)
Overlap
No overlap
Novel loci RNAs
(Gene : 506 / Transcripts : 572)
Candidate novel RNAs (length >= 200nt)
(Gene : 485 / Transcripts : 551)
Overlap with reference lncRNA
(Gene : 67 / Transcripts : 89)
Coding potentail calculation & filtering
(CPC & CPAT)
Novel loci
Noncoding : 393 transcripts
Ambiguous : 84 transcripts
Putative coding : 74 transcripts
Known gene loci
Noncoding : 65 transcripts
Ambiguous : 23 transcripts
Putative coding : 1 transcripts
Novel APA isoforms
Protein coding gene (n=569)
lncRNA gene (n=20)
Pseudogene (n=2)
Integration with
BDGP6.22 gene model

Figure II-7. Identification of novel RNAs and APA isoforms from hybrid transcriptome

Overview of pipeline to filter and classify transcript and identify novel RNAs and APA isoforms. Among the Nanopore tier 1 transcripts that were reported by GFFcompare to match the structure of annotated isoforms from BDGP6.22 gene annotation, expressed isoforms (transcripts that overlap with genes that are not lncRNAs: CPM ≥ 3 and isoform fraction ≥ 0.2; transcripts overlap with lncRNA genes: CPM ≥ 3 and isoform fraction ≥ 0.2) were selected. Then, those expressed transcripts whose 3'ends are at least more than 15 nt away from 3'end of the reference transcripts were selected and assigned as novel APA isoforms.

II-３.４ Short- and long-read RNA-seq data are highly reproducible and able to capture biological diversity

Based on the comprehensive gene annotation of *Drosophila* larvae, we measured gene expressions in short- and long-read sequencing data. For Nanopore sequencing data, counts per million mapped reads (CPM) were calculated using NanoCount (Gleeson et al., 2022), recently introduced tool that is specialized in estimating transcript abundances from Nanopore sequencing data using expectation-maximization (EM) approach. RNA-seq by expectation maximization (RSEM (B. Li & Dewey, 2011)), which is well known tool to estimate transcript abundance in short-read sequencing data also using EM approach, was used for Illumina sequencing data. At first, we tested if replicates of Nanopore sequencing data can be distinguished by gene expressions. Through principal component analysis (PCA), we could observe that replicates of lymph glands or circulating hemocytes under WT and wasp-infested conditions were clearly grouped according to their biological signatures (Figure II-8). For example, samples from the lymph gland and circulating blood cells of the same time point and condition (either WT or wasp infested) were found to cluster closely together in PCA plot, despite originating from different tissues. Interestingly, circulating hemocytes of 48 hPI were clustered together and most distant from all other samples, implying distinct biological signatures of cells in this specific condition.

Next, we assessed if gene expressions from two different sequencing platforms are correlated to each other (Figure II-9). In all samples, we could observe that gene expression levels (Nanopore CPM and Illumina TPM) are highly and positively correlated, Pearson correlation coefficients ranging from 0.79 to 0.97. In summary, RNA expressions estimated from Nanopore and Illumina sequencing data are highly reproducible and effective in capturing biological signatures of different tissues under different conditions.

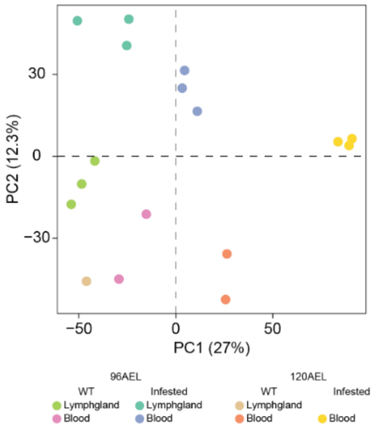


Figure II-9. Nanopore sequencing data demonstrate high reproducibility and capture variability arising from various conditions

PCA plot was generated by utilizing $\log_2$ transformed gene expressions. The plot shows principal component 1 and 2 (PC1 and PC2). Conditions in lymph gland and circulating hemocytes are colored accordingly.

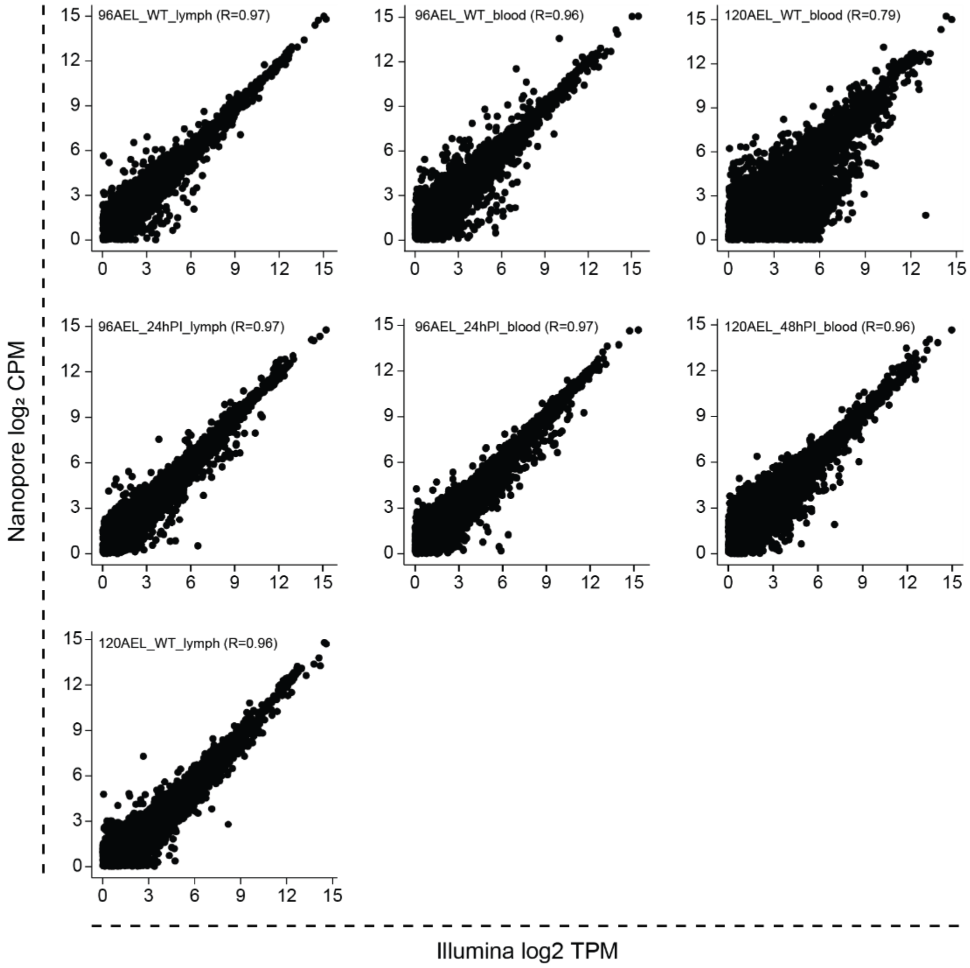


Figure II-8. Gene expressions estimated from Nanopore and Illumina platforms are highly correlated to each other in all samples

Gene expressions in nanopore ($\log_2$ transformed CPM) and illumine sequencing data ($\log_2$ transformed TPM) are projected onto y- and x-axis, respectively. On the top-left corner of each plot, name of condition and tissue in which gene expressions are estimated is written along with Pearson correlation coefficients.

II-３.５ Novel lncRNAs exhibit low coding-potential, conservation and expression levels compared to protein coding genes

To assess whether novel lncRNAs are indeed likely to be non-coding, we applied three complementary approaches to test their coding capabilities. By Comparing the CPC scores (Kong et al., 2007) and CPAT probabilities (L. Wang et al., 2013), both of which estimate coding potential of transcript using in silico prediction, we were able to conclude that novel lncRNAs have a very low coding potential, some of them even lower than previously annotated lncRNAs (Figure II-10A and B). In addition, using BLASTx, we tested if any predicted open reading frame (ORF) in transcripts could correspond to known protein or protein domain sequences annotated in *Drosophila* proteome (DROME). Consequently, both known and novel lncRNAs display a minimal likelihood of their predicted ORFs aligning with known protein sequences, which further substantiate the low coding capabilities of these newly discovered novel lncRNAs (Figure II-10C). In addition to assessing their coding potential, we measured evolutionary conservation of transcript sequences using PhastCons across 27 insect species (Figure II-10D). Conservation scores of novel lncRNAs were comparable to those of known lncRNAs and lower than those protein-coding genes, indicating more rapid evolution of lncRNAs compared to protein coding genes as previously reported(Ulitsky & Bartel, 2013).

Next, we explored the expression levels of protein coding, known and novel lncRNAs across lymph gland and circulating hemocytes of WT and wasp infested conditions (Figure II-11). As expected, expression levels of lncRNAs were generally lower compared to protein coding genes in both sequencing platforms. Expression levels of novel lncRNAs were comparable to those of known lncRNAs in most samples except for 120 hr AEL circulating hemocytes under WT condition of Illumina sequencing data. Lastly, expression specificity of lncRNAs was evaluated using Tau specificity index (Yanai et al., 2005), which serves as a measure of the extent to which a particular gene is expressed in a specific condition. In both sets of sequencing data, known and novel lncRNAs exhibited a higher likelihood of specific expression patterns compared to protein coding genes (Figure II-12). This finding aligns well with previous research, which has reported that lncRNA expression is typically more variable between tissues (Cabili et al., 2011; Derrien et al., 2012; Pauli et al., 2012). To summarize, novel lncRNAs identified in *Drosophila* larvae under WT and wasp infested conditions exhibit characteristics of those of well-established non-coding RNAs, confirming the validity of lncRNA discovery and analysis pipeline utilized in this study.

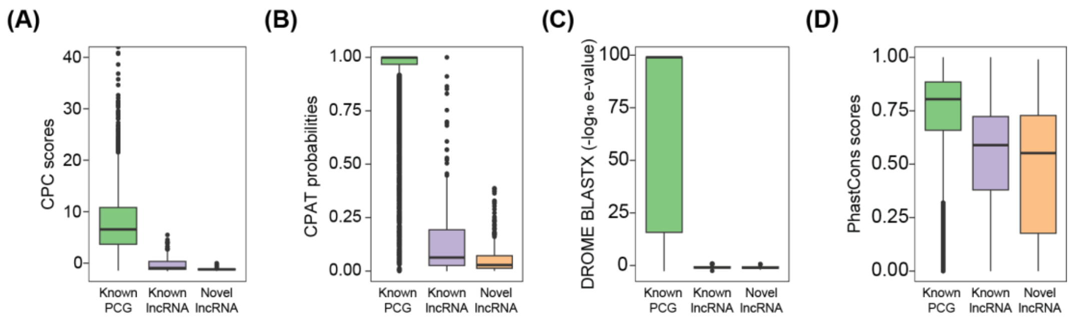


Figure II-12. Novel lncRNAs exhibit relatively low coding potential and conservation compared to known protein coding genes

Boxplot showing CPC scores **(A)**, CPAT probabilities **(B)**, e-values of BLASTX run against *Drosophila* proteome (DROME) **(C)**, and PhastCons scores **(D)**. For the analysis, single longest isoform was selected per gene and analyzed. PhastCons scores were derived from UCSC genome browser, which contain measurements of evolutionary conservation using PhastCons for 27 insect species including *Drosophila* melanogaster.

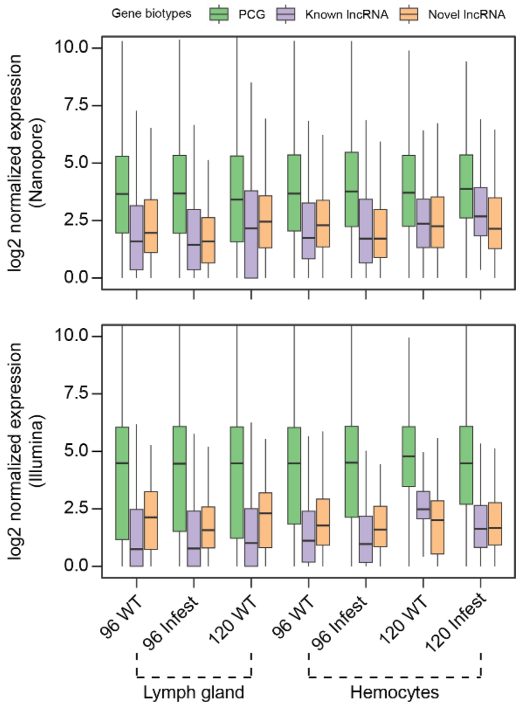


Figure II-11. Gene expression levels of protein-coding genes (PCGs) and known lncRNAs annotated in BDGP6.22 gene annotation, and novel lncRNAs

Read counts from Nanopore and Illumine sequencing data were normalized by the method implemented in edgeR and transformed into $\log_2$ scale.

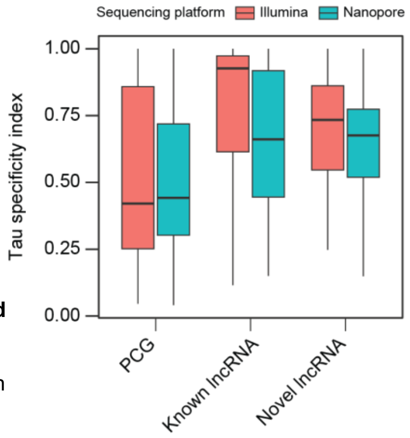


Figure II-10. lncRNAs exhibit a higher degree of condition- and tissue-specific expression compared to PCGs

Tau specificity index was calculated for each gene of PCG, known lncRNA, and novel lncRNAs.

II-３.６ Identification of differentially expressed known and novel lncRNAs

To investigate lncRNAs of functional importance in immune response against parasitic wasp infestation or developments in *Drosophila* larvae, we searched for lncRNAs that are differentially expressed between conditions. Expression levels of lncRNAs were compared between 1. WT and infested condition of same tissue at same time point and 2. different time points of same tissues under same condition (WT or wasp infested). Then, only the lncRNAs that were observed to be significantly and differentially expressed in both sequencing data were selected for further analysis (Figure II-13). Majority of these lncRNAs were already known ones but a few lncRNAs exhibit dynamic expression across different conditions. Interestingly, most of lncRNAs selected were the ones that are highly expressed in 120 hr AEL circulating hemocytes under wasp infestation, implying their functional importance in immune response against parasitic wasp. It might be the case that these lncRNAs contribute to suddent surge in lamellocyte levels within circulating hemocytes during wasp infestation (Markus et al., 2005; Rizki & Rizki, 1992). We are in a progress to experimentally validate their expression in *Drosophila* larvae.

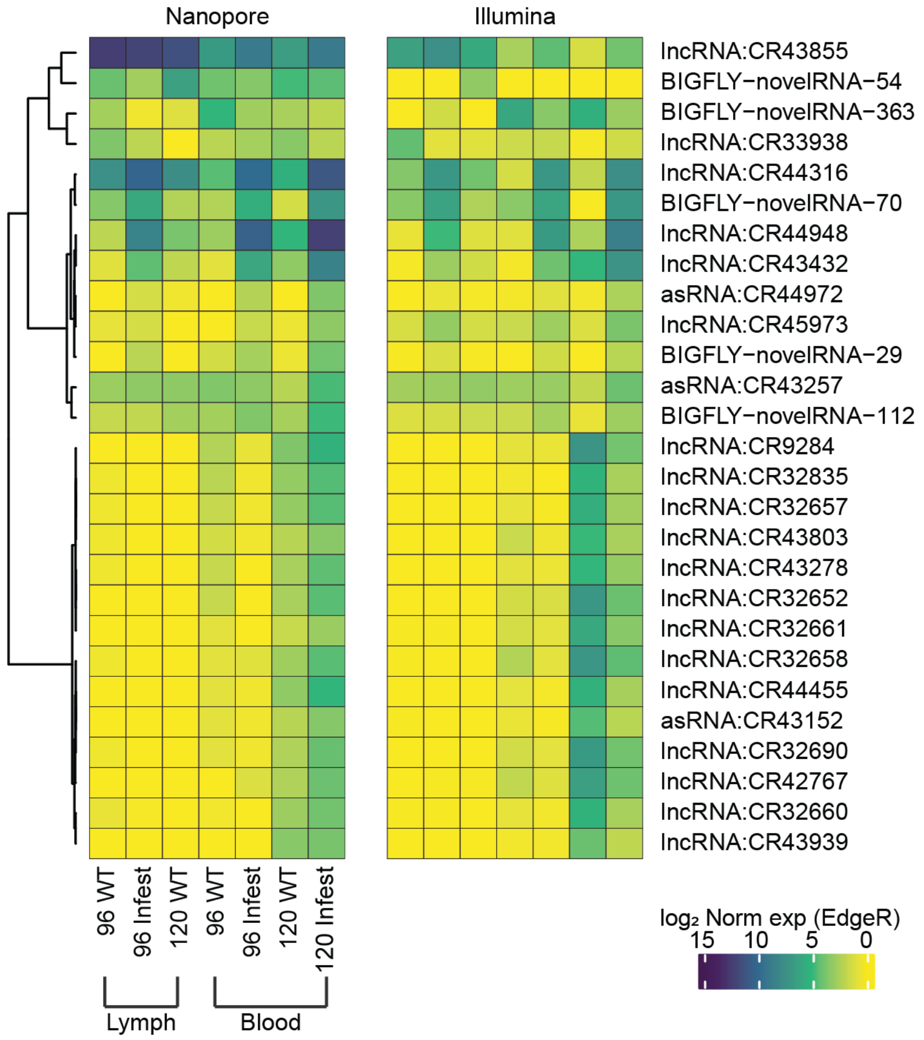


Figure II-13. List of known and novel lncRNAs that are consistently and differentially expressed between different condition of *Drosophila* larvae in nanopore and illumine sequencing data

In the heatmap, lncRNAs that exhibit concordant expression alternations (either up- or down-regulated) identified through both nanopore and illumina sequencing data are shown. Expression values are same as in Figure II-11.

II-３.７ scRNA-seq analysis of Drosophila larvae using comprehensive gene annotation

To explore the transcriptome landscape, particularly focusing on lncRNAs across various cell types at the single-cell level, we analyzed Drop-seq data from published studies, which examined and built consensus of hemocytes in lymph gland and circulating blood cells of Drosophila larvae (Cho et al., 2020; Tattikota et al., 2020), using our extensive gene annotation. Cell types extensively annotated in a preprint work(Sang-Ho Yoon, 2023) were used for cell type assignment and only cells whose mitochondrial contents are lower than 10% were utilized for further analysis, resulting in 36,007 cells (Figure II-14). Expression levels of single-cell RNA-seq (scRNA-seq) data were highly and positively correlated with Illumina and Nanopore sequencing data in all conditions (Figure II-15). As Drop-seq enables single-cell transcriptome profiling through capture 3’end of RNA molecules, we evaluated whether our extensive gene annotation, which includes novel RNAs and novel 3’ ends of newly discovered APA isoforms (Figure II-16A) contributed to increase in correct assignment of scRNA-seq reads to genes. The number of unique molecular identifiers (UMIs) assigned to each was generally higher when scRNA-seq data were analyzed using the extensive gene annotation, particularly in those genes with newly discovered APA isoforms (Figure II-16B). This result demonstrates the accurate annotation of novel 3’end using the hybrid transcriptome approach.

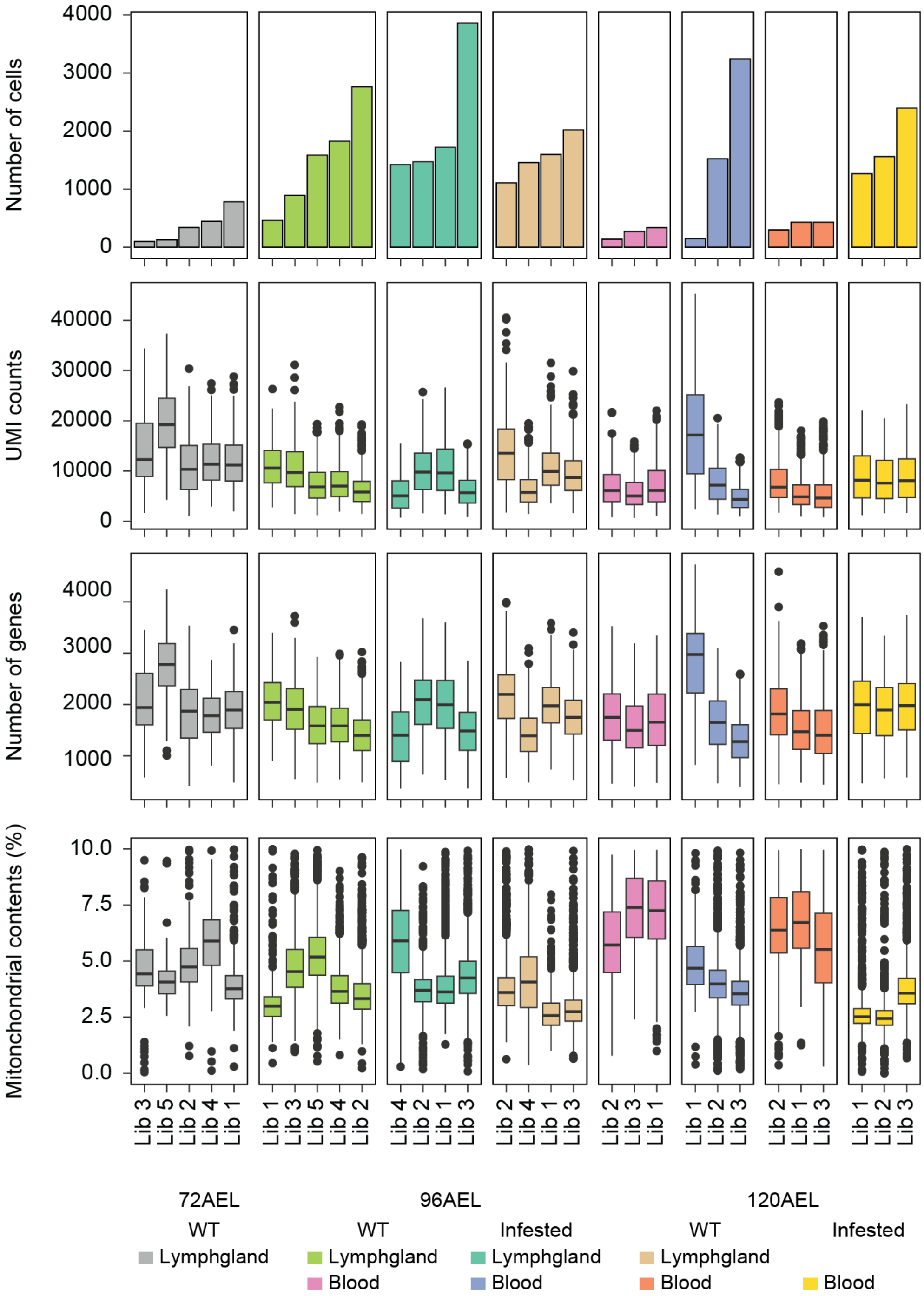


Figure II-14. Quality assessment of scRNA-seq data of *Drosophila* larvae

From top to bottom, number of cells, UMI counts, number of genes and mitochondrial contents of each library of 8 different conditions from *Drosophila* larvae. Data shown here are the ones that have gone through pre-processing and basic filtering steps.

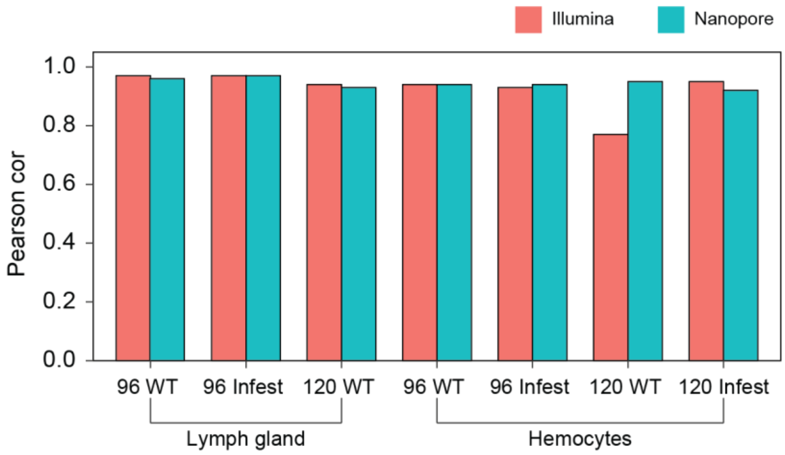


Figure II-15. Single-cell RNA-seq data are highly correlated with bulk sequencing data

Pearson correlation coefficients of pseudo bulk, gene expression levels of single-cell RNA-seq data (CPM) with those of illumina (TPM, red) and nanopore (CPM, blue) sequencing data.

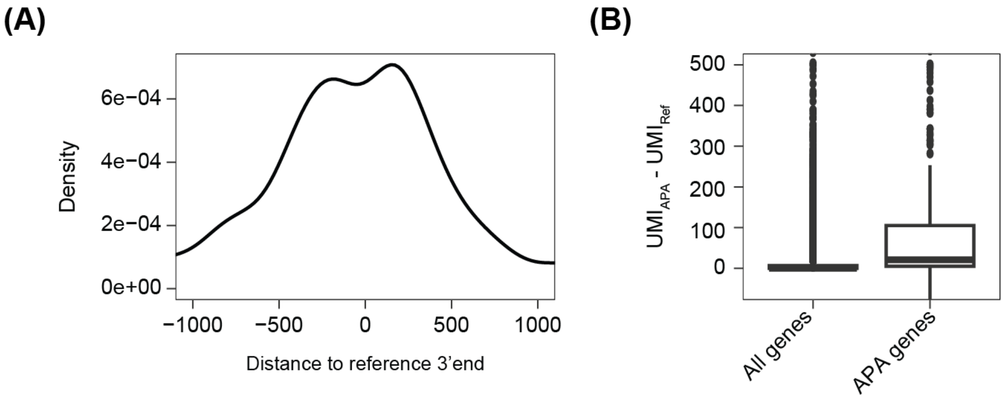


Figure II-16. Identification of alternatively poly-adenylated isoforms led to increase in UMI coverage

(A) Density plot depicting relative distance of 3'ends of APA to those of reference genes. **(B)** Relative UMI difference between the data analyzed based on BDGP6.22 gene annotation and the one with APA isoforms and novel RNAs added. APA genes indicate the genes with newly identified APA isoforms.

II-３.８ Recapitulation of major cell types using comprehensive gene annotation in single cell level

After filtering and validating the increase in UMI coverage across genes using extensive gene annotation, we projected cells into two-dimensional t-distributed stochastic neighbor embedding (t-SNE) plots to assess if the major cell types in both lymph gland and circulating blood of *Drosophila* larvae can be recapitulated (Figure II-17). In the first t-SNE plot, cells were labeled by major cell types, which consist of adipohemocyte, crystal cell (CC), GST-rich, lamellocyte (LM), prohemocyte (PH), plasmatocyte (PM), and posterior signaling center (PSC), that were comprehensively annotated in the preprinted work(Sang-Ho Yoon, 2023) (Figure II-17A). We could observe that cells were grouped according to their cell types in t-SNE projections, which suggest that single cell transcriptome profiling based on the extensive gene annotation recapitulate biological difference between different cell types as in previous research (Cho et al., 2020; Tattikota et al., 2020). Among the cell types, burst of lamellocyte population at 48 hPI in circulating hemocytes could be observed as previously reported (Lanot, Zachary, Holder, & Meister, 2001) (Figure II-17B). In addition, expression patterns of known markers corresponded well with distribution of cell types (Figure II-18). For example, prophenoloxidase 1 (PPO1), which encodes a protein produced by crystal cells and is known to be involved in the melanization reaction, was specifically expressed in crystal cells. Well known marker of lamellocyte, atilla (Evans, Liu, & Banerjee, 2014), was also shown to be specifically expressed in lamellocytes. Expression pattern of other marker genes such as Ance, Hml, and NimC1 also correspond well with distribution of cells in t-SNE projections. Number and composition of different cell types in lymph gland and circulating hemocytes under WT and wasp infested conditions are summarized in Figure II-19. In summary, single-cell analysis using our extensive gene annotation recapitulated major cell types in *Drosophila* larvae.

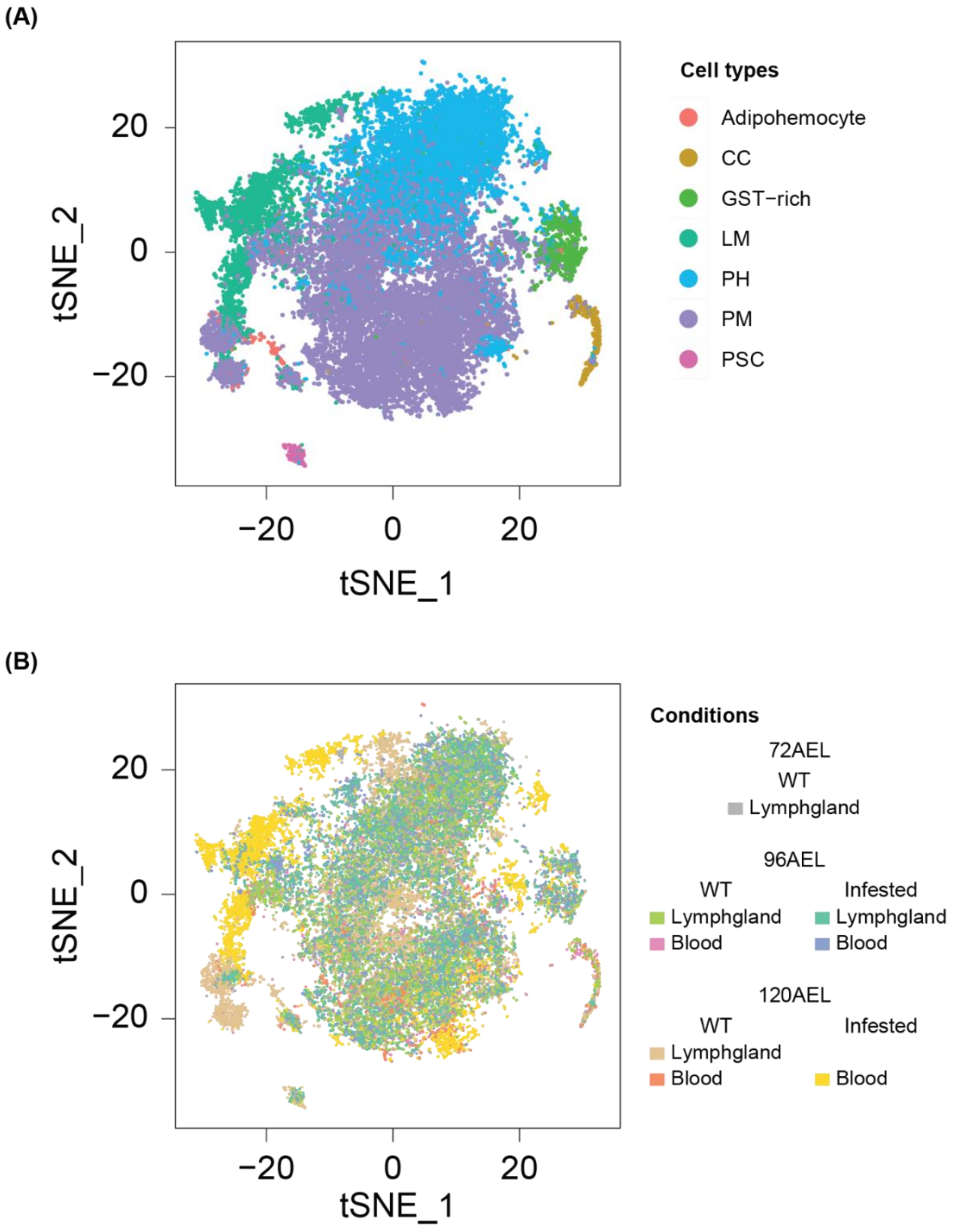


Figure II-17. Various cell types capture in single-cell RNA-seq of multiple conditions of *Drosophila* larvae

tSNE projections of cells captured in single-cell RNA-seq were labeled by cell types (A) and conditions (B).

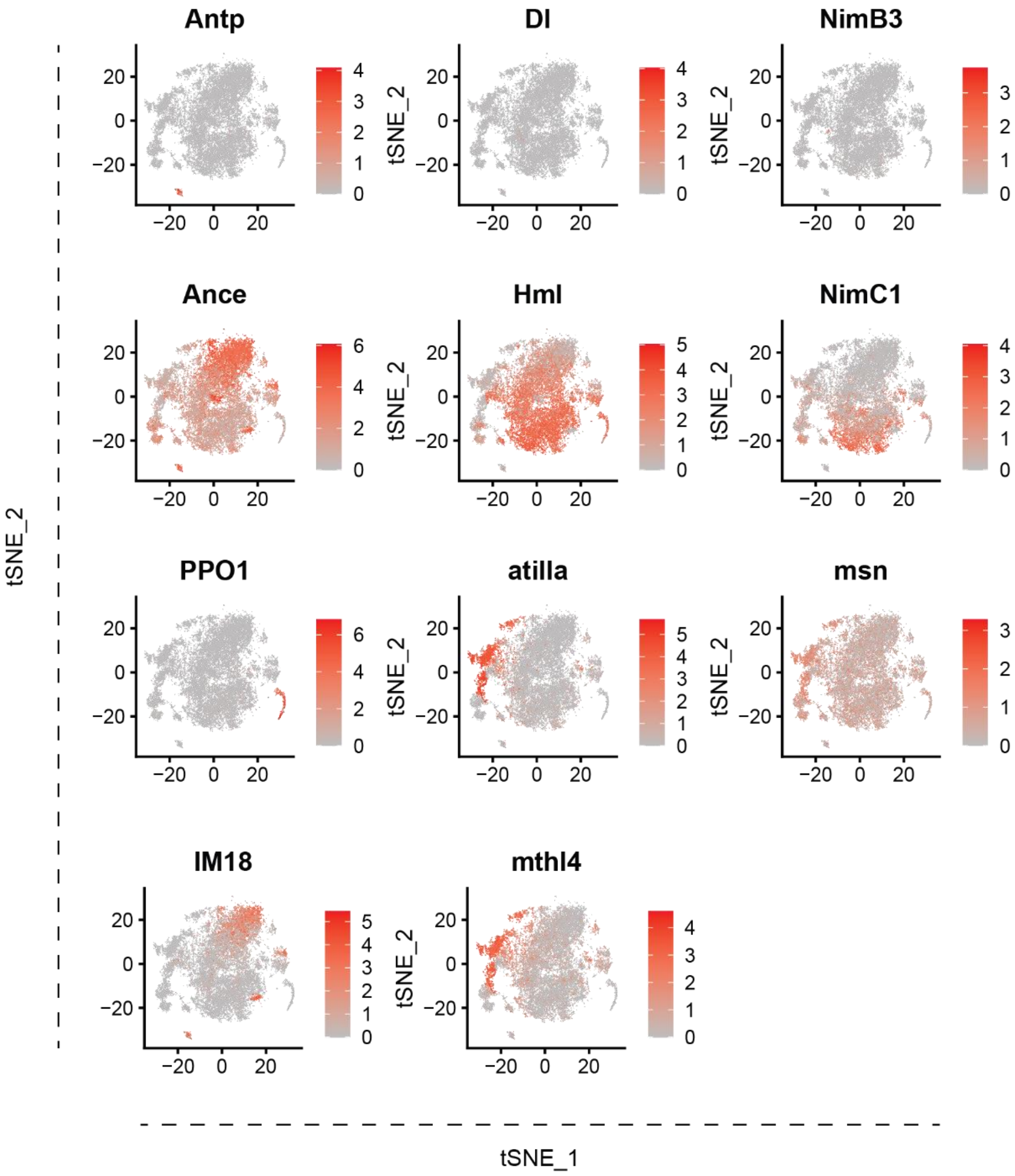


Figure II-18. Expression levels of known marker genes in the major cell types

Color bars on the right indicate level of scaled gene expression. Each marker gene is known to highly expressed in the following cell types (marker gene : cell type). Antp : PSC. Dl : PH. NimB3 : PH. IM18 : PH. Hml : PM and CC. Ance : PH. NimC1 : PM. PPO1 : CC. msn : LM. atilla : LM.

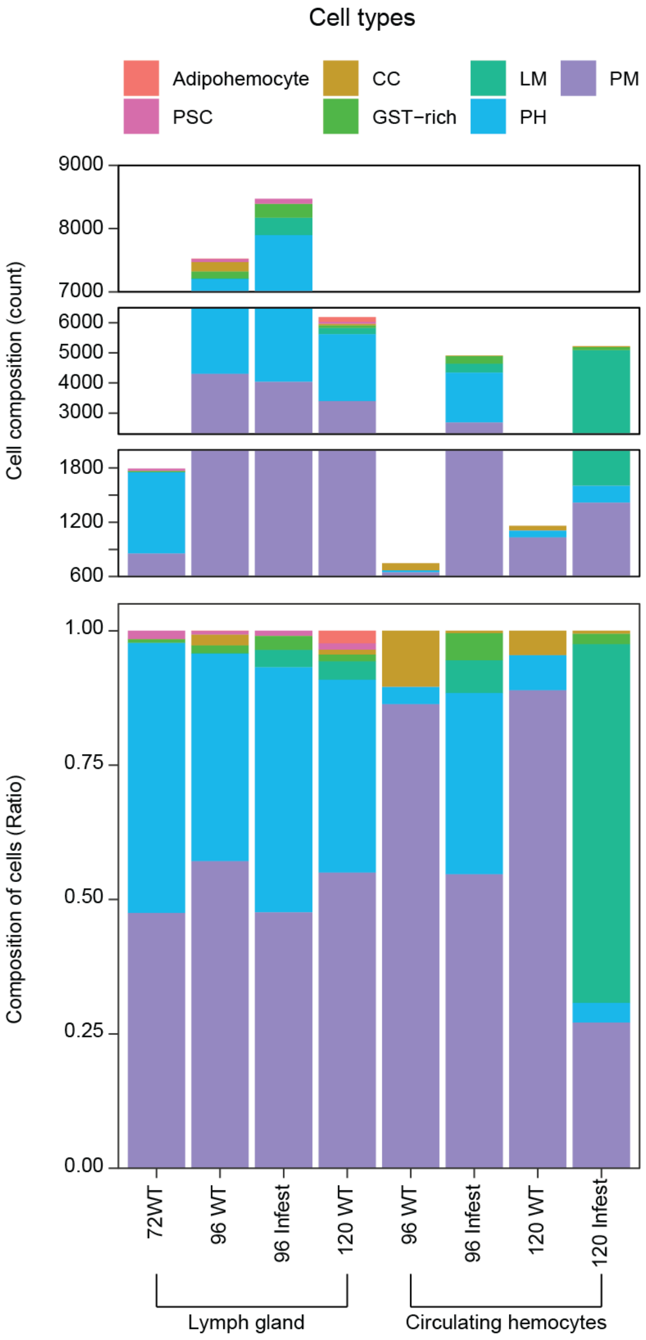


Figure II-19. Composition of cell types in each condition of *Drosophila* larvae

Number of (top) and ratio of (bottom) multiple cell types in single-cell RNA-seq of each condition of *Drosophila* larvae.

II-３.９ Discovery of known and novel lncRNA markers expressed in cell type specific manner

We next explored cell type specific marker genes, particularly lncRNAs, in scRNA-seq data using the extensive gene annotation. In addition to identification of protein coding marker genes that have been previously reported (Markus et al., 2005; Rizki & Rizki, 1992), we were able to identify substantial number of lncRNA markers expressed in cell type specific manner (Figure II-20). Among these lncRNA markers, previously reported lncRNA markers such as “lncRNA:CR43432” and “lncRNA:CR44948” were also detected in our analysis. Notably, lamellocytes exhibited strong and highly specific expression of many lncRNAs, suggesting functional roles of these lncRNAs in development and differentiation of lamellocytes during immune responses against parasitic wasp infestation.

In addition to exploring differentially expressed marker genes between cell types, we searched for lncRNA markers that are differentially expressed along the trajectory of specific cell lineage. Through trajectory analysis of 72 hr AEL WT lymph gland cells and circulating blood cells under WT and wasp infested conditions using Monocle 3 (Cao et al., 2019), we identified three main trajectories originating from PH cell types to GST-rich, PM, and LM cell types (Figure II-21). As we observed strong expression of novel and known lncRNA genes in lamellocyte, we anticipated that there could be more non-coding RNAs that play important roles in differentiation and development of lamellcytes. Therefore, we focused on cell type lineage involving lamellocytes (Figure II-22A) and identified 12 lncRNA genes whose expressions are correlated with pseudo time (Figure II-22B). In addition to the lncRNAs marker already identified in Figure II-20, we were able to discover additional lncRNA markers whose expression levels change dynamically along the lineage. For example, “lncRNA:CR43855” is highly expressed at PH and its expression gradually decrease as the cells differentiate into lamellocytes. We are currently working to validate expression of these lncRNA markers in specific cell types and investigate on their functions in cell type differentiation and development, especially in cell types that exert important functions in immune response such as lamellocyte.

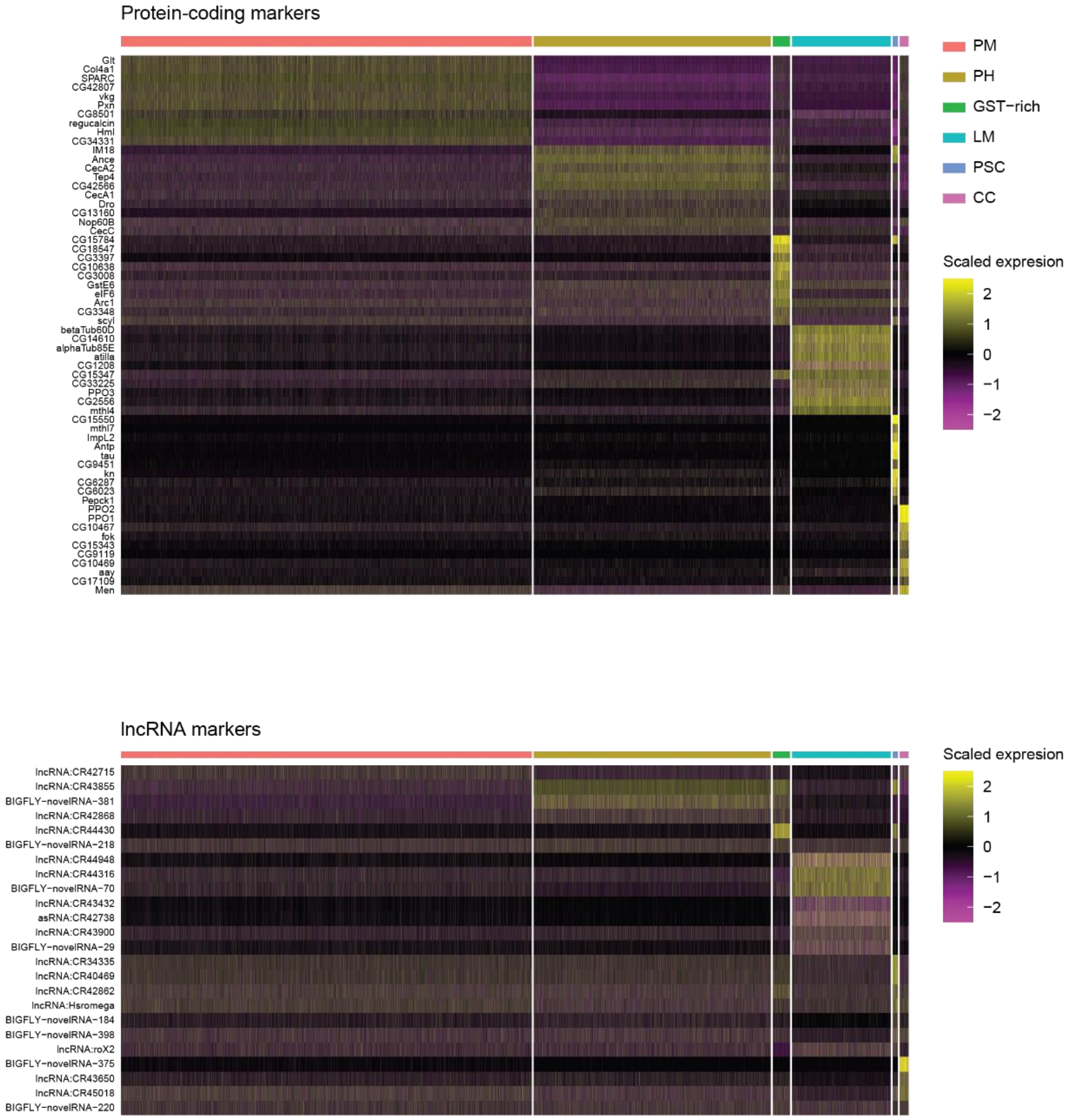


Figure II-20. Protein-coding and lncRNA gene markers highly expressed in different cell types

Scaled expression levels of top 10 highly expressed protein-coding (top) and all lncRNA gene markers (bottom) in each cell type are plotted as heatmap.

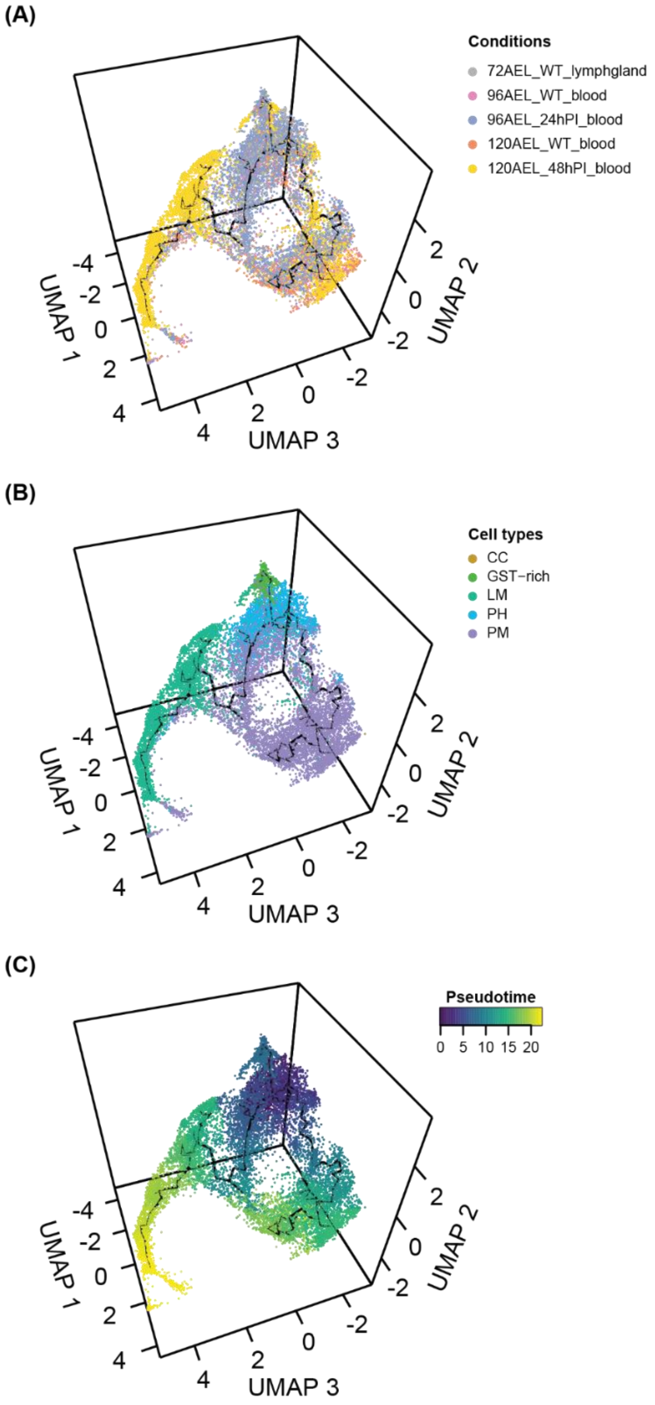


Figure II-21. Trajectory analysis of hemocyte cells reveal three main trajectories starting from PH cells to GST-rich, PH, and LM cells

72AEL WT lymph gland cells and hemocytes of WT and infested conditions were utilized for trajectory analysis using monocle3. Three dimensional UMAP projections of cells are labeled by conditions **(A)**, cell types **(B)**, and pseudotime **(C)**.

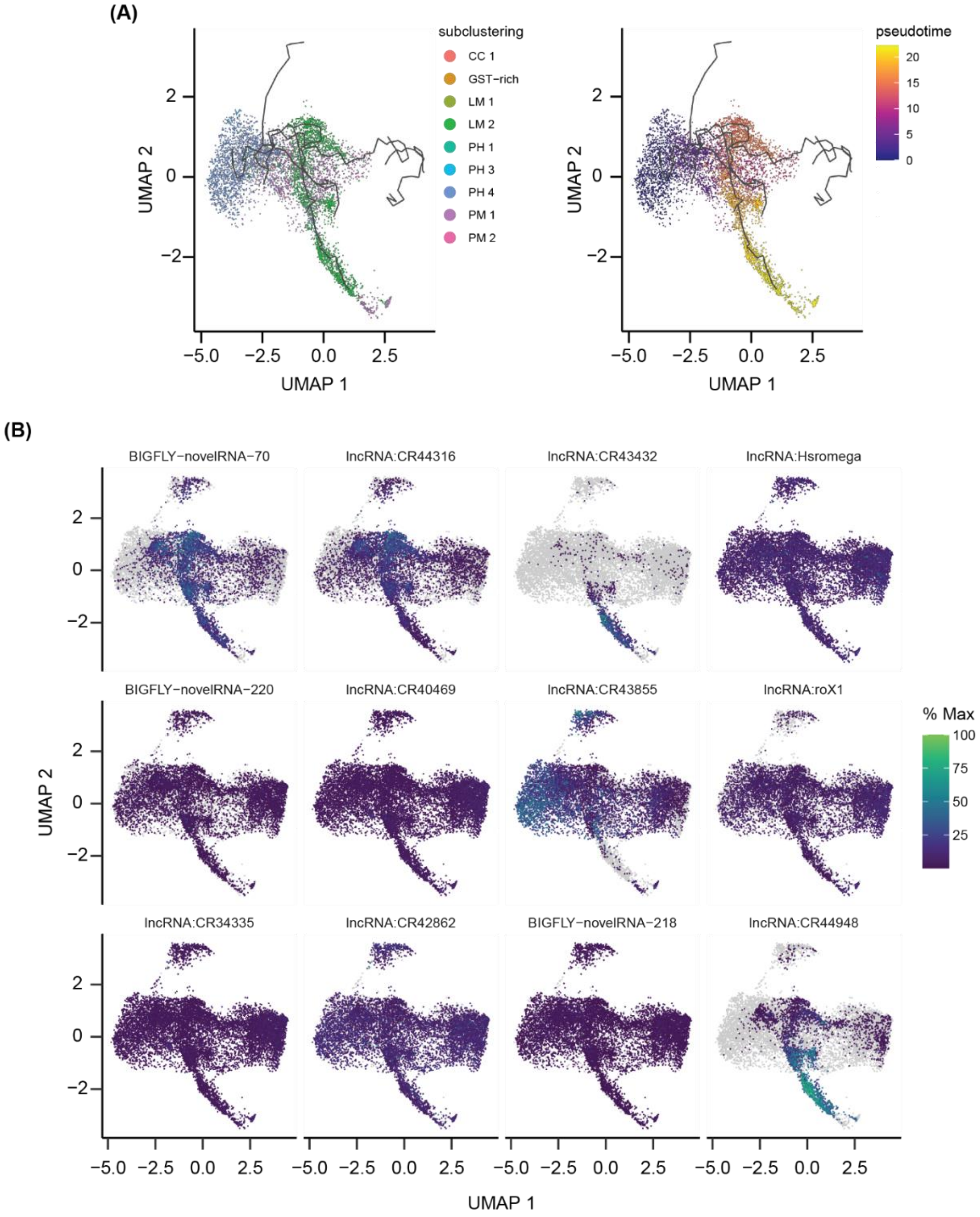


Figure II-22. Identification of known and novel lncRNAs that are dynamically expressed in the lineage of hemocytes from PH to LM

(A) Subset lineage that starts from PH to LM is depicted and cells are labeled by subcluster (left) and pseudotime (right). (B) Scale expressions (0 to 100 %) of known and novel lncRNAs whose expression levels are significantly correlated with pseudotime are plotted along the subset lineage that starts from PH to LM.

II-３.１０ Isoform switching and alternative splicing actively occur in circulating hemocytes under wasp infestation conditions

In addition to discovery of lncRNA markers in lamellocyte populations, we explored global isoform switching and associated RNA alternative splicing (AS) events between different conditions of *Drosophila* larvae. Firstly, we searched for isoform switching and associated AS events between wild-type and wasp infested conditions of both lymph gland and circulating hemocytes (Figure II-23). Around 20 isoform switching events passing the statistical significance (FDR ≤ 0.05 and Δ isoform fraction ≥ 0.1) were detected and these events were observed to take place uniquely according to their origin of tissues and time points in Drosophila larvae. Next, we explored differential switching events between different time points of circulating hemocytes (Figure II-24). Compared to circulating hemocytes under normal conditions, those under wasp infestation conditions exhibited active isoform switching between 96 and 120 hr AEL (Figure II-24A). Majority of these isoform switching events were associated with differential usage of alternative transcription start site, implying isoform switching events arising from alternative selection of first exon in circulating hemocytes under wasp infestation conditions (Figure II-24B). Among these switching events, we discovered some genes were affected in their protein sequences and domains (Figure II-25). For example, in HisRS gene, which is predicted to be involved in histidyl-tRNA-aminoacylation and mitochondrial translation, transcript “FBtr0333804” was shown to be highly expressed in 96hr AEL 24hPI lymph gland compared to its wild-type counterpart, leading to loss of WHEP-TRS protein domain (Figure II-25A). in gish gene, which encodes a plasma membrane-associated kinase that regulates Hedgehog and Wingless signaling activity, transcript “FBtr0100331” was shown to be highly expressed in 120hr AEL 48hPI blood compared to 96hr AEL 24hPI blood, leading to loss of CK1gamma_C protein domain. Functional consequences of these isoform switching events need further examination.

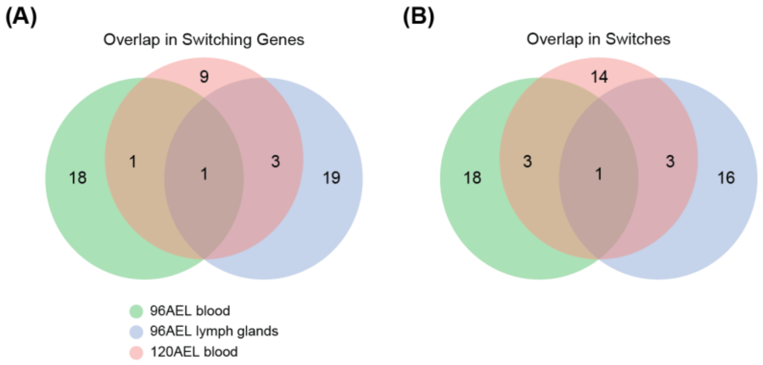


Figure II-24. Overlap of isoform switching events and associated genes between wild-type and wasp infestation conditions

Overlap of isoform switching events **(A)** and associated genes **(B)** between normal and wasp infestation conditions of lymph gland and circulating hemocytes at 96 and 120 hr AEL.

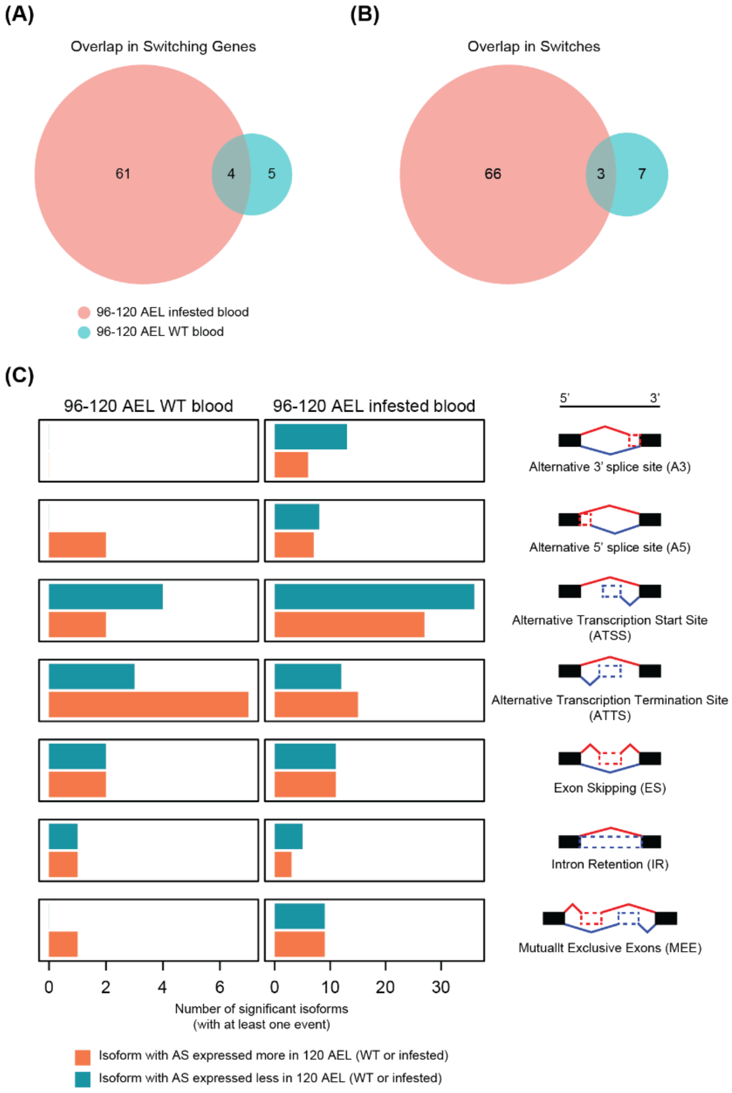


Figure II-23. Isoforms switching events more prevalent in circulating hemocytes under wasp infestation conditions

Overlap of isoform switching events between 96 and 120hr AEL of circulating hemocytes under normal and wasp infestation conditions **(A)** and associated genes **(B)**. **(C)** Number of isoforms switching events associated with alternative splicing types specified on the right.

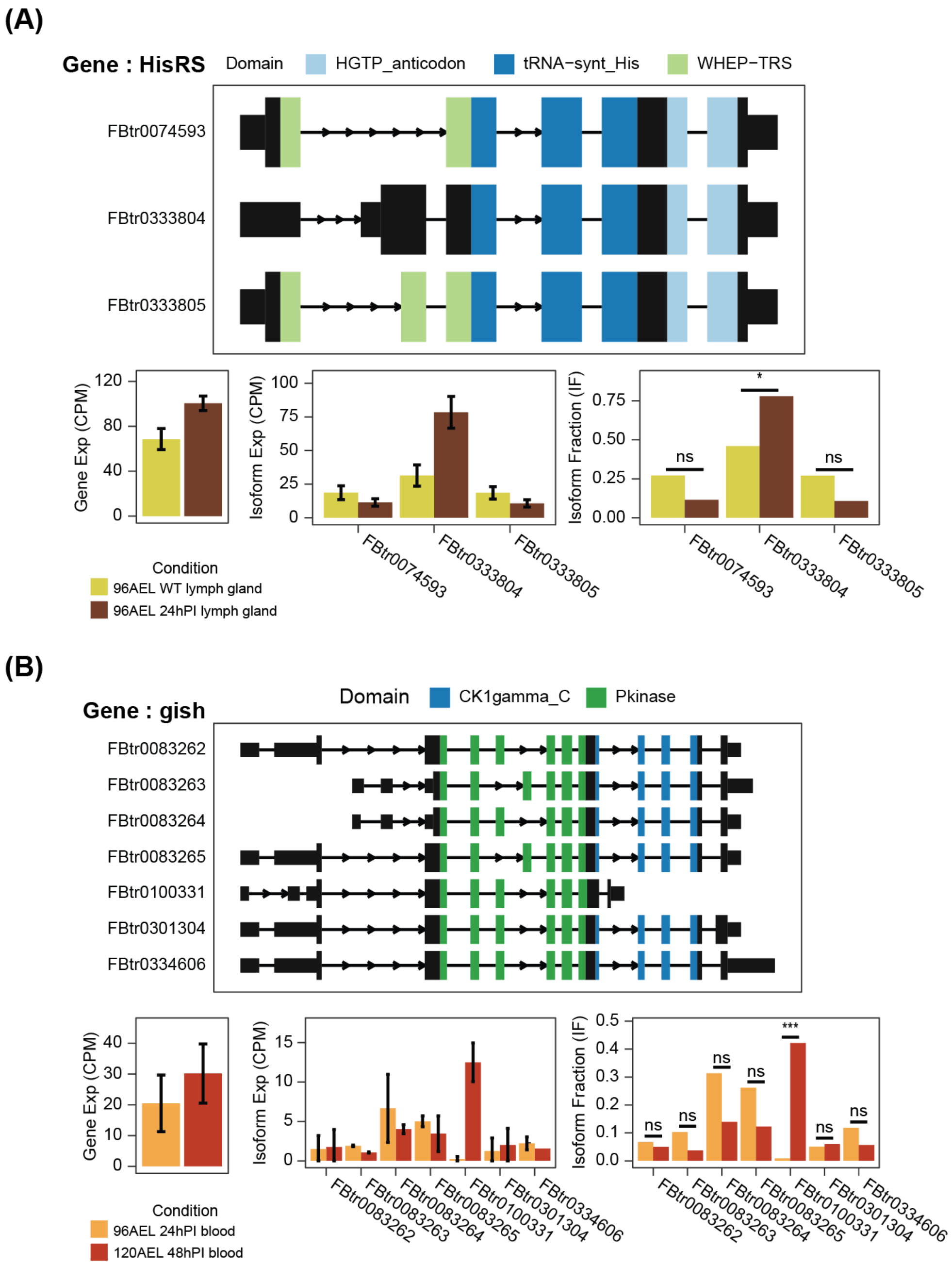


Figure II-25. Isoform switching leading to changes in protein sequences and domains. Schematic of isoform structures and changes in their expression and isoform fraction between 96hr AEL normal and wasp-infested lymph gland in HisRS gene **(A)**, and between 96hr AEL 24hPI and 120hr AEL 48hPI circulating hemocytes in gish gene **(B)**

II-３.１１ Fusion genes detected through long-read RNA-seq

Through long-read RNA, we were able to identify genes that are fused together (fusion genes) in both circulating hemocytes and lymph glands (Figure Ⅱ-26). As a result, 30 fusion events with at least 10 reads supporting the gene fusion were detected. Some of genes involved in fusion events were not expressed at all (exonic expression), suggesting fusion events taking place in intronic regions of the genes. Among these events, most prominent one was the fusion of PPO2 and CG13743 (Figure Ⅱ-26). Short, 5' part of PPO2 was observed to be fused to intronic sequence in CG13743 gene (Figure Ⅱ-27). These events were observed across all conditions being analyzed, implying that this fusion event is not a product of immune response against wasp infestation. As PPO2 is crucial gene in biological functions of crystal cells, further investigation is required to validate and explore the functions of this fusion gene. Genomic locations and structure of other fusion genes are depicted in Figure Ⅱ-27 and experimental validation of their formation and functions are being tested through experimental procedures.

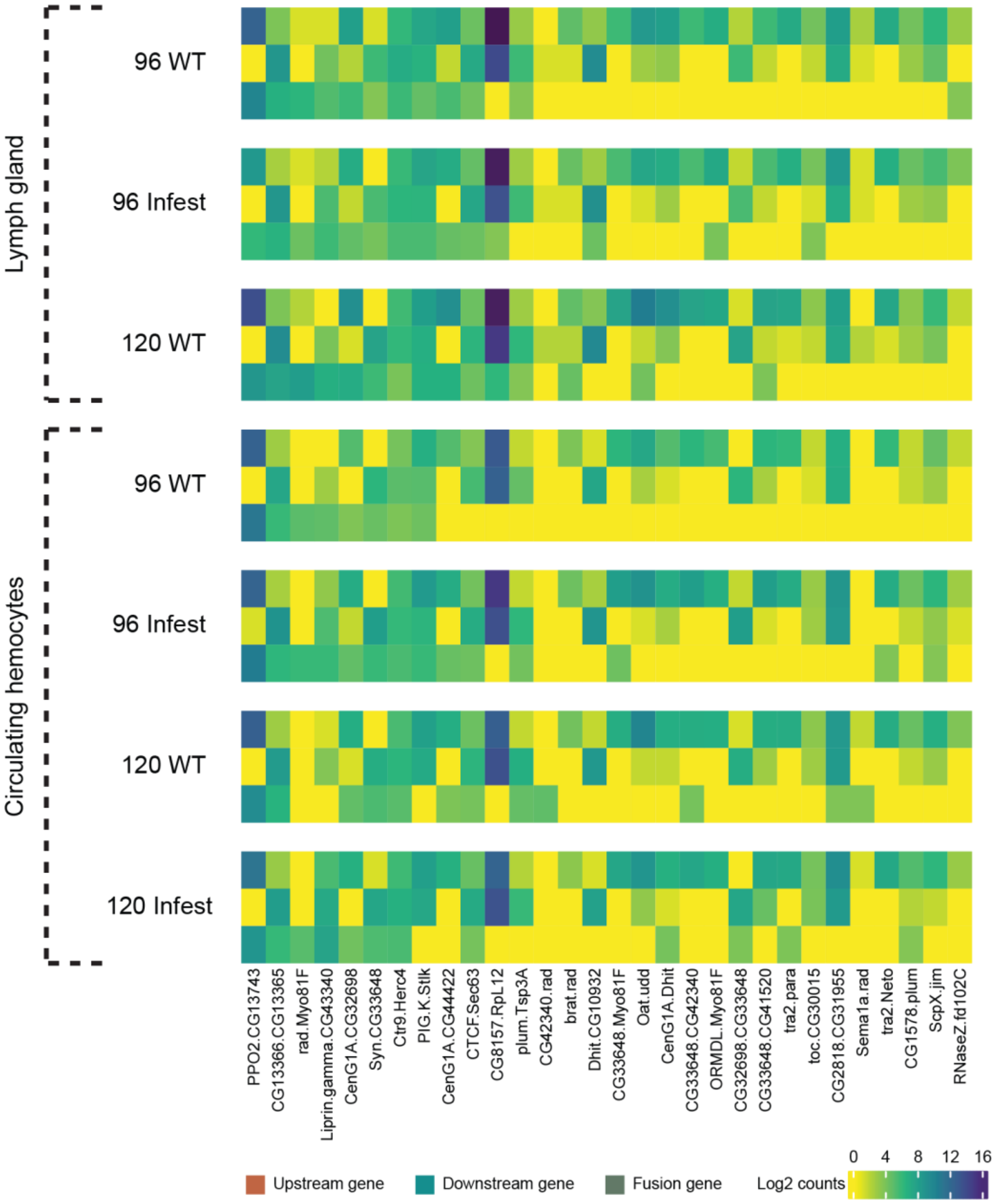


Figure II-26. Fusion genes detected by analyzing long-read RNA-seq data

Heatmap visualizing log2 counts of genes in upstream and downstream of break points and fusion genes themselves.

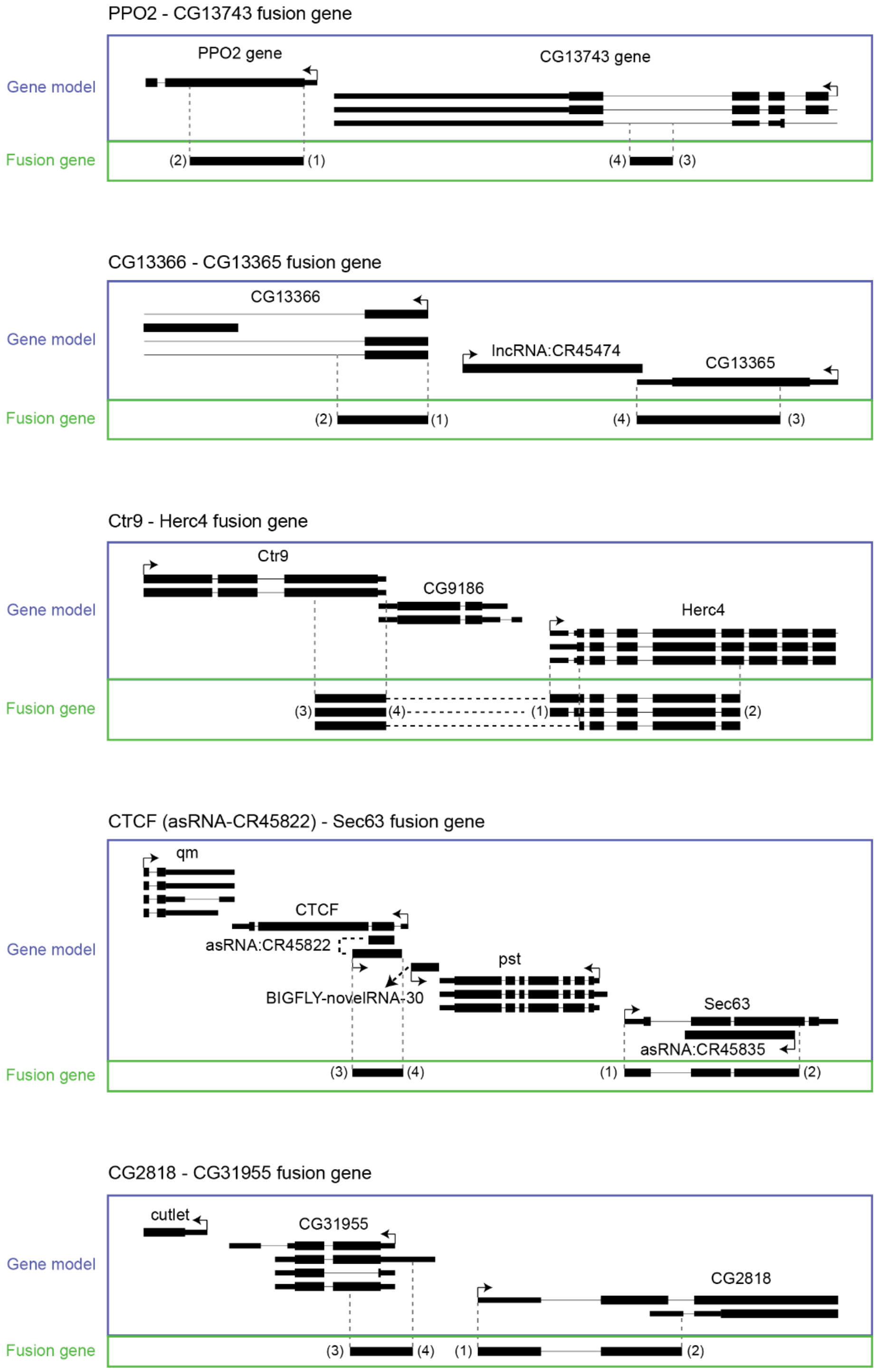


Figure II-27. Genomic location and structure of fusion genes

Genomic locations and structure of 5 fusion gene candidates are shown. Numbers in parenthesis indicate the order of fusion genes. Break points are (2) and (3).

II- 4 Discussion

In this study, we developed hybrid sequencing approach utilizing both short- and long-read RNA-seq data. Using the approach, we were able to overcome limitations of each sequencing data: Relatively high error rate of Nanopore sequencing data and low coverage at 5’ and 3’ ends of RNA molecules and short read length of Illumina sequencing data that make them difficult resolve accurate structure of long and complex transcripts.

The transcripts that were identified from two sequencing platforms were categorized into tier 1 and 2 based on their structural similarity. For tier 1 transcripts, only the Nanopore tier1 transcripts were used for further analysis in this study as the long-read RNA-seq data were shown to provide more superior coverage across gene body, especially at 3’end regions (Figure II-3). Nucleotide composition at the 3’ ends of assembled transcripts of each sequencing data further strengthened our decision in selection Nanopore tier1 over Illumina tier 1 transcripts (Figure II-4).

Through this hybrid sequencing approach, we have a complied 640 novel transcripts including novel lncRNAs originated from both known and unknown gene loci in *Drosophila* larvae. We are currently working to validate their expression in *Drosophila* larvae under WT and wasp-infested condition, which will provide more insights into their cellular localization and biological function in different tissues under normal or immune activated condition. In addition to lncRNAs, we also discovered 591 novel APA isoforms. Identification of these novel APA have led to increase in UMI detection and assignment to genes. This reflects the importance of correct annotation of 3’end in conventional scRNA-seq analysis as major types of scRNA-seq data capture 3’end of genes for expression profiling. This novel APA isoforms, moreover, could be expressed in specific condition such as in immune response against parasitic wasp infestation and part of key regulatory axis in controlling expression levels of essential genes. Further studies are required to dig into such possibilities.

Using the extensive gene annotation model, we analzyed scRNA-seq data from previous studies (Cho et al., 2020; Tattikota et al., 2020) on lymph gland and circulating hemocytes of *Drosophila* larvae under WT and wasp infested condition. In addition to previously identified lncRNAs that were highly expressed in specific cell type such as lamellocytes, which dramatically increase in their number and functions to exert immune response against parasitic egg laid by wasp, we were able to discover both known and novel lncRNAs highly expressed in specific cell type through comparing gene expression levels of different cell types and trajectory analysis. Experimental validation of expression of these lncRNAs in specific cell types and functional roles are being conducted, which could give us hints in functional roles of these non-coding RNAs in immune response of *Drosophila* larvae.

Through long-read RNA-seq, we discovered isoforms switching and their associated RNA AS and fusion gene events. Isoform switching events were observed to be most active between different time points of circulating hemocytes under wasp infestation conditions, possibly due to changes in cell population and immune defense mechanisms. Utilizing the advantage of long-read RNA-seq in identifying full-length RNA, we discovered 30 fusion gene events and most prominent one was the one between PPO2 and CG13743 gene. As this PPO2 gene is very crucial in crystal cell activity, further studies are required to validate and explore functions of these fusion gene.

In summary, using both short- and long-read RNA-seq, we have devised computational pipeline to assemble more accurate transcriptome. Using the update gene model, we have analyzed isoform switching and fusion gene events across developmental stages of *Drosophila* larvae under normal and immune challenges. Some of these events are being validated through experimental approaches. Lastly, both known and novel lncRNA markers expressed in specific cell types have been identified in single-cell resolution. Prominently, lncRNA markers were observed to be strongly expressed in lamellocytes in response to wasp infestation and are being experimentally tested of their association in cellular immunity of *Drosophila* larvae (Figure II-28).

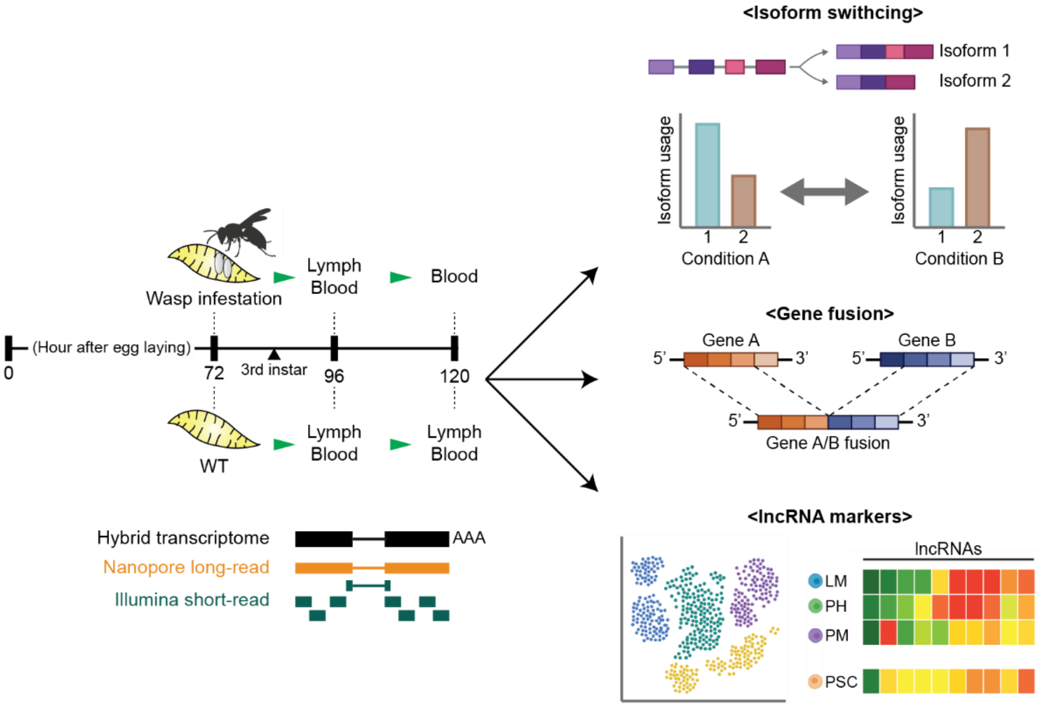


Figure II-28. Summary of Chapter II: Hybrid transcriptome analysis of *Drosophila* larvae under immune responses

II- 5 Materials & methods

II- 5 . 1 Experimental procedures

II- 5 . 1 . 1 Preparation of Nanopore cDNA sequencing library

PCR cDNA barcoding kit, SQK-PCB109, from Oxford Nanopore technologies were used for long-read RNA-seq analysis. Except for 120AEL WT lymph gland sample, which was sequenced only once without barcoding, RNAs from 6 conditions of Drosophila larvae (96AEL WT lymph gland, 96AEL 24hPI lymph gland, 96AEL WT blood, 96AEL 24hPI blood, 120AEL WT blood, and 120AEL 48hPI blood) were extracted twice and sequence three times to generate three replicates. For each run of Nanopore sequencing, 300 ng RNA of each condition X 6 were extracted and used for sequencing. Briefly, 300 ng RNA of each condition was first incubated at 65 °C for 5 minutes and then snapped cool on a pre-chilled freezer block. Strand switching buffer consisting of 4 μl 5x RT Buffer (ThermoFisher, cat # EP0751), 1 μl RNaseOUT (Life Technologies, cat # 10777019), 1 μl Nuclease-free water, and 2 μl Strand-Switching Primer (SSP, at 10 μM) was made and added to snap-cooled, annealed mRNA, followed by mixing by flicking the tube and spinning down. Then the tube was incubated at 42 °C for 2 minutes. Next, 1 μl of Maxima H Minus Reverse Transcriptase (ThermoFisher, cat # EP0751) was added and reverse transcription and strand-switching reaction was conducted for 90 minutes at 42 °C, followed by heat inactivation for 5 minutes at 85 °C. For each sample, following reaction reagents were prepared twice at room temperature (RT): 5 μl Reverse-transcribed RNA sample, 1.5 μl Barcode Primers (BP01-BP12), 18.5 μl Nuclease-free water, and 25 μl 2x LongAmp Taq Master Mix (NEB M0287). Reverse-transcribed RNA samples were amplified using the following cycling conditions: Initial denaturation for 30 seconds at 95° C (1 cycle), denaturation for 15 seconds at 95° C (12 cycles), annealing for 15 seconds at 62° C (12 cycles), extension for 10 minutes at 65° C (12 cycles), and final extension for 6 minutes at

65° C (1 cycle). After the amplifying reaction, 1 µl of NEB Exonuclease 1 (20 units, NEB, Cat # M0293) directly to each PCR tube and it was incubated for 15 minutes at 37° C, followed by incubation for 15 minutes at 80 ° C. PCR reactions of same barcodes (same condition) were pooled in a clean 1.5 ml Eppendorf DNA LoBind tube. Next, 0.8X equivalents of resuspended AMPure XP beads (Agencounrt) to the reaction and mix by pipetting. Incubate on a Hula mixer (rotator mixer) for 5 minutes at RT. Samples were spinned down and placed on a magnet stand (pellet visible at this step), during which the supernatant was removed by pipette. Tubes were kept on the magnet and beads were washed with 200 µl of freshly-prepared 70% ethanol without disturbing the pellet. The ethanol was removed, and the washing step was repeated once more. Tubes were spinned down and placed back on the magnet. Any residual 70% ethanol we pipetted off and tubes were briefly allow to dry for 1 to 5 minutes. After drying the tubes were removed from the magnetic rack and the pellet was resuspended in 12 µl of Elution Buffer (EB). The tube was Incubated on a Hula mixer for 10 minutes at RT. Beads were pelleted on magnet until the eluate is clear and colourless. 12 µl of clear and colourless eluate which contains the DNA library was retained in a clean 1.5 ml Eppendorf DNA LoBind tube and pelleted beads were disposed. 1 µl of the amplified DNA was analyzed by NanoDrop and subjected to 0.8 % agarose gel running for size, quantity and quality. 67~135 ng of amplified cDNA of each condition was pooled together to a final volume of 11 µl in Elution Buffer (EB). 1 µl of Rapid Adapter (RAP) was added to the amplified cDNA library and incubated for 5 minutes at RT. We then opened the MinION Mk1B lid and slide the flow cell under the clip. We slide the priming port cover clockwise to open the priming port and after opening it, we draw back a small volume to remove any bubbles (20~30 µl) using a P1000 pipette. Next, flow cell priming mix was made as follows: add 30 µl of thawed and mixed Flush Tether (FLT) directly to the tube of thawed and mixed Flush Buffer (FB) and mix by vortexing. 800 µl of the priming mix was loaded into the flow cell via the priming port and waited for 5 minutes. During this time, loading library was made as follows: 37.5 µl Sequencing Buffer (SQB), 25.5 µl Loading Beads (LB) that was throughly mixed immediately before use, and 12 µl DNA library. The flow cell priming was then done by gently lifting the SpotON sample port cover to make the SpotON sample port accessible and loading 200 µl of the priming mix into the flow cell via the priming port (not the SpotON sample port). Prepared loading library was gently mixed by pipetting up and down just prior to loading. The library was loaded by adding 75 µl of sample to the flow cell via the SpotON sample port in a dropwise fashion. The SpotON sample port cover was gently replaced, priming port was closed and finally, the MinION Mk1B lid was replaced.

II-５.１.２ Bulk RNA-seq of the circulating hemocyte

At 96 or 120 hr AEL, with or without wasp infestation, 100 to 150 larvae were dissected. Larvae were vortexed with glass beads (Sigma G9268) for one minute to get the entire larval circulating hemocytes. Ten larvae were dissected together in 20 l of Schneider's medium (Gibco, 21720024) and transferred to 100ul of cold Schneider's medium. Hemocyte samples were centrifuged at 7,000 rpm and 4 °C for 5 minutes. Supernatant was removed and 1000ul of Trizol (MRC, TR118) was added for RNA extraction. More than 1 g of RNA was prepared for each experiment. Both library preparation and sequencing were performed by the Macrogen (Macrogen, Inc., Seoul, South Korea). Illumina short-read RNA-seq of other samples were obtained from the previous study (Cho et al., 2020).

II-５.２ Computation and statistical procedures

II-５.２.１ Sequencing and base calling

Sequencing of Nanopore cDNA sequencing was performed on the laptop equipped with Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz 1.99GHz, 16GB RAM, and 1TB SSD. Fast5 files, which contain the raw electrical signal levels masured by the nanopores, were generated from MinKNOW software installed on the laptop. Resulting fast5 were merged and used for base calling by Guppy software (version 3.6.1-1) in high accuracy mode.

II-５.２.２ Hybrid transcriptome assembly pipeline

A. *Illumina RNA-seq data:* Illumina RNA-seq reads were first assessed of their quality using FastQC (version 0.11.8) (Andrews, 2010) and were trimmed based on Phred-scaled quality scores using Sickle (version 1.33) (Joshi NA, 2011). Reads were then aligned to genome sequence of Berkeley Drosophila Genome Project (BDGP) assembly release 6 (July 2014) using STAR (version 2.5.3a(Dobin et al., 2013)). Initial transcriptome assembly was performed using StringTie (version 2.1.3b) (Kovaka et al., 2019) without reference annotation for guiding the assembly (de novo assembly), setting minimum isoform fraction (-f) to 0.1--rf -f 0.1, minimum reads per bp coverage to consider for multi-exon transcript (-c) to 2.5, and fraction of bundle allowed to be covered by multi-hit reads (-M) to 0.95. Transcriptomes assembled from each condition were merged into single transcriptome using stringtie –merge, setting minimum input transcript TPM to include in the merge (-T) to 0.5.

B. *Nanopore cDNA sequencing data:* Nanopore cDNA reads were filtered on minimum average read quality score of 7 using NanoFilt (version 2.8.0) (De Coster, D'Hert, Schultz, Cruts, & Van Broeckhoven, 2018). Next, adapter sequences were trimmed off and reads were re-oriented by Pychopper. Reads were then corrected of sequencing error using Lordec (version 0.9) (Salmela & Rivals, 2014), utilizing k-mers from Illumina RNA-seq reads of same condition. Transcriptome assembly was performed using these filtered, re-oriented, and corrected Nanopore cDNA reads through FLAIR (version 1.5.1) (Tang et al., 2020) with slight modifications: When counting number of full-length reads that support the structure of assembled transcriptome, we used in-house script that calculate exonic overlap based on genomic coordinates of reads aligned to genome. Only the reads that show more than 80% exonic overlap with the assembled transcript were deemed as full-length reads and transcripts that show at least 5 full-length reads supporting their structure were kept for further analysis. Transcriptomes assembled from each condition were merged into single transcriptome using in-house script, as described in fly lncRNA Figure 1B.

II-５.２.３ Generation of hybrid transcriptome

Transcriptomes assembled and merged from Illumina and Nanopore sequencing data were compared against each other and classified as follows: Transcripts with single-exon exhibiting exonic overlap, and multi-exon transcripts with identical intron structure across both sequencing platforms, were categorized as “Tier 1”. Those that did not meet these criteria were classified as “Tier 2”. To improve the quality of Illumina Tier 2 transcripts, they were subjected to CAFÉ pipeline(You, Yoon, & Nam, 2017) as follows: 1. updating exon-junctions based on splicing reads ; 2. Updating 5’ and 3’end based on cap analysis of gene expression (CAGE) -seq (Brown et al., 2014) and 3p-seq data (TargetScan 7.2 (Lewis, Burge, & Bartel, 2005)), respectively.

II-５.２.４ Identification of novel RNAs and APA isoforms

This section explains detailed method implemented in the pipeline depicted in fly Figure II-7. The hybrid transcriptome was first filtered to only include transcripts whose expression level exceed the cutoff in both sequencing platform in a minimum of one condition (Nanopore CPM ≥ 1 and Illumina TPM ≥ 1). The filtered transcriptome were compared to BDGP6.22 reference gene annotation using GFFcompare (version 0.11.6) Those that do not overlap with any of reference gene was defined as “Novel loci RNAs” (transcript classification codes are "x", "i", "y","p", or "u" in GFFcompare). Those overlap with known lncRNAs were also defined (transcript classification codes are “j”, “o”, or “k” in GFFcompare). These two sets of RNAs were assessed of coding potential using coding potential calculator (CPC (Kong et al., 2007)) and coding potential assessment tool (CPAT (L. Wang et al., 2013)). When both tools determined RNAs to be non-coding or protein coding RNAs, they were defined as “Noncoding” and “Putative coding”, respectively. Otherwise, they were defined as “Ambiguous”.

To identify novel APA isoforms, assembled transcripts identified to be originated from reference transcripts by GFFcompare (transcript classification code is “=”) were inspected. For assembled transcripts that match the structure of reference transcript except for lncRNA, those whose expression levels and isoform fraction exceed the cutoff (Nanopore CPM ≥ 3, isoform fraction ≥ 0.2) were selected. For assembled transcripts that match the structure of known lncRNA transcripts, following expression and isoform fraction cutoffs were applied (Nanopore CPM ≥ 1, isoform fraction ≥ 0.2). Finally, among the selected transcripts, those with 3’ end at least 15 nt distant from reference transcripts were defined as novel APA isoforms.

II-５.２.５ PCA and correlation of gene expression across different high-throughput sequencing platforms

For PCA of multiple replicates of Nanopore sequencing data, transcript abundance, CPM, was estimated by NanoCount. Gene expression levels were determined by aggregating the CPM of all transcripts that originated from each respective gene. They were then transformed into log2 scale and corrected of batch effects using “removeBatchEffect” function implemented in limma R package (version 3.54.2). PCA was performed by factoextra R package (version 1.0.7).

To assess the correlation of gene expression levels between Nanopore, Illumina and scRNA-seq data, we calculated pseudo bulk CPM in each condition of scRNA-seq data, as follows: Relative read count of each gene was multiplied by scaling factor of 10^6 and average value of each gene from all cells of specific condition was calculated to derive pseudo bulk CPM. Expression levels from all sequencing platforms were transformed into log2 scale and used to calculate Pearson correlations.

II-５.２.６ Conservation of transcripts across multiple insect species

PhastCons scores (Siepel et al., 2005) generated from multiple sequence alignment of 27 insect species were downloaded from UCSC browser. (Kent et al., 2002) Phylogenetic tree of 27 insect species including Drosophila melanogaster is shown below.

II-５.２.７ Identification of differentially expressed lncRNAs

To investigate differentially expressed lncRNAs, we used edgeR R package (version 3.40.2) to discover differentially expressed genes (DEGs) in following comparisons : 96AEL WT lymph gland and 96AEL 24hPI lymph gland ; 96AEL WT blood and 96AEL 24hPI blood ; 120AEL WT blood and 120AEL 48hPI blood ; 96AEL WT blood and 120AEL WT blood ; 96AEL 24hPI blood and 120AEL 48hPI blood. As Illumina sequencing data were not replicated, read counts normalized by trimmed mean of M values (TMM) method implemented in edgeR R package and normalized read counts were used to calculate $\log_2$ fold changes. Among the lncRNAs, those that were observed to be differentially expressed in same direction in both sequencing data (Nanopore: $|$ log2 fold change $| \geq 2$ and FDR ≤ 0.05, Illumina: $|$ log2 fold change $| \geq 2$) were selected and plotted (fly lncRNA Figure 13).

II-５.２.８ Pre-processing of scRNA-seq data

Raw scRNA-seq data from the published studies (Cho et al., 2020; Tattikota et al., 2020) were generated in paired-end reads following single-cell capture using Drop-seq. Mapping of scRNA-seq data to genome sequence of BDGP assembly release 6 (July 2014) and extraction of digital gene expression (DGE) was performed by following the Drop-seq Core Computational Protocol version 2.0.0, which describe detailed workflow of Drop-seq tool version 2.4.0. Briefly, meta files needed for Drop-seq alignment were generated by running “create_Drop-seq_reference_metadata.sh” program implemented in Drop-seq tool with default parameters. Drop-seq alignment was conducted using “Drop-seq_alignment.sh” program implemented in Drop-seq tool with default parameters. Then, cell barcodes were sorted by number of reads assigned to them and filtered based on manual inspection as recommended in Drop-seq Core Computational Protocol. Number of cells retained in each Drop-seq library is summarized below. DGE for each sample was extracted by using “DigitalExpression” program implemented in Drop-seq tool and DGEs from all samples were merged into single DGE matrix using in-house script.

Sample	Library	Retained cells
96AEL_WT_lymphgland	lib1	500
96AEL_WT_lymphgland	lib2	3000
96AEL_WT_lymphgland	lib3	1000
96AEL_WT_lymphgland	lib4	2000
96AEL_WT_lymphgland	lib5	1800
96AEL_24hPI_lymphgland	lib1	2000
96AEL_24hPI_lymphgland	lib2	1600
96AEL_24hPI_lymphgland	lib3	4500
96AEL_24hPI_lymphgland	lib4	1700
96AEL_WT_blood	lib1	500
96AEL_WT_blood	lib2	200
96AEL_WT_blood	lib3	500
96AEL_24hPI_blood	lib1	200
96AEL_24hPI_blood	lib2	2000
96AEL_24hPI_blood	lib3	4000
120AEL_WT_blood	lib1	800
120AEL_WT_blood	lib2	500
120AEL_WT_blood	lib3	600
120AEL_48hPI_blood	lib1	1500
120AEL_48hPI_blood	lib2	2000
120AEL_48hPI_blood	lib3	3000
72AEL_WT_lymphgland	lib1	1000
72AEL_WT_lymphgland	lib2	400
72AEL_WT_lymphgland	lib3	200
72AEL_WT_lymphgland	lib4	500
72AEL_WT_lymphgland	lib5	200
120AEL_WT_lymphgland	lib1	2000
120AEL_WT_lymphgland	lib2	1600
120AEL_WT_lymphgland	lib3	4000
120AEL_WT_lymphgland	lib4	2000

Table II-1. Summary of number of cells retained in each Drop-seq library.

II-５.２.９ scRNA-seq data analysis using Seurat

The resulting DGE matrix was analyzed by Seurat R package (version 4.3.0). First, using the cell type labels extensively annotated in the preprinted work (Sang-Ho Yoon, 2023), we assigned each cell to correct hemocyte type and removed those without cell type information. Secondly, each library was filtered by mitochondrial contents (<10%) to remove low-quality cells, resulting in 36,007 cells. UMI counts of filtered cells were normalized, log-transformed and scaled using the functions "NormalizeData" and "ScaleData" that are implemented in Seurat R package for proper data integration. PCA was performed and degree of explained variability by each principal component (PC) was inspected through JackStraw plot. 54 PCs were selected to explain the variability of the scaled UMI counts across cells. For further analysis of dimension reduction of single-cell data using t-distributed stochastic neighbor embedding (t-SNE) and UMAP plot, Harmony R package (version 0.1.1) (Korsunsky et al., 2019) was used for integration of single-cell data from multiple conditions. t-SNE and UMAP analysis were performed using "RunTSNE" and "RunUMAP" functions implemented in Seurat R package. Detailed workflow of the Seurat is well explained on the Seurat website (https://satijalab.org/seurat/).

II-５.２.１０ Identification of lncRNA markers in single-cell levels

To explore lncRNA marker genes expressed in specific cell type, normalized, and scaled UMI counts across all genes were utilized. Using the "FindAllMarkers" function implemented in Seurat R package, we identified set of lncRNA markers that are detected in minimum fraction of 0.25 in either of the two populations being compared and whose average $\log_2$ fold change against all other cell populations is at least 0.5.

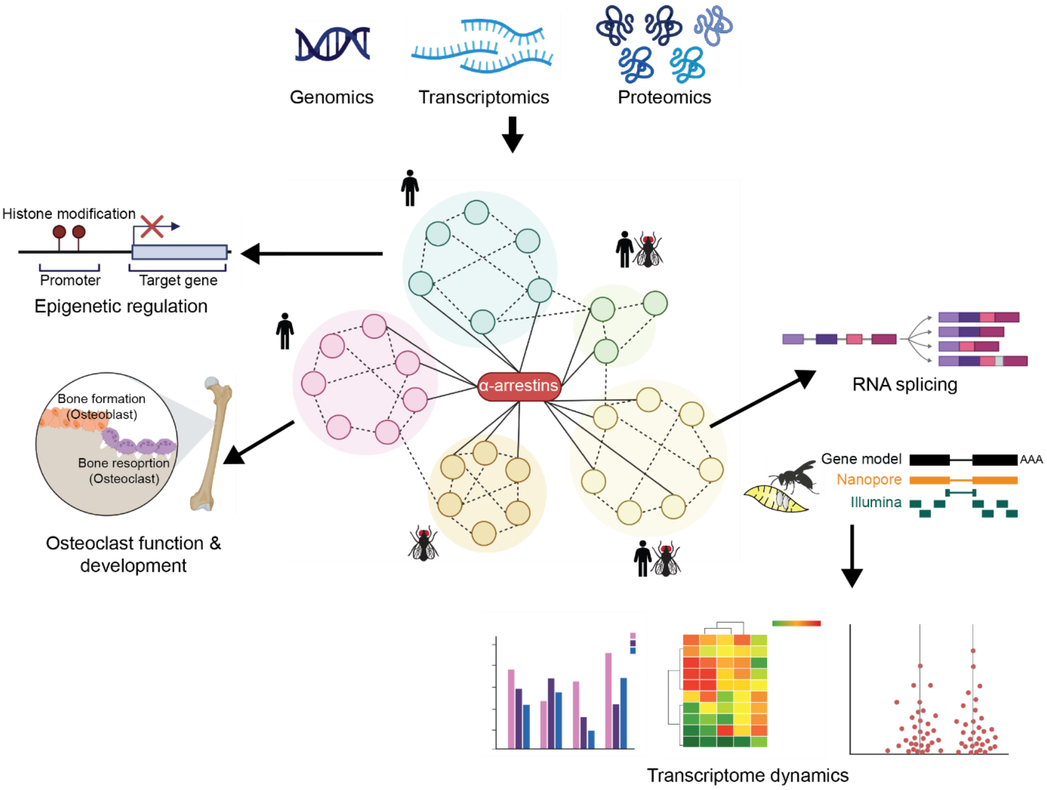
For trajectory analysis, Monocle 3 (version 3_1.3.1) was utilized. Seurat object from "scRNA-seq data analysis using Seurat" was converted to monocle3 object using the function, "as.cell_data_set". Principal graph was learned from UMAP space constructed in the analysis using Seurat R package and cells were ordered according to pseudotime by setting cells from 72 hr AEL WT lymp gland as starting point (root node). To extract cell lineage involving lamellocyte, “choose_graph_segments” function was used by setting start and end nodes of lamellocyte lineage that were determined from manual inspection. lncRNAs that are differentially expressed along the LM lineage were identified through “graph_test” function, which conduct spatial correlation analysis using the Morna’s I test. Whether cells at nearby positions on a trajectory will have similar or dissimilar expressions is determined from statistics from the test. We selected lncRNAs whose q value from the test is below 0.01, are expressed in a minimum of 0.1 fraction of cells being analyzed with minimum UMI count 5.

General discussion

In this study, we have constructed high-confidence interactomes of α-arrestins from human and *Drosophila*, which greatly expanded previously known PPIs involving α-arrestins (Summary Figure1). The interactomes hint toward many uncharacterized aspects of α-arrestins's biology and suggest conserved roles between the two species. Additionally, we discovered conserved functions, such as RNA splicing and novel cellular functions specific to human α-arrestins (Summary Figure1). The investigation of specific interacting protein complexes and their functions in α-arrestins could further our understanding of their roles in various disease models.

We also have developed a hybrid sequencing approach utilizing both short- and long-read RNA-seq data to overcome the limitations of each sequencing data (Summary Figure1), such as the relatively high error rate of Nanopore sequencing data and the low coverage at 5' and 3' ends of RNA molecules and short read length of Illumina sequencing data. This approach has allowed us to resolve the accurate structure of long and complex transcripts, which in turn has led to the identification of 640 novel transcripts including novel lncRNAs and 591 novel APA isoforms in *Drosophil*a larvae. By combining our extensive gene annotation model with scRNA-seq data from previous studies on lymph gland and circulating hemocytes (Cho et al., 2020; Tattikota et al., 2020) and *Drosophila* larvae under WT and wasp-infested conditions, we have discovered both known and novel lncRNAs highly expressed in specific cell types (Summary Figure1). Further experimental validation and investigation of the functional roles of these non-coding RNAs in immune response of *Drosophila* larvae are currently underway.

Our study provides a comprehensive resource for the community, offering detailed α-arrestins interactome maps and gene annotations to facilitate future research. Furthermore, our hybrid sequencing approach could be applied to other organisms and biological systems, providing valuable insights into the transcriptomic landscape and functional roles of lncRNAs in diverse biological pathway, including immune responses we study here.



Summary Figure 1. Overall summary of the study

References

Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature, 537*(7620), 347-355. doi:10.1038/nature19949

Akalin, A., Franke, V., Vlahovicek, K., Mason, C. E., & Schubeler, D. (2015). Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics, 31*(7), 1127-1129. doi:10.1093/bioinformatics/btu775

Alberts, B. (1998). The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell, 92*(3), 291-294. doi:Doi 10.1016/S0092-8674(00)80922-8

Alvarez, C. E. (2008). On the origins of arrestin and rhodopsin. *BMC Evol Biol, 8*, 222. doi:10.1186/1471-2148-8-222

Andoh, T., Hirata, Y., & Kikuchi, A. (2002). PY motifs of Rod1 are required for binding to Rsp5 and for drug resistance. *FEBS Lett, 525*(1-3), 131-134. doi:10.1016/s0014-5793(02)03104-6

Andrews, S. (2010). A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Eppig, J. T. (2000). Gene ontology: tool for the unification of biology. *Nature genetics, 25*(1), 25.

Banerjee, U., Girard, J. R., Goins, L. M., & Spratford, C. M. (2019). Drosophila as a Genetic Model for Hematopoiesis. *Genetics, 211*(2), 367-417. doi:10.1534/genetics.118.300223

Batista, T. M., Dagdeviren, S., Carroll, S. H., Cai, W., Melnik, V. Y., Noh, H. L., . . . Lee, R. T. (2020). Arrestin domain-containing 3 (Arrdc3) modulates insulin action and glucose metabolism in liver. *Proc Natl Acad Sci U S A, 117*(12), 6733-6740. doi:10.1073/pnas.1922370117

Benovic, J. L., DeBlasi, A., Stone, W. C., Caron, M. G., & Lefkowitz, R. J. (1989). Beta-adrenergic receptor kinase: primary structure delineates a multigene family. *Science, 246*(4927), 235-240. doi:10.1126/science.2552582

Bier, E. (2005). Drosophila, the golden bug, emerges as a tool for human genetics. *Nature Reviews Genetics, 6*(1), 9-23. doi:10.1038/nrg1503

Boase, N. A., & Kelly, J. M. (2004). A role for creD, a carbon catabolite repression gene from Aspergillus nidulans, in ubiquitination. *Mol Microbiol, 53*(3), 929-940. doi:10.1111/j.1365-2958.2004.04172.x

Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., . . . Celniker, S. E. (2014). Diversity and dynamics of the Drosophila transcriptome. *Nature, 512*(7515), 393-399. doi:10.1038/nature12962

Byrne, A., Beaudin, A. E., Olsen, H. E., Jain, M., Cole, C., Palmer, T., . . . Vollmers, C. (2017). Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications, 8*. doi:ARTN 16027

10.1038/ncomms16027

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development, 25*(18), 1915-1927. doi:10.1101/gad.17446611

Cao, J. Y., Spielmann, M., Qiu, X. J., Huang, X. F., Ibrahim, D. M., Hill, A. J., . . . Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature, 566*(7745), 496-+. doi:10.1038/s41586-019-0969-x

Cattenoz, P. B., Sakr, R., Pavlidaki, A., Delaporte, C., Riba, A., Molina, N., . . . Giangrande, A. (2020). Temporal specificity and heterogeneity of Drosophila immune cells. *Embo Journal, 39*(12), e104486. doi:10.15252/embj.2020104486

Chen, K. S., & DeLuca, H. F. (1994). Isolation and characterization of a novel cDNA from HL-60 cells treated with 1,25-dihydroxyvitamin D-3. *Biochim Biophys Acta, 1219*(1), 26-32. doi:10.1016/0167-4781(94)90242-9

Chen, Y., Ning, J., Cao, W., Wang, S., Du, T., Jiang, J., . . . Zhang, B. (2020). Research Progress of TXNIP as a Tumor Suppressor Gene Participating in the Metabolic Reprogramming and Oxidative Stress of Cancer Cells in Various Cancers. *Front Oncol, 10*, 568574. doi:10.3389/fonc.2020.568574

Cho, B., Yoon, S. H., Lee, D., Koranteng, F., Tattikota, S. G., Cha, N., . . . Shim, J. (2020). Single-cell transcriptome maps of myeloid blood cell lineages in Drosophila. *Nat Commun, 11*(1), 4483. doi:10.1038/s41467-020-18135-y

Choi, H., Larsen, B., Lin, Z. Y., Breitkreutz, A., Mellacheruvu, D., Fermin, D., . . . Nesvizhskii, A. I. (2011). SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nature Methods, 8*(1), 70-73. doi:10.1038/nmeth.1541

Colland, F., Jacq, X., Trouplin, V., Mougin, C., Groizeleau, C., Hamburger, A., . . . Gauthier, J. M. (2004). Functional proteomics mapping of a human signaling pathway. *Genome Research, 14*(7), 1324-1332. doi:10.1101/gr.2334104

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics, 34*(15), 2666-2669. doi:10.1093/bioinformatics/bty149

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., . . . Guigo, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research, 22*(9), 1775-1789. doi:10.1101/gr.132159.111

DeWire, S. M., Ahn, S., Lefkowitz, R. J., & Shenoy, S. K. (2007). Beta-arrestins and cell signaling. *Annu Rev Physiol, 69*, 483-510. doi:10.1146/annurev.physiol.69.022405.154749

Diaz, L. K., Cristofanilli, M., Zhou, X., Welch, K. L., Smith, T. L., Yang, Y., . . . Gilcrease, M. Z. (2005). Beta4 integrin subunit gene expression correlates with tumor size and nuclear grade in early breast cancer. *Mod Pathol, 18*(9), 1165-1175. doi:10.1038/modpathol.3800411

Dinkel, H., Van Roey, K., Michael, S., Kumar, M., Uyar, B., Altenberg, B., . . . Behrendt, A. (2015). ELM 2016—data update and new functionality of the eukaryotic linear motif resource. *Nucleic acids research, 44*(D1), D294-D300.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics, 29*(1), 15-21. doi:10.1093/bioinformatics/bts635

Domingues, A., Jolibois, J., Marquet de Rouge, P., & Nivet-Antoine, V. (2021). The Emerging Role of TXNIP in Ischemic and Cardiovascular Diseases; A Novel Marker and Therapeutic Target. *Int J Mol Sci, 22*(4). doi:10.3390/ijms22041693

Dores, M. R., Lin, H., N, J. G., Mendez, F., & Trejo, J. (2015). The alpha-arrestin ARRDC3 mediates ALIX ubiquitination and G protein-coupled receptor lysosomal sorting. *Mol Biol Cell, 26*(25), 4660-4673. doi:10.1091/mbc.E15-05-0284

Dotimas, J. R., Lee, A. W., Schmider, A. B., Carroll, S. H., Shah, A., Bilen, J., . . . Yoshioka, J. (2016). Diabetes regulates fructose absorption through thioredoxin-interacting protein. *Elife, 5*, e18313.

Draheim, K. M., Chen, H. B., Tao, Q., Moore, N., Roche, M., & Lyle, S. (2010). ARRDC3 suppresses breast cancer progression by negatively regulating integrin beta4. *Oncogene, 29*(36), 5032-5047. doi:10.1038/onc.2010.250

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., . . . Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res, 47*(D1), D427-D432. doi:10.1093/nar/gky995

Eng, J. K., Jahan, T. A., & Hoopmann, M. R. (2013). Comet: an open-source MS/MS sequence database search tool. *Proteomics, 13*(1), 22-24. doi:10.1002/pmic.201200439

Evans, C. J., Liu, T., & Banerjee, U. (2014). Drosophila hematopoiesis: Markers and methods for molecular genetic analysis. *Methods, 68*(1), 242-251. doi:10.1016/j.ymeth.2014.02.038

Feng, S. M., Deng, L. F., Chen, W., Shao, J. Z., Xu, G. L., & Li, Y. P. (2009). Atp6v1c1 is an essential component of the osteoclast proton pump and in F-actin ring formation in osteoclasts. *Biochemical Journal, 417*, 195-203. doi:10.1042/Bj20081073

Fields, S., & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature, 340*(6230), 245-246. doi:10.1038/340245a0

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., . . . Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res, 47*(D1), D766-D773. doi:10.1093/nar/gky955

Fu, Y., Huang, X., Zhang, P., van de Leemput, J., & Han, Z. (2020). Single-cell RNA sequencing identifies novel cell types in Drosophila blood. *J Genet Genomics, 47*(4), 175-186. doi:10.1016/j.jgg.2020.02.004

Girard, J. R., Goins, L. M., Vuu, D. M., Sharpley, M. S., Spratford, C. M., Mantri, S. R., & Banerjee, U. (2021). Paths and pathways that generate cell-type heterogeneity and developmental progression in hematopoiesis. *Elife, 10*. doi:10.7554/eLife.67516

Gleeson, J., Leger, A., Prawer, Y. D. J., Lane, T. A., Harrison, P. J., Haerty, W., & Clark, M. B. (2022). Accurate expression quantification from nanopore direct RNA sequencing with NanoCount. *Nucleic Acids Res, 50*(4), e19. doi:10.1093/nar/gkab1129

Gold, K. S., & Bruckner, K. (2015). Macrophages and cellular immunity in Drosophila melanogaster. *Semin Immunol, 27*(6), 357-368. doi:10.1016/j.smim.2016.03.010

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet, 17*(6), 333-351. doi:10.1038/nrg.2016.49

Gramates, L. S., Marygold, S. J., Santos, G. D., Urbano, J. M., Antonazzo, G., Matthews, B. B., . . . the FlyBase, C. (2017). FlyBase at 25: looking to the future. *Nucleic Acids Res, 45*(D1), D663-D671. doi:10.1093/nar/gkw1016

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics, 32*(18), 2847-2849. doi:10.1093/bioinformatics/btw313

Guruharsha, K. G., Rual, J. F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., . . . Artavanis-Tsakonas, S. (2011). A protein complex network of Drosophila melanogaster. *Cell, 147*(3), 690-703. doi:10.1016/j.cell.2011.08.047

Han, S. H., Jeon, J. H., Ju, H. R., Jung, U., Kim, K. Y., Yoo, H. S., . . . Choi, I. (2003). VDUP1 upregulated by TGF-beta1 and 1,25-dihydorxyvitamin D3 inhibits tumor cell growth by blocking cell-cycle progression. *Oncogene, 22*(26), 4035-4046. doi:10.1038/sj.onc.1206610

Han, S. O., Kommaddi, R. P., & Shenoy, S. K. (2013). Distinct roles for beta-arrestin2 and arrestin-domain-containing proteins in beta2 adrenergic receptor trafficking. *EMBO Rep, 14*(2), 164-171. doi:10.1038/embor.2012.187

Hartwell, L. H., Hopfield, J. J., Leibler, S., & Murray, A. W. (1999). From molecular to modular cell biology. *Nature, 402*(6761 Suppl), C47-52. doi:10.1038/35011540

Hellemans, J., Mortier, G., De Paepe, A., Speleman, F., & Vandesompele, J. (2007). qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data. *Genome Biol, 8*(2), R19. doi:10.1186/gb-2007-8-2-r19

Herranz, S., Rodriguez, J. M., Bussink, H. J., Sanchez-Ferrero, J. C., Arst, H. N., Jr., Penalva, M. A., & Vincent, O. (2005). Arrestin-related proteins mediate pH signaling in fungi. *Proc Natl Acad Sci U S A, 102*(34), 12141-12146. doi:10.1073/pnas.0504776102

Honti, V., Csordas, G., Kurucz, E., Markus, R., & Ando, I. (2014). The cell-mediated immunity of Drosophila melanogaster: hemocyte lineages, immune compartments, microanatomy and regulation. *Developmental and Comparative Immunology, 42*(1), 47-56. doi:10.1016/j.dci.2013.06.005

Hosack, D. A., Dennis, G., Jr., Sherman, B. T., Lane, H. C., & Lempicki, R. A. (2003). Identifying biological themes within lists of genes with EASE. *Genome Biol, 4*(10), R70. doi:10.1186/gb-2003-4-10-r70

Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., Berger, B., Perrimon, N., & Mohr, S. E. (2011). An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics, 12*(1), 357.

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res, 37*(1), 1-13. doi:10.1093/nar/gkn923

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc, 4*(1), 44-57. doi:10.1038/nprot.2008.211

Hultmark, D., & Ando, I. (2022). Hematopoietic plasticity mapped in Drosophila and other insects. *Elife, 11*. doi:10.7554/eLife.78906

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., . . . Gygi, S. P. (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell, 162*(2), 425-440. doi:10.1016/j.cell.2015.06.043

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., . . . Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol, 36*(4), 338-345. doi:10.1038/nbt.4060

Jan, C. H., Friedman, R. C., Ruby, J. G., & Bartel, D. P. (2011). Formation, regulation and evolution of Caenorhabditis elegans 3 ' UTRs. *Nature, 469*(7328), 97-U114. doi:10.1038/nature09616

Jin Lee, G. C., Grey Christoforo, CS Foo, Chris Probert, Anshul Kundaje, Nathan Boley, kohpangwei, Mike Dacre, Daniel Kim. (2016). kundajelab/atac_dnase_pipelines: 0.3.3.

Joshi NA, F. J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files

(Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.

Jung, S. H., Evans, C. J., Uemura, C., & Banerjee, U. (2005). The Drosophila lymph gland as a developmental model of hematopoiesis. *Development, 132*(11), 2521-2533. doi:10.1242/dev.01837

Junn, E., Han, S. H., Im, J. Y., Yang, Y., Cho, E. W., Um, H. D., . . . Choi, I. (2000). Vitamin D3 up-regulated protein 1 mediates oxidative stress via suppressing the thioredoxin function. *J Immunol, 164*(12), 6287-6295. doi:10.4049/jimmunol.164.12.6287

Keller, A., Eng, J., Zhang, N., Li, X. J., & Aebersold, R. (2005). A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol Syst Biol, 1*, 2005 0017. doi:10.1038/msb4100024

Keller, A., Nesvizhskii, A. I., Kolker, E., & Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem, 74*(20), 5383-5392.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research, 12*(6), 996-1006. doi:10.1101/gr.229102

Kim, S. K., Choe, J. Y., & Park, K. Y. (2019). TXNIP-mediated nuclear factor-kappaB signaling pathway and intracellular shifting of TXNIP in uric acid-induced NLRP3 inflammasome. *Biochem Biophys Res Commun, 511*(4), 725-731. doi:10.1016/j.bbrc.2019.02.141

Kim, S. Y., Lee, E. H., Park, S. Y., Choi, H., Koh, J. T., Park, E. K., . . . Kim, J. E. (2019). Ablation of Stabilin-1 Enhances Bone-Resorbing Activity in Osteoclasts In Vitro. *Calcif Tissue Int, 105*(2), 205-214. doi:10.1007/s00223-019-00552-x

Kim, Y. M., & Benovic, J. L. (2002). Differential roles of arrestin-2 interaction with clathrin and adaptor protein 2 in G protein-coupled receptor trafficking. *J Biol Chem, 277*(34), 30760-30768. doi:10.1074/jbc.M204528200

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., & Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res, 35*(Web Server issue), W345-349. doi:10.1093/nar/gkm391

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., . . . Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods, 16*(12), 1289-+. doi:10.1038/s41592-019-0619-0

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology, 20*(1). doi:ARTN 278

10.1186/s13059-019-1910-1

Kuhn, H., Hall, S. W., & Wilden, U. (1984). Light-induced binding of 48-kDa protein to photoreceptor membranes is highly enhanced by phosphorylation of rhodopsin. *FEBS Lett, 176*(2), 473-478. doi:10.1016/0014-5793(84)81221-1

Kwon, H. J., Won, Y. S., Suh, H. W., Jeon, J. H., Shao, Y., Yoon, S. R., . . . Choi, I. (2010). Vitamin D3 upregulated protein 1 suppresses TNF-alpha-induced NF-kappaB activation in hepatocarcinogenesis. *J Immunol, 185*(7), 3980-3989. doi:10.4049/jimmunol.1000990

Kwon, Y., Vinayagam, A., Sun, X., Dephoure, N., Gygi, S. P., Hong, P., & Perrimon, N. (2013). The Hippo signaling pathway interactome. *Science, 342*(6159), 737-740. doi:10.1126/science.1243971

Kyriakakis, P., Tipping, M., Abed, L., & Veraksa, A. (2008). Tandem affinity purification in Drosophila: the advantages of the GS-TAP system. *Fly (Austin), 2*(4), 229-235. doi:10.4161/fly.6669

Lanot, R., Zachary, D., Holder, F., & Meister, M. (2001). Postembryonic hematopoiesis in Drosophila. *Developmental Biology, 230*(2), 243-257. doi:DOI 10.1006/dbio.2000.0123

Leitao, A. B., Arunkumar, R., Day, J. P., Geldman, E. M., Morin-Poulard, I., Crozatier, M., & Jiggins, F. M. (2020). Constitutive activation of cellular immunity underlies the evolution of resistance to infection in Drosophila. *Elife, 9*. doi:10.7554/eLife.59095

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell, 120*(1), 15-20. doi:10.1016/j.cell.2004.12.035

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics, 12*, 323. doi:10.1186/1471-2105-12-323

Li, K. Q., Tian, Y. L. Z., Yuan, Y., Fan, X. L., Yang, M. Y., He, Z., & Yang, D. Y. (2019). Insights into the Functions of LncRNAs in Drosophila. *International Journal of Molecular Sciences, 20*(18). doi:ARTN 4646

10.3390/ijms20184646

Lin, C. H., MacGurn, J. A., Chu, T., Stefan, C. J., & Emr, S. D. (2008). Arrestin-related ubiquitin-ligase adaptors regulate endocytosis and protein turnover at the cell surface. *Cell, 135*(4), 714-725. doi:10.1016/j.cell.2008.09.025

Lohse, M. J. (1992). Stable overexpression of human beta 2-adrenergic receptors in mammalian cells. *Naunyn Schmiedebergs Arch Pharmacol, 345*(4), 444-451. doi:10.1007/BF00176623

Lu, S., Simin, K., Khan, A., & Mercurio, A. M. (2008). Analysis of integrin beta4 expression in human breast cancer: association with basal-like tumors and prognostic significance. *Clin Cancer Res, 14*(4), 1050-1058. doi:10.1158/1078-0432.CCR-07-4116

Lundgren, D. H., Hwang, S.-I., Wu, L., & Han, D. K. (2010). Role of spectral counting in quantitative proteomics. *Expert review of proteomics, 7*(1), 39-53.

Macias, M. J., Hyvonen, M., Baraldi, E., Schultz, J., Sudol, M., Saraste, M., & Oschkinat, H. (1996). Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide. *Nature, 382*(6592), 646-649. doi:DOI 10.1038/382646a0

Malik, R., & Marchese, A. (2010). Arrestin-2 interacts with the endosomal sorting complex required for transport machinery to modulate endosomal sorting of CXCR4. *Mol Biol Cell, 21*(14), 2529-2541. doi:10.1091/mbc.E10-02-0169

Mandal, L., Martinez-Agosto, J. A., Evans, C. J., Hartenstein, V., & Banerjee, U. (2007). A Hedgehog- and Antennapedia-dependent niche maintains Drosophila haematopoietic precursors. *Nature, 446*(7133), 320-324. doi:10.1038/nature05585

Markus, R., Kurucz, T., Rus, F., & Ando, I. (2005). Sterile wounding is a minimal and sufficient trigger for a cellular immune response in Drosophila melanogaster. *Immunology Letters, 101*(1), 108-111. doi:10.1016/j.imlet.2005.03.021

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nat Rev Genet, 12*(10), 671-682. doi:10.1038/nrg3068

Matsumoto, N., & Nakanishi-Matsui, M. (2019). Proton pumping V-ATPase inhibitor bafilomycin A1 affects Rab7 lysosomal localization and abolishes anterograde trafficking of osteoclast secretory lysosomes. *Biochemical and Biophysical Research Communications, 510*(3), 421-426. doi:10.1016/j.bbrc.2019.01.118

Mellacheruvu, D., Wright, Z., Couzens, A. L., Lambert, J. P., St-Denis, N. A., Li, T., . . . Nesvizhskii, A. I. (2013). The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nature Methods, 10*(8), 730-736. doi:10.1038/nmeth.2557

Mo, J. S., Park, H. W., & Guan, K. L. (2014). The Hippo signaling pathway in stem cell biology and cancer. *EMBO Rep, 15*(6), 642-656. doi:10.15252/embr.201438638

Mohankumar, K. M., Currle, D. S., White, E., Boulos, N., Dapper, J., Eden, C., . . . Gilbertson, R. J. (2015). An in vivo screen identifies ependymoma oncogenes and tumor-suppressor genes. *Nat Genet, 47*(8), 878-887. doi:10.1038/ng.3323

Nabhan, J. F., Hu, R., Oh, R. S., Cohen, S. N., & Lu, Q. (2012). Formation and release of arrestin domain-containing protein 1-mediated microvesicles (ARMMs) at plasma membrane by recruitment of TSG101 protein. *Proc Natl Acad Sci U S A, 109*(11), 4146-4151. doi:10.1073/pnas.1200448109

Nabhan, J. F., Pan, H., & Lu, Q. (2010). Arrestin domain-containing protein 3 recruits the NEDD4 E3 ligase to mediate ubiquitination of the beta2-adrenergic receptor. *EMBO Rep, 11*(8), 605-611. doi:10.1038/embor.2010.80

Nesvizhskii, A. I., Keller, A., Kolker, E., & Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem, 75*(17), 4646-4658.

Nishinaka, Y., Masutani, H., Oka, S., Matsuo, Y., Yamaguchi, Y., Nishio, K., . . . Yodoi, J. (2004). Importin alpha1 (Rch1) mediates nuclear translocation of thioredoxin-binding protein-2/vitamin D(3)-up-regulated protein 1. *J Biol Chem, 279*(36), 37559-37565. doi:10.1074/jbc.M405473200

Nishiyama, A., Matsui, M., Iwata, S., Hirota, K., Masutani, H., Nakamura, H., . . . Yodoi, J. (1999). Identification of thioredoxin-binding protein-2/vitamin D-3 up-regulated protein 1 as a negative regulator of thioredoxin function and expression. *Journal of Biological Chemistry, 274*(31), 21645-21650. doi:DOI 10.1074/jbc.274.31.21645

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology, 302*(1), 205-217. doi:10.1006/jmbi.2000.4042

Oka, S., Masutani, H., Liu, W., Horita, H., Wang, D., Kizaka-Kondoh, S., & Yodoi, J. (2006). Thioredoxin-binding protein-2-like inducible membrane protein is a novel vitamin D3 and peroxisome proliferator-activated receptor (PPAR)gamma ligand target protein that regulates PPARgamma signaling. *Endocrinology, 147*(2), 733-743. doi:10.1210/en.2005-0679

Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., . . . Ahn, N. G. (2005). Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics, 4*(10), 1487-1502. doi:10.1074/mcp.M500084-MCP200

Patwari, P., Higgins, L. J., Chutkow, W. A., Yoshioka, J., & Lee, R. T. (2006). The interaction of thioredoxin with Txnip. Evidence for formation of a mixed disulfide by disulfide exchange. *J Biol Chem, 281*(31), 21884-21891. doi:10.1074/jbc.M600427200

Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., . . . Schier, A. F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Research, 22*(3), 577-591. doi:10.1101/gr.133009.111

Pei, T. M., Li, Y. J., Wang, J. B., Wang, H. L., Liang, Y. J., Shi, H. W., . . . Liu, L. X. (2015). YAP is a critical oncogene in human cholangiocarcinoma. *Oncotarget, 6*(19), 17206-17220. doi:DOI 10.18632/oncotarget.4043

Puca, L., & Brou, C. (2014). Alpha-arrestins - new players in Notch and GPCR signaling pathways in mammals. *J Cell Sci, 127*(Pt 7), 1359-1367. doi:10.1242/jcs.142539

Puca, L., Chastagner, P., Meas-Yedid, V., Israel, A., & Brou, C. (2013). Alpha-arrestin 1 (ARRDC1) and beta-arrestins cooperate to mediate Notch degradation in mammals. *J Cell Sci, 126*(Pt 19), 4457-4468. doi:10.1242/jcs.130500

Qayyum, N., Haseeb, M., Kim, M. S., & Choi, S. (2021). Role of Thioredoxin-Interacting Protein in Diseases and Its Therapeutic Outlook. *Int J Mol Sci, 22*(5). doi:10.3390/ijms22052754

Qin, A., Cheng, T. S., Pavlos, N. J., Lin, Z., Dai, K. R., & Zheng, M. H. (2012). V-ATPases in osteoclasts: structure, function and potential inhibitors of bone resorption. *Int J Biochem Cell Biol, 44*(9), 1422-1435. doi:10.1016/j.biocel.2012.05.014

Quinn, J. J., & Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet, 17*(1), 47-62. doi:10.1038/nrg.2015.10

Rauch, S., & Martin-Serrano, J. (2011). Multiple interactions between the ESCRT machinery and arrestin-related proteins: implications for PPXY-dependent budding. *J Virol, 85*(7), 3546-3556. doi:10.1128/JVI.02045-10

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res, 47*(W1), W191-W198. doi:10.1093/nar/gkz369

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res, 43*(7), e47. doi:10.1093/nar/gkv007

Rizki, T. M., & Rizki, R. M. (1992). Lamellocyte Differentiation in Drosophila Larvae Parasitized by Leptopilina. *Developmental and Comparative Immunology, 16*(2-3), 103-110. doi:Doi 10.1016/0145-305x(92)90011-Z

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. doi:10.1093/bioinformatics/btp616

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., . . . Carroll, J. S. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature, 481*(7381), 389-393. doi:10.1038/nature10730

Salmela, L., & Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics, 30*(24), 3506-3514. doi:10.1093/bioinformatics/btu538

Sang-Ho Yoon, B. C., Daewon Lee,Hanji Kim,Jiwon Shim,Jin-Wu Nam. (2023). Molecular Traces of Drosophila Hemocyte Evolution. doi:Available at SSRN: https://ssrn.com/abstract=4387698 or http://dx.doi.org/10.2139/ssrn.4387698

Saxena, G., Chen, J., & Shalev, A. (2010). Intracellular shuttling and mitochondrial function of thioredoxin-interacting protein. *J Biol Chem, 285*(6), 3997-4005. doi:10.1074/jbc.M109.034421

Schutte, U., Bisht, S., Heukamp, L. C., Kebschull, M., Florin, A., Haarmann, J., . . . Feldmann, G. (2014). Hippo Signaling Mediates Proliferation, Invasiveness, and Metastatic Potential of Clear Cell Renal Cell Carcinoma. *Translational Oncology, 7*(2), 309-321. doi:10.1016/j.tranon.2014.02.005

Sedlazeck, F. J., Lee, H., Darby, C. A., & Schatz, M. C. (2018). Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet, 19*(6), 329-346. doi:10.1038/s41576-018-0003-4

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research, 13*(11), 2498-2504. doi:10.1101/gr.1239303

Shea, F. F., Rowell, J. L., Li, Y., Chang, T. H., & Alvarez, C. E. (2012). Mammalian alpha arrestins link activated seven transmembrane receptors to Nedd4 family e3 ubiquitin ligases and interact with beta arrestins. *PLoS One, 7*(12), e50557. doi:10.1371/journal.pone.0050557

Shenoy, S. K., & Lefkowitz, R. J. (2011). beta-Arrestin-mediated receptor trafficking and signal transduction. *Trends Pharmacol Sci, 32*(9), 521-533. doi:10.1016/j.tips.2011.05.002

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M. M., Rosenbloom, K., . . . Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research, 15*(8), 1034-1050. doi:DOI 10.1101/gr.3715005

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313. doi:10.1093/bioinformatics/btu033

Steijger, T., Abril, J. F., Engstrom, P. G., Kokocinski, F., Consortium, R., Hubbard, T. J., . . . Bertone, P. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods, 10*(12), 1177-1184. doi:10.1038/nmeth.2714

Sterne-Weiler, T., Weatheritt, R. J., Best, A. J., Ha, K. C. H., & Blencowe, B. J. (2018). Efficient and Accurate Quantitative Profiling of Alternative Splicing Patterns of Any Complexity on a Laptop. *Molecular Cell, 72*(1), 187-+. doi:10.1016/j.molcel.2018.08.018

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., . . . von Mering, C. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res, 43*(Database issue), D447-452. doi:10.1093/nar/gku1003

Tang, A. D., Soulette, C. M., van Baren, M. J., Hart, K., Hrabeta-Robinson, E., Wu, C. J., & Brooks, A. N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications, 11*(1). doi:ARTN 1438

10.1038/s41467-020-15171-6

Tattikota, S. G., Cho, B., Liu, Y., Hu, Y., Barrera, V., Steinbaugh, M. J., . . . Perrimon, N. (2020). A single-cell survey of Drosophila blood. *Elife, 9*. doi:10.7554/eLife.54818

Teo, G., Liu, G., Zhang, J., Nesvizhskii, A. I., Gingras, A. C., & Choi, H. (2014). SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *J Proteomics, 100*, 37-43. doi:10.1016/j.jprot.2013.10.023

The UniProt, C. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res, 45*(D1), D158-D169. doi:10.1093/nar/gkw1099

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H. Y., Karlak, B., Daverman, R., . . . Narechania, A. (2003). PANTHER: A library of protein families and subfamilies indexed by function. *Genome Research, 13*(9), 2129-2141. doi:10.1101/gr.772403

Tian, X., Kang, D. S., & Benovic, J. L. (2014). beta-arrestins and G protein-coupled receptor trafficking. *Handb Exp Pharmacol, 219*, 173-186. doi:10.1007/978-3-642-41199-1_9

Toyomura, T., Murata, Y., Yamamoto, A., Oka, T., Sun-Wada, G. H., Wada, Y., & Futai, M. (2003). From lysosomes to the plasma membrane - Localization of vacuolar type H+-ATPase with the a3 isoform during osteoclast differentiation. *Journal of Biological Chemistry, 278*(24), 22023-22030. doi:10.1074/jbc.M302436200

Trimpert, C., Wesche, D., de Groot, T., Pimentel Rodriguez, M. M., Wong, V., van den Berg, D. T. M., . . . Deen, P. M. T. (2017). NDFIP allows NEDD4/NEDD4L-induced AQP2 ubiquitination and degradation. *PLoS One, 12*(9), e0183774. doi:10.1371/journal.pone.0183774

Tsubaki, H., Tooyama, I., & Walker, D. G. (2020). Thioredoxin-Interacting Protein (TXNIP) with Focus on Brain and Neurodegenerative Diseases. *Int J Mol Sci, 21*(24). doi:10.3390/ijms21249357

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., . . . Ponten, F. (2015). Proteomics. Tissue-based map of the human proteome. *Science, 347*(6220), 1260419. doi:10.1126/science.1260419

Ulitsky, I., & Bartel, D. P. (2013). lincRNAs: Genomics, Evolution, and Mechanisms. *Cell, 154*(1), 26-46. doi:10.1016/j.cell.2013.06.020

UniProt Consortium, T. (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Res, 46*(5), 2699. doi:10.1093/nar/gky092

Vidal, M., Cusick, M. E., & Barabasi, A. L. (2011). Interactome networks and human disease. *Cell, 144*(6), 986-998. doi:10.1016/j.cell.2011.02.016

Vijaymeena, M., & Kavitha, K. (2016). A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal, 3*(2), 19-28.

Vinayagam, A., Hu, Y., Kulkarni, M., Roesel, C., Sopko, R., Mohr, S. E., & Perrimon, N. (2013). Protein complex-based analysis framework for high-throughput data sets. *Sci Signal, 6*(264), rs5. doi:10.1126/scisignal.2003629

Vinayagam, A., Kulkarni, M. M., Sopko, R., Sun, X., Hu, Y., Nand, A., . . . Perrimon, N. (2016). An Integrative Analysis of the InR/PI3K/Akt Network Identifies the Dynamic Response to Insulin Signaling. *Cell Rep, 16*(11), 3062-3074. doi:10.1016/j.celrep.2016.08.029

Wang, K. C., & Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Molecular Cell, 43*(6), 904-914. doi:10.1016/j.molcel.2011.08.018

Wang, L., Park, H. J., Dasari, S., Wang, S. Q., Kocher, J. P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research, 41*(6). doi:ARTN e74

10.1093/nar/gkt006

Wang, Q. Y., Yu, J. J., Kadungure, T., Beyene, J., Zhang, H., & Lu, Q. (2018). ARMMs as a versatile platform for intracellular delivery of macromolecules. *Nature Communications, 9*. doi:ARTN 960

10.1038/s41467-018-03390-x

Wang, Y., Dong, Q. Z., Zhang, Q. F., Li, Z. X., Wang, E. H., & Qiu, X. S. (2010). Overexpression of yes-associated protein contributes to progression and poor prognosis of non-small-cell lung cancer. *Cancer Science, 101*(5), 1279-1285. doi:10.1111/j.1349-7006.2010.01511.x

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., . . . Morris, Q. (2010). The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res, 38*(Web Server issue), W214-220. doi:10.1093/nar/gkq537

Weber, J., Polo, S., & Maspero, E. (2019). HECT E3 Ligases: A Tale With Multiple Facets. *Front Physiol, 10*, 370. doi:10.3389/fphys.2019.00370

Wilden, U., Wust, E., Weyand, I., & Kuhn, H. (1986). Rapid affinity purification of retinal arrestin (48 kDa protein) via its light-dependent binding to phosphorylated rhodopsin. *FEBS Lett, 207*(2), 292-295. doi:10.1016/0014-5793(86)81507-1

Wondafrash, D. Z., Nire'a, A. T., Tafere, G. G., Desta, D. M., Berhe, D. A., & Zewdie, K. A. (2020). Thioredoxin-Interacting Protein as a Novel Potential Therapeutic Target in Diabetes Mellitus and Its Underlying Complications. *Diabetes Metab Syndr Obes, 13*, 43-51. doi:10.2147/DMSO.S232221

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., . . . Timp, W. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods, 16*(12), 1297-+. doi:10.1038/s41592-019-0617-2

Wu, N., Zheng, B., Shaywitz, A., Dagon, Y., Tower, C., Bellinger, G., . . . McGraw, T. E. (2013). AMPK-dependent degradation of TXNIP upon energy stress leads to enhanced glucose uptake via GLUT1. *Molecular Cell, 49*(6), 1167-1175.

Xiao, J., Shi, Q., Li, W., Mu, X., Peng, J., Li, M., . . . Fan, J. (2018). ARRDC1 and ARRDC3 act as tumor suppressors in renal cell carcinoma by facilitating YAP1 degradation. *Am J Cancer Res, 8*(1), 132-143.

Xiao, W., Adhikari, S., Dahal, U., Chen, Y. S., Hao, Y. J., Sun, B. F., . . . Yang, Y. G. (2016). Nuclear m(6)A Reader YTHDC1 Regulates mRNA Splicing. *Molecular Cell, 61*(4), 507-519. doi:10.1016/j.molcel.2016.01.012

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., . . . Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics, 21*(5), 650-659. doi:10.1093/bioinformatics/bti042

You, B. H., Yoon, S. H., & Nam, J. W. (2017). High-confidence coding and noncoding transcriptome maps. *Genome Research, 27*(6), 1050-1062. doi:10.1101/gr.214288.116

Yu, G., Wang, L. G., & He, Q. Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics, 31*(14), 2382-2383. doi:10.1093/bioinformatics/btv145

Zbieralski, K., & Wawrzycka, D. (2022). alpha-Arrestins and Their Functions: From Yeast to Human Health. *Int J Mol Sci, 23*(9). doi:10.3390/ijms23094988

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., . . . Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol, 9*(9), R137. doi:10.1186/gb-2008-9-9-r137

Zhi, X., Zhao, D., Zhou, Z. M., Liu, R., & Chen, C. S. (2012). YAP Promotes Breast Cell Proliferation and Survival Partially through Stabilizing the KLF5 Transcription Factor. *American Journal of Pathology, 180*(6), 2452-2461. doi:10.1016/j.ajpath.2012.02.025

Zhou, R., Tardivel, A., Thorens, B., Choi, I., & Tschopp, J. (2010). Thioredoxin-interacting protein links oxidative stress to inflammasome activation. *Nat Immunol, 11*(2), 136-140. doi:10.1038/ni.1831

Zuckerman, R., & Cheasty, J. E. (1986). A 48 kDa protein arrests cGMP phosphodiesterase activation in retinal rod disk membranes. *FEBS Lett, 207*(1), 35-41. doi:10.1016/0014-5793(86)80008-4

Authors contributions

Kyung-Tae Lee performed computational analyses; Inez K.A. Pranoto from University of Washington performed all AP/MS experiments; Soon-Young Kim and Jung-Eun Kim from Kyungpook National University contributed to ARRDC5-related functional study, Hee-Joo Choi, Ngoc Bao To, and Jeong-Yeon Lee from Hanyang University performed or contributed to ChIP, ATAC-seq and RNA-seq related experiments; Young Kwon from University of Washington contributed to writing the manuscript and supervised the first project related with α-arrestin. Bumsik Cho and Jiwon Shim from Hanyang University contributed to Nanopore cDNA and Illumina sequencing, and Vupd1 related experiments. Sang-Ho Yoon contributed to computational analysis of scRNA-seq data. PPO2-CG13743 fusion gene was discovered and analyzed by DaeWon Lee from Hanyang University. Jin-Wu Nam from Hanyang University, as a principal investigator, supervised and contributed to writing of all projects specified in this manuscript.

국문 요지

초파리 및 인간 유래 α-arrestin 의 단백질 상호작용 네트워크 분석과 초파리 애벌레 발달 및 면역 반응에서의 전사체 다이내믹스 연구

한양대학교 대학원
자연과학대학 생명과학과
이경태

본 연구에서는 대량 신속 처리된 multi omics 데이터를 활용하여 arrestin 패밀리 단백질 중 하나인 α-arrestin 에 대한 연구를 인간과 초파리에서 진행하고 초파리 애벌레 발달 과정 및 면역 반응에서 생기는 hemocyte 의 전사체 다이내믹스를 연구하였음. α-arrestin 은 다목적 단백질로서 여러 신호 전달 체계, 특히 G-protein coupled receptor 를 조절하는 것으로 보고됨. 몇몇 α-arrestin 에 대한 연구가 진행됐지만 제한적인 부분에서만 깊게 연구 돼있음. 해당 단백질은 여러 종에서 보존 돼있기 때문에 보존돼 있는, 혹은 종 특이적인 α-arrestin 의 기능 연구는 이들에 대한 이해를 높일 것이고 다양한 신호 전달 체계를 조절하고 여러 질병과 연관돼 있는 만큼 이들에 대한 통합적인 연구는 기초 연구로서뿐만 아니라 치료학 부분으로 까지 응용이 가능할 것으로 기대됨. 이를 위해 대량 신속 처리된 α-arrestin 의 affinity purification 후 질량 분광 분석법을 통해 생산된 데이터를 이용해 상호작용하는

단백질을 동정하였고 전산학적 및 통계적 분석을 통해 신뢰도 높은 단백질 상호 작용 네트워크를 구성하였음.

인간과 초파리에서의 α-arrestin 상호 작용 네트워크 비교 분석을 통해 보존돼 있는 기능들을 확인하였고, 기존에 알려져 있던 ubiquitination 에 의한 단백질 분해 기작 뿐만 아니라 RNA splicing 과 같은 새로운 기능이 두 종 모두에서 보존 돼있다는 것을 확인하였음. 이중 인간 유래 α-arrestin 중 하나인 ARRDC3 가 RNA splicing 과 연관 돼있음을 확인하였고 이를 public data 와 본 연구진에서 생산한 대량 신속 처리된 RNA sequencing 데이터로 간접적이지만 연관성을 검증하였음.

다음으로 인간 특이적인 기능에 대한 검증을 진행하였음. 인간 유래 α-arrestin 중 하나인 TXNIP과 histone deacetylase 복합체와의 상호작용 검증 및 연관성을 조사하기 위해 대량 신속 처리된 Assay for transposase-accessible chromatin using sequencing (ATAC-seq) 및 RNA-seq 데이터를 생산하였고, TXNIP 의 유전자 발현이 억제됐을 때 상당 수 유전자의 발현이 억제되고 프로모터 부분의 chromatin accessibility 가 감소하는 것을 확인하였으며, CD22 과 L1CAM 두 유전자의 프로모터에서 HDAC2 단백질의 결합이 유의미하게 증가함을 밝혀 냈음. 또다른 인간 유래 α-arrestin 인 ARRDC5 는 V-type ATPase 들과 강하게 상호작용하는 것을 확인하였고, 이들이 중요한 역할을 하는 마우스 유래 osteoclast 세포의 분화와 뼈 재흡수 기능 및 plasma membrane 으로 의 이동을 유도한다는 것을 실험적으로 검증하였음.

초파리 유래 α-arrestin 중 하나인 Vdup1 은 초파리 애벌레에서 발현하는 전구 세포에서 강하게 발현함을 확인하였고 RNA splicing 복합체와 상호작용하고 있음을 확인하였기 때문에 해당 모델 생물에서 전사체 다이내믹스 연구를 진행함. 또한 기존 선행 연구에서 초파리 hemocyte (초파리 혈액 세포) 중 하나인 lamellocyte 가 면역 반응 (말벌에 의한 감염)에 의해 급격히 증가할 때 일부 비 번역 RNA (lncRNA)를 강하게 발현함을 확인하였음. 알려져 있는 비 번역 RNA 뿐만 아니라 추가로

알려지지 않은 비 번역 RNA 들이 면역 반응에서 증가하는 lamellocyte 의 기능과 연관돼 있을 것이라는 가설을 세웠고 이를 검증하고자 하였음. 분석을 위해 차세대 염기 서열 분석 (next-generation sequencing) 데이터와 3 세대 염기 서열 분석 (3rd generation sequencing) 데이터를 동시에 활용하는 하이브리드 전사체 구축 파이프라인을 고안하였고 두 분석 기술의 단점을 보완하고 장점을 극대화함으로써 보다 정확한 transcript 구조를 예측할 수 있었음. 해당 파이프라인을 통해 업데이트 된 유전지 모델을 바탕으로 단일 세포 레벨에서 다양한 세포 타입 특이적인 lncRNA 마커를 발굴할 수 있었고 현재 Lamellocyte lncRNA 마커에 대한 발현 및 기능을 실험적으로 검증하고 있음.

3 세대 염기서열 분석은 RNA 의 전장 분석이 가능하다는 장점이 있고 이를 활용하여 이소체 수준에서의 차등 발현 분석 및 fusion gene 분석을 진행하였음. 말벌 감염 상태에서 혈액 유래 hemocyte 간의 이소체 변화가 가장 크게 나타난 것을 확인하였고 이들의 패턴과 일부 후보군의 검증을 진행하고 있음. 마지막으로 fusion gene 분석을 통해 2 개 이상의 유전자가 fusion 되는 현상을 확인하였고 약 30 개정도의 fusion gene 을 검출할 수 있었음. 특히 Prophenoloxidase 2 와 CG13743 유전자와의 fusion 현상이 모든 샘플에서 높은 빈도로 관측됐고 이를 포함해 5 개 fusion gene 에 대해 Polymerase chain reaction 을 이용한 실험적 검증을 진행하고 있음.

정리하면 대량 신속 처리된 multi-omics 데이터를 활용해 α-arrestin 단백질의 상호작용 네트워크 구축 및 진화적 관점에서의 보존 및 분화된 기능에 대한 분석을 진행하였고, 초파리 애벌레의 면역 반응에 의해 변화하는 전사체 다이내믹스를 초파리 면역세포에서 체계적으로 분석하였고 fusion gene, 이소체 차등 발현 및 lncRNA 와의 연관성을 연구하였음.

감사의 글

2015 년에 처음으로 남진우 교수님의 생물정보 및 유전체 연구실에 들어와 학부 인턴 2 년, 석박 통합 6 년이라는 기간 동안 공부하였습니다. 방황하던 시기에 저를 받아 주시고, 저의 가능성과 잠재력을 이끌어내 주시고, 연구 뿐만 아니라 여러 부분에서 저의 스승으로서 훌륭한 지도를 해 주신 남진우 교수님께 정말 감사드립니다. 남진우 교수님과 생활하며 정말 많은 것을 배웠고 앞으로 나아가야 할 방향에 대해 많은 도움을 받을 수 있었습니다. 긴 연구실 생활동안 항상 같은 위치에서 저를 이끌어 주신 남진우 교수님께 다시 한번 감사드립니다. 지금은 졸업하여 국내 회사 및 해외 연구자로서 활동하고 있는 연구실 선배, 후배 및 동기인 이현주 누나, 남경우 형, 홍효선, 서좌원 형, 유보현 형, 손장일 박사님, Vipin menon, 표성진, 최우희, 이은경, 김한지 에게도 감사의 말을 전합니다. 연구실 생황을 하는 동안 많은 연구 관련 조언을 해주었고 힘든 시간을 이겨 내주는 활력소가 돼 주었습니다. 현재 연구실에 같이 생활하고 있는 박석주, 윤상호, 최서원, 최민학 형, 김가영, 이승은, 김지훈, 송현석, 김솔빈, 한준섭, 이동은 선생님, 나한솔, 이도헌, 정가영 선생님, Ngoc Bao To, 왕가문, 이민욱 에게도 감사의 말을 전합니다. 연구실 일원들과 함께 일할 수 있어서 영광이었고 앞으로 모두 하고자 하는 일 이루어지기를 기원합니다.

석박 통합 기간에 진행한 첫 번째 연구 주제인 alpha-arrestin 연구의 아이디어 및 초기 데이터를 제공해주시고 1 년동안의 해외 연구자 경험 제공 및 연구 기간동안 훌륭한 연구지도를 해주셨으며 학위 심사위원을 맡아 주신 University of Washington 의 권영 교수님과 최근 박사학위를 취득한 Inez K. Pranoto 에게 감사의 말씀을 전합니다. 함께 연구할 수 있어서 영광이었고 연구가 잘 마무리되기 위해 열심히 하도록 하겠습니다. alpha-arrestin 연구의 실험적 검증을 맡아 주시고 학위

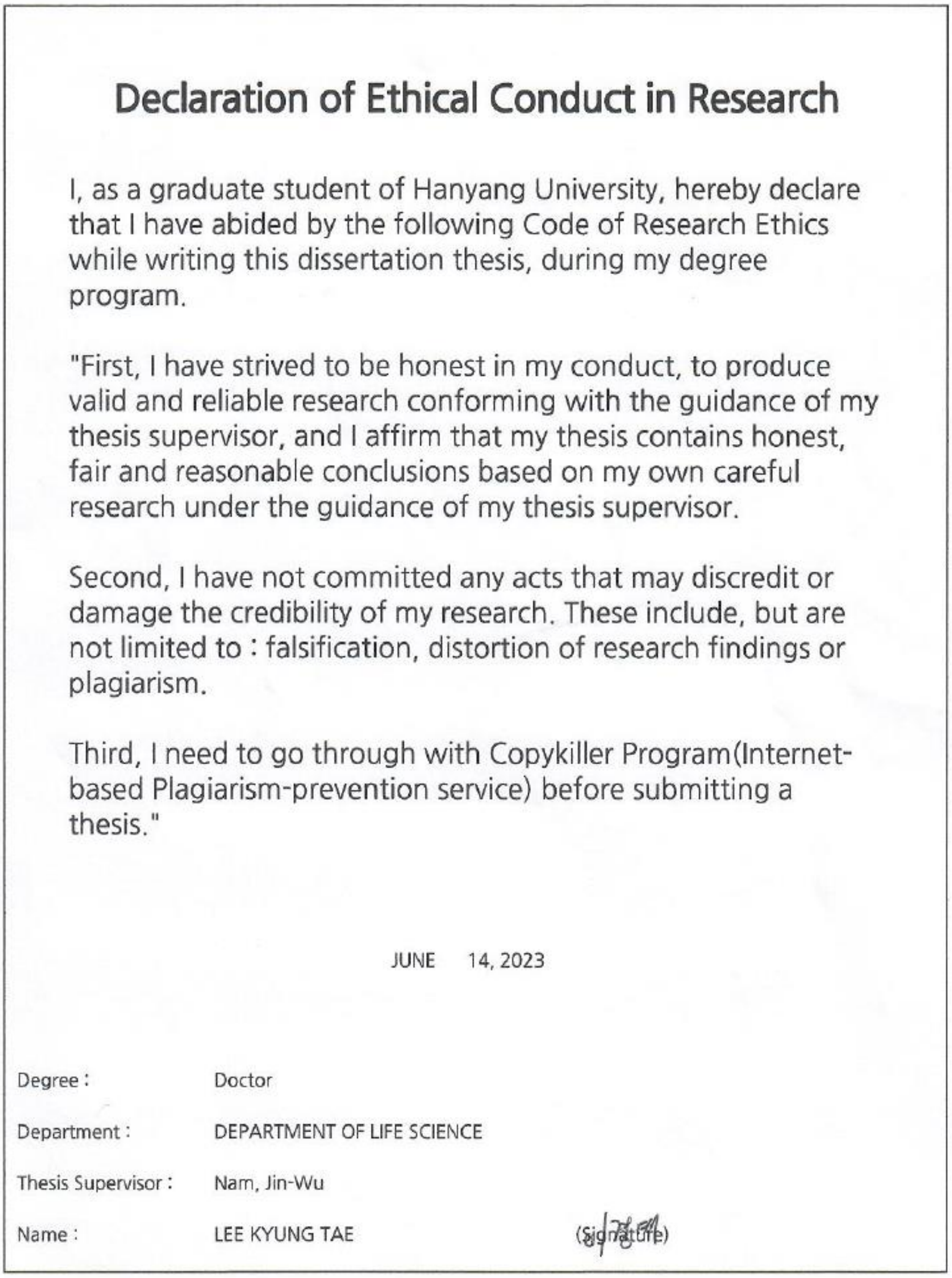
심사 위원을 맡아 주신 한양대학교 이정연 교수님과 최희주 박사님께도 감사의 말씀을 전합니다. 해당 연구 결과들은 alpha-arrestin 연구의 향상에 지대한 영향을 주었습니다. alpha-arrestin 연구에서 ARRDC5 의 실험 분석을 진행해주신 경북대학교 김정은 교수님과 김수영 연구원님께도 감사의 말씀을 전합니다.

두 번째 연구 주제인 초파리 larvae 의 lncRNA 연구를 진행해주시고 학위 심사위원장을 맡아 주신 한양대학교 심지원 교수님과 연구실 일원인 조범식, 이대원, 윤성규, 그리고 연구를 같이 진행해보진 못하였지만 항상 응원해주었던 신민규, 차누리 에게도 감사의 말씀을 전합니다. 아직 연구가 더 진행되야 할 부분이 있지만 최선을 다해 연구가 좋은 성과를 맺도록 노력하겠습니다. 그리고 심사위원을 맡아 주신 한양대학교 IBB 정효빈 교수님께도 감사의 말씀을 드립니다. 학위 심사 동안 주신 소중한 의견 및 조언을 잘 반영하여 좋은 연구하도록 하겠습니다.

오랜 기간동안 저를 응원해주시고 지원해주신 가족들에게도 감사의 말씀을 전합니다. 어머니 아버지의 도움 없이는 결코 박사 과정을 헤쳐 나가지 못했을 것입니다. 방황하던 저를 믿어 주시고 끝까지 응원해 주셔서 정말 감사드리고 사랑합니다. 항상 관심과 지원을 아끼지 않았던 누나, 매형에게도 감사의 말씀을 전합니다. 저의 앞날을 항상 응원해주시고 기다려 주신 할머니, 할아버지께도 감사의 말씀을 전합니다. 누구보다 저의 박사 학위 취득을 기뻐해 주신 만큼 앞으로도 더 열심히 인생을 살아가도록 하겠습니다. 아직 많이 부족한 사위를 받아 주시고 믿어 주시며 응원해주신 장모님, 장인 어른께도 감사의 말씀드립니다. 말씀해주신 대로 빠른 속도보다는 올바른 방향으로 인생을 살아가도록 하겠습니다. 지면에 언급하지는 못했지만 항상 저의 앞길을 응원해주신 친인척, 친구들에게도 감사의 말씀 전합니다. 받은 사랑과 응원을 보답하며 사는 사람이 되도록 하겠습니다.

마지막으로 박사 학위 기간동안 여자친구로서, 그리고 지금은 아내로서 저를 가장 많이 사랑해주고 응원해주며 물심양면 지원해준 사랑하는 한을 이에게 감사의

말을 전합니다. 불확실한 저의 미래를 믿어주고 저의 가장 큰 버팀목이 된 지금의 아내가 없었다면 지금의 저도 없었을 것입니다. 앞으로 더 많이 사랑하고 서로 의지하며 살도록 하겠습니다.



Declaration of Ethical Conduct in Research

I, as a graduate student of Hanyang University, hereby declare that I have abided by the following Code of Research Ethics while writing this dissertation thesis, during my degree program.

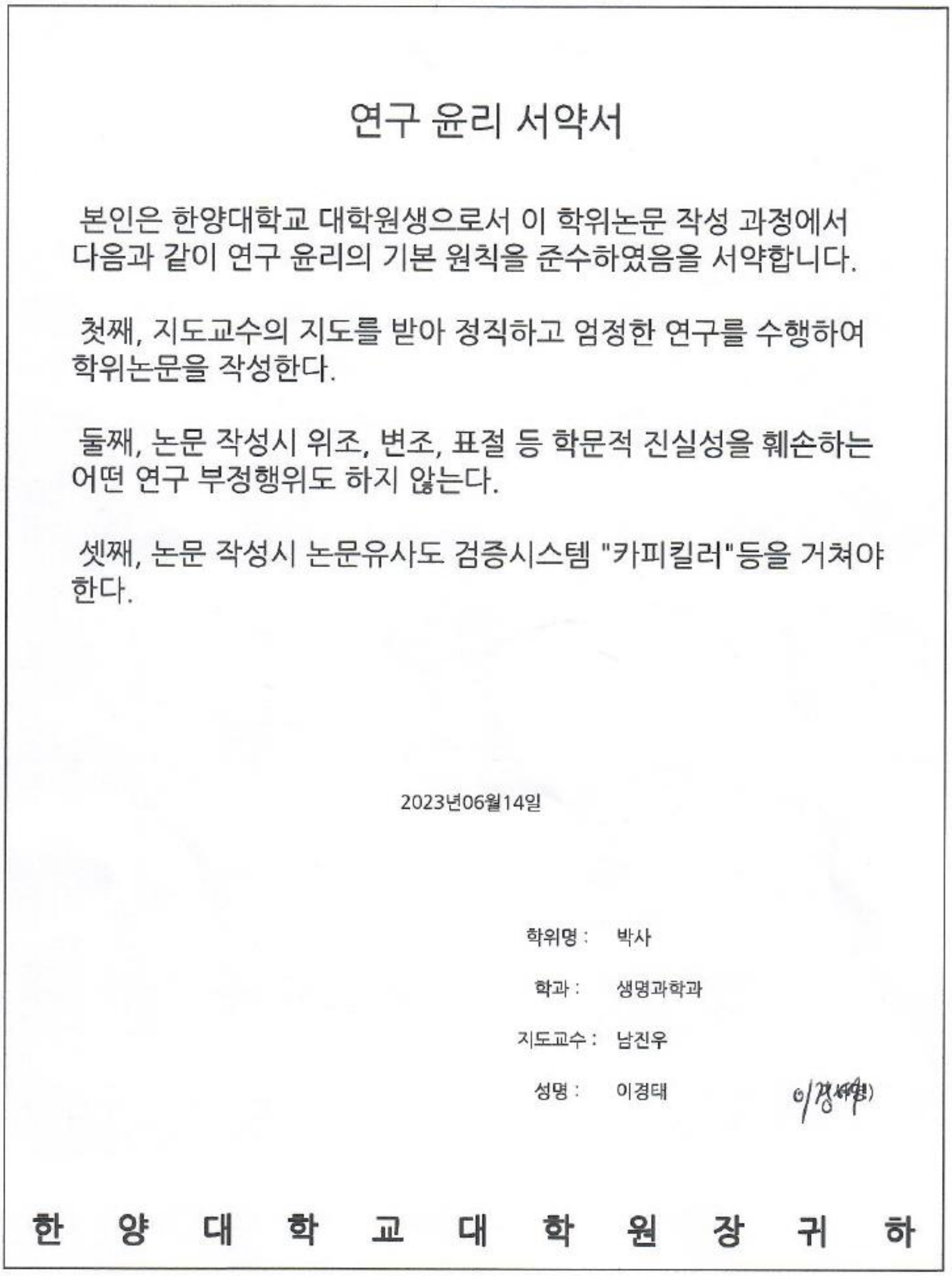
"First, I have strived to be honest in my conduct, to produce valid and reliable research conforming with the guidance of my thesis supervisor, and I affirm that my thesis contains honest, fair and reasonable conclusions based on my own careful research under the guidance of my thesis supervisor.

Second, I have not committed any acts that may discredit or damage the credibility of my research. These include, but are not limited to : falsification, distortion of research findings or plagiarism.

Third, I need to go through with Copykiller Program(Internet-based Plagiarism-prevention service) before submitting a thesis."

JUNE 14, 2023

Degree :	Doctor	
Department :	DEPARTMENT OF LIFE SCIENCE	
Thesis Supervisor :	Nam, Jin-Wu	
Name :	LEE KYUNG TAE	(Signature)



연구 윤리 서약서

본인은 한양대학교 대학원생으로서 이 학위논문 작성 과정에서 다음과 같이 연구 윤리의 기본 원칙을 준수하였음을 서약합니다.

첫째, 지도교수의 지도를 받아 정직하고 엄정한 연구를 수행하여 학위논문을 작성한다.

둘째, 논문 작성시 위조, 변조, 표절 등 학문적 진실성을 훼손하는 어떤 연구 부정행위도 하지 않는다.

셋째, 논문 작성시 논문유사도 검증시스템 "카피킬러"등을 거쳐야 한다.

2023년06월14일

학위명 : 박사

학과 : 생명과학과

지도교수 : 남진우

성명 : 이경태 이경(서명)

한 양 대 학 교 대 학 원 장 귀 하