The Company of Biologists

TECHNIQUES AND RESOURCES   **RESEARCH ARTICLE**

Identification and classification of cis-regulatory elements in the amphipod crustacean *Parhyale hawaiensis*

Dennis A. Sun[1],*, Jessen V. Bredeson[1], Heather S. Bruce[2] and Nipam H. Patel[2,3],*

ABSTRACT

Emerging research organisms enable the study of biology that cannot be addressed using classical 'model' organisms. New data resources can accelerate research in such animals. Here, we present new functional genomic resources for the amphipod crustacean *Parhyale hawaiensis*, facilitating the exploration of gene regulatory evolution using this emerging research organism. We use Omni-ATAC-seq to identify accessible chromatin genome-wide across a broad time course of *Parhyale* embryonic development. This time course encompasses many major morphological events, including segmentation, body regionalization, gut morphogenesis and limb development. In addition, we use short- and long-read RNA-seq to generate an improved *Parhyale* genome annotation, enabling deeper classification of identified regulatory elements. We discover differential accessibility, predict nucleosome positioning, infer transcription factor binding, cluster peaks based on accessibility dynamics, classify biological functions and correlate gene expression with accessibility. Using a Minos transposase reporter system, we demonstrate the potential to identify novel regulatory elements using this approach. This work provides a platform for the identification of novel developmental regulatory elements in *Parhyale*, and offers a framework for performing such experiments in other emerging research organisms.

KEY WORDS: ATAC-seq, *Parhyale hawaiensis*, RNA-seq, Evolution, Functional genomics

INTRODUCTION

Advances in genomic techniques have facilitated genetic research in emerging research organisms. Genome sequencing has become substantially less expensive over time (https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data), enabling researchers to study genome composition. Moreover, the establishment of genetic perturbation techniques, such as RNAi and CRISPR-Cas9 mutagenesis, has enabled the rapid characterization of gene function in a range of different research organisms. RNAi knockdown has been applied to various insects (Christian et al., 2015; Mito et al., 2011), chelicerates (Sharma et al., 2013), flatworms (Rouhana et al., 2013; Srivastava et al., 2014) and numerous other organisms to study gene function. More recently, CRISPR-Cas9 mutagenesis has enabled both targeted ablation and targeted transgenesis in organisms as diverse as cephalopods (Crawford et al., 2020) and reptiles (Rasys et al., 2019). Using these tools, it is now possible to identify and perturb the function of genes across diverse organisms to understand how genes and genomes evolve across the tree of life.

Although the study of gene function in diverse systems has grown more tractable, the study of gene regulation has historically proven difficult even in model organisms. Unlike protein-coding genes or noncoding RNAs, which can often be predicted with gene sequence alone, cis-regulatory elements (CREs) are composed of complex assemblies of transcription factor binding sites which can be computationally challenging to identify (Li et al., 2015; Wasserman and Sandelin, 2004). These sites can change dramatically over short periods of time; for example, in the context of the *even skipped* stripe 2 enhancer across several *Drosophila* species, the position of crucial transcription factor binding sites shifted in as short a span as 1-2 million years divergence (Ludwig et al., 1998). Such rapid evolutionary flexibility, although important for driving short- and long-term evolutionary processes, makes the identification of such elements in novel research organisms a considerable challenge.

In recent years, novel techniques for identifying CREs have been developed that rely on the fact that CREs often occur in regions of accessible chromatin. The earliest, including DNAse I hypersensitivity and Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq) enabled researchers to locate accessible chromatin regions and discover novel regulatory elements in a range of research organisms (Giresi et al., 2007; Lai et al., 2018; Pérez-Zamorano et al., 2017). However, such techniques were hampered by low signal-to-noise ratios, the requirement for very large amounts of tissue and technically challenging protocols. In 2015, Buenrostro et al. published papers describing the Assay for Transposase-Accessible Chromatin with next-generation sequencing (ATAC-seq), a novel technique using a hyperactive version of the Tn5 transposase, as a way of identifying accessible chromatin regions genome-wide (Buenrostro et al., 2015). In this paper, they demonstrated that ATAC-seq performed comparably with previous accessibility techniques, with reduced tissue requirements and increased convenience and speed.

ATAC-seq has proved to be an effective technique for traditional model systems, and has offered a true revolution for emerging model systems. The low material input requirements of ATAC-seq are of particular use in emerging research organisms, for which generating the millions of cells required for previous techniques would prove challenging. Researchers have applied ATAC-seq to gar embryos in order to study the evolution of limb development (Gehrke et al., 2015) and acoel flatworms and hydra to study regeneration (Cazet et al., 2021; Gehrke et al., 2019), among numerous other examples. Thus, ATAC-seq is now established as a generalizable tool that is particularly well-suited to accelerating work in emerging research organisms.

[1]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. [2]Marine Biological Laboratory, Woods Hole, MA 02543, USA. [3]Department of Organismal Biology and Anatomy, University of Chicago, Chicago, IL 60637, USA.

*Authors for correspondence (dennis.a.sun@berkeley.edu; npatel@mbl.edu)

D.A.S., 0000-0003-1000-7276; J.V.B., 0000-0001-5489-8512; H.S.B., 0000-0003-3199-8224; N.H.P., 0000-0003-4328-654X

DEVELOPMENT

In 2017, Corces et al. reported a further-improved version of ATAC-seq, Omni-ATAC-seq, which provides several improvements to the standard ATAC-seq protocol (Corces et al., 2017). In particular, it has been reported to have a greater signal-to-noise ratio, decreased mitochondrial read contamination and applicability to fixed and frozen tissues, with only minor modifications to the buffers and detergents used in the standard ATAC-seq protocol.

In this paper, we apply the Omni-ATAC-seq protocol to embryos of the amphipod crustacean *Parhyale hawaiensis*, an emerging research organism for the study of arthropod development, evolution and regeneration (Paris et al., 2021 preprint; Stamataki and Pavlopoulos, 2016; Sun and Patel, 2019). We identify dynamic regions of open chromatin ('peaks') across a broad swath of *Parhyale* developmental time. We comprehensively analyze our Omni-ATAC-seq data to predict the position of nucleosomes along the genome and infer the footprints of transcription factors bound to peaks. Using fuzzy clustering (Kumar and E Futschik, 2007), we partition our peaks into groups based on similar accessibility trajectories, revealing groups of peaks with different transcription factor footprint enrichment and nucleosome occupancy that are active at different points in development. In addition, we use short- and long-read RNA-seq to improve the *Parhyale* genome annotation and investigate the relationship between accessibility and gene expression over time during development.

Parhyale has served as a platform for foundational discoveries about such processes as body plan evolution (Martin et al., 2016), the evolution of arthropod limbs (Bruce and Patel, 2020) and the evolution of regeneration (Konstantinides and Averof, 2014). By facilitating the identification of regulatory elements in this research organism, our work will enable other researchers to investigate the complexities of gene regulatory evolution in these processes and others. Furthermore, by enabling the assessment of cis-regulatory elements in *Parhyale*, we open avenues to investigate fundamental mechanisms of gene regulation in the understudied non-insect crustacean clade.

Although this work primarily focuses on *Parhyale*, our methods can provide examples of how one can perform thorough analyses of ATAC-seq and RNA-seq data using existing tools to generate hypotheses about gene expression dynamics and regulatory element function. Such approaches can be applied to a diverse range of organisms, and will facilitate deeper understanding of gene regulation across the tree of life.

RESULTS

Omni-ATAC-seq identifies open chromatin across *Parhyale* developmental stages

To identify developmental regulatory elements, we performed Omni-ATAC-seq on duplicate libraries of 15 stages of *Parhyale* embryonic development (Fig. 1A-C; Fig. S1A). We evaluated the quality of our libraries using a variety of standard tests, such as fragment size analysis and enrichment of Omni-ATAC signal at promoters (see 'Omni-ATAC sequencing quality control' in Materials and Methods and Fig. 2A,B). Our analyses indicated that Omni-ATAC-seq performed as expected in identifying regulatory elements genome-wide, with low mitochondrial read contamination.

To assess whether our libraries were capable of capturing significant variation over time, we performed principal component analysis (PCA) on our libraries (Fig. 2C). PCA revealed that our libraries were primarily separated along two principal components (PCs), PC1 (60% of variation) and PC2 (18% of variation), with a considerable drop in variation explained in other PCs (Fig. S2D). PC1 appeared to be associated with developmental time, with earlier

developmental stage libraries showing a negative loading and later developmental stages showing a positive loading. Within the PCA plots, samples from the S21 and S22 time points appeared to be more separated in PC2 relative to the other time points. A discussion of this observation and its implications can be found in the supplementary Materials and Methods and Figs S3 and S4C.

We used Genrich (https://github.com/jsh58/Genrich) to call significant peaks (merged *q*-value between replicates<0.05) in each of our 15 time points independently, and merged overlapping peaks across time points using bedtools (Quinlan and Hall, 2010), yielding 190,078 genomic regions which we used as our 'peaks' in downstream analyses, including the ImpulseDE2 (IDE2) pipeline. The dynamic accessibility of these peaks is illustrated in a heatmap in Fig. 2D. Of these peaks, 163,227 (85.87%) were classified as having statistically significant variation (padj<0.01) over our time-course; 60,909 (37.32%) were classified as having transient expression dynamics; and 88,231 (54.05%) were classified as showing positively or negatively monotonic expression dynamics (Fig. 2E). These results indicate that we were able to identify many dynamically accessible regions across the *Parhyale* genome.

To identify regions of dynamically accessible chromatin, we used the IDE2 pipeline (Fischer et al., 2018) (see Materials and Methods for a brief summary of the software's advantages). IDE2 produces a fitted model of accessibility for each of the peaks used in the analysis (Fischer et al., 2018) (see Fig. S4C,D for examples). We used these model fits to summarize the global properties of the dynamic peaks we identified. We observed that our model fits tended to achieve maximum accessibility at early, middle and late developmental stages (Fig. 2F). These time points also appeared to be associated with peaks that showed strong loading in PCA (Fig. S2G), indicating that early, middle and late time points are the primary drivers of variation in our dataset.

These global analysis results indicate that our Omni-ATAC-seq experiments captured information normally found in ATAC-seq data, including strong enrichment at promoters. Overall, the low mitochondrial read contamination, large library size and low fraction of duplicated reads suggest that our Omni-ATAC-seq data are of high quality. In addition, the results of the IDE2 differential accessibility analyses indicate that the vast majority of accessible regions in the *Parhyale* genome show dynamic accessibility over developmental time.

Improving the *Parhyale* genome annotation using short- and long-read RNA-seq

Although Omni-ATAC-seq signal showed enrichment across annotated mRNA starts in a genome-wide analysis, careful examination of individual genes and gene models indicated that many of the MAKER gene annotations are fragmented (see Fig. S5A). A more accurate genome annotation would improve both genome-wide analyses and enable the precise classification of candidate regulatory elements as promoters, exonic and intronic regulatory elements, and intergenic regulatory elements. To generate a more complete genome annotation, we performed RNA-seq using two approaches: short-read sequencing using the Illumina NovaSeq platform and long-read sequencing using Oxford Nanopore technology. For four developmental stages (S13, S19, S21 and S23), we generated triplicate RNA-seq libraries (Fig. 3A,B; representative embryo images in Fig. S1B).

We assembled multiple transcriptomes from each of the sequencing approaches (Fig. 3B; see Materials and Methods). To generate a more complete genome annotation, we used the Mikado pipeline to merge our assembled transcriptomes with
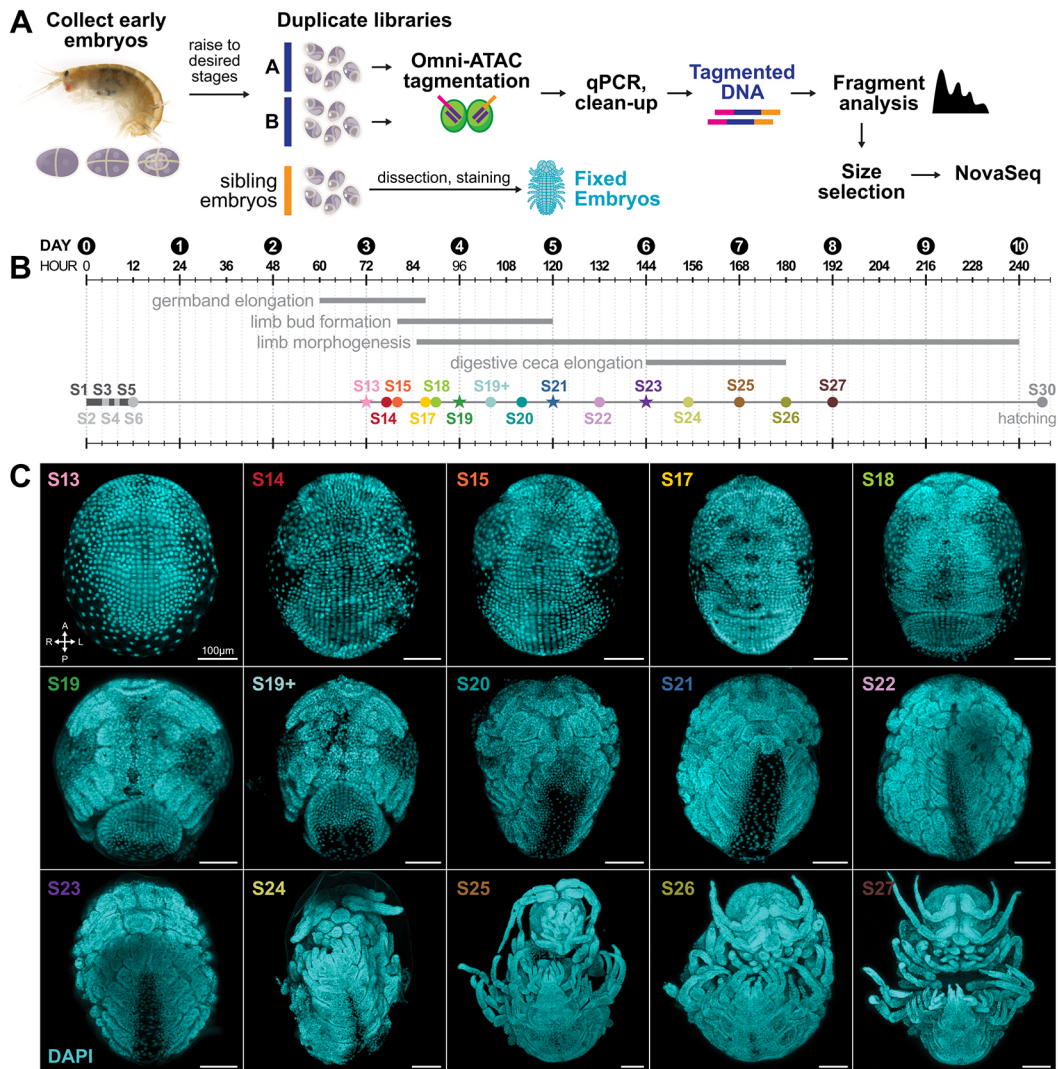
Fig. 1. Time-course ATAC-seq in *P. hawaiensis* embryos. (A) Overview of ATAC-seq protocol. Embryos were collected at early developmental stages and raised to specific developmental time points. Duplicate libraries were generated by tagmenting five embryos each from a single clutch of sibling embryos. qPCR analysis was used to determine the optimum number of additional cycles of PCR. Tagmented DNA was cleaned and fragment analysis was performed to assess library quality. Final libraries were pooled in equal concentrations, size-selected, and sequence on an Illumina NovaSeq short read sequencer. (B) Timeline of developmental stages. RNA-seq libraries were also generated for time points marked with a star. (C) Representative embryo images from clutches used for Omni-ATAC-seq. Embryos are stained with DAPI and mounted ventral-side up. A, anterior; P, posterior; L, left; R, right. Scale bars: 100 µm.

the previously-generated Kao et al. (2016) transcriptome and the MAKER genome annotation. The Mikado pipeline and others (e.g. EvidentialGene) use a variety of metrics to compare transcripts from different sources to determine the 'best' gene models for each gene region in the genome (Gilbert, 2019 preprint; Venturini et al., 2018). We evaluated the Mikado-merged transcriptome (hereafter 'Mikado transcriptome') using BUSCO, and observed a score comparable with the best of the individual transcriptomes (90.9% complete), and a marked improvement compared with the MAKER annotation (80.9% complete) (Fig. 3C).

To further evaluate the completeness of the genome annotation, we examined individual genes and loci of interest. Using previous Rapid Amplification of cDNA Ends (RACE) data, we observed that the Mikado transcriptome has a more complete Hox cluster than the MAKER annotation (Fig. S5A-C), including a more complete *Hox3* gene than has been previously reported. To more comprehensively evaluate gene model completeness between transcriptomes, we generated a custom script to compare transcriptome data

with previous RACE data for additional developmental genes (see Materials and Methods). When we compared the completeness measures between the Mikado transcriptome and the MAKER genome annotation, we observed a marked improvement. Notably, the Mikado transcriptome had a greater proportion of RACE genes with promoter-peak overlap (Mikado 0.84>MAKER 0.71), as well as a greater fraction of 'single' transcripts (Mikado 0.73>MAKER 0.55) (Fig. 3D).

To further evaluate the quality of the Mikado transcriptome as a reference, we compared its performance with the MAKER annotation using a variety of metrics (Fig. 3E). The Mikado transcriptome produced a larger number of total gene models (n=21,218) than are found in the MAKER annotation (n=15,105). Using bedtools, we attempted to assign a candidate promoter peak to each transcript in each genome annotation (see Materials and Methods) and observed that the Mikado transcriptome had a larger number of transcripts with an assigned promoter (Mikado n=18,946; MAKER n=11,909), further suggesting improved genome-wide
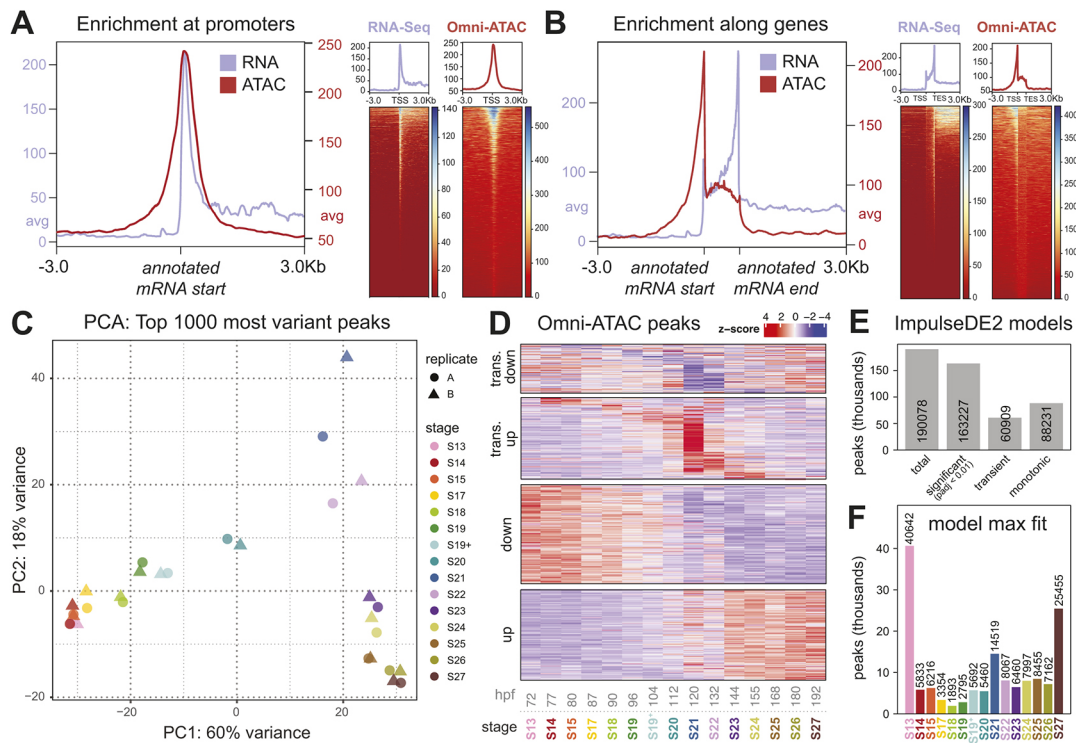
DEVELOPMENT

3

Fig. 2. Whole-genome analysis of ATAC-seq reveals enrichment of promoter signals, dynamic accessibility over developmental time. (A) Omni-ATAC and RNA-seq signal enrichment at promoters. Annotated mRNA start positions are aligned and average signal across genomic positions is plotted for either Omni-ATAC or RNA-seq signal. Axes reflect the average signal of RNA-seq or Omni-ATAC reads across all mRNA start positions in the genome. A strong enrichment of RNA-seq signal is observed 3′ of mRNA start sites, whereas enrichment of Omni-ATAC-seq signal appears to be symmetric around mRNA start sites. (B) Omni-ATAC and RNA-seq signal enrichment across gene bodies. Omni-ATAC signal is greatly enriched 5′ of gene bodies and slightly enriched across genes, whereas RNA-seq reads are enriched across gene bodies, with an increase in enrichment towards the 3′ end of gene bodies. (C) PCA of Omni-ATAC libraries. PCA loadings were assigned to the top 1000 variant peaks, where variance was measured as row variance across all libraries. PC1 (60% of variance) appears to correlate with developmental time. PC2 (18% variance) appears to be associated with middle developmental time points. (D) Heatmap of Omni-ATAC accessibility dynamics, generated from IDE2 model fits. Colors reflect a z-score calculated from IDE2, where red (high z) indicates increase in accessibility and blue (low z) indicates decrease in accessibility. (E) Bar chart indicating number of peaks classified as significant, transient and monotonic by IDE2. (F) Bar chart quantifying stage of maximum accessibility for significant IDE2 model fits. Peaks predominantly achieved maximum accessibility at the start and end of the time course.

gene completeness. We also observed stronger enrichment of Omni-ATAC-Signal at promoters of Mikado genes compared with MAKER genes (Fig. 3F,G), as well as improved performance using genome annotation software (see Materials and Methods). Given these results indicating an improved genome annotation, we used the Mikado transcriptome as our reference for downstream analyses.

Using the Mikado transcriptome as a reference, we assigned Omni-ATAC-seq peaks to a number of different spatial categories ('peak types') along the genome (Fig. 3H; see Materials and Methods). We observed approximately equivalent numbers of peaks over gene bodies (34.4%: 6.2% exonic, 28.2% intronic) and in intergenic regions (37.2%: 7% proximal intergenic, 30.2% distal intergenic). We further partitioned the intergenic peaks into proximal and distal segments, with distal intergenic peaks representing those peaks >10 kb away from the nearest gene (see Materials and Methods for rationale for this cutoff). The majority of intergenic peaks in our dataset were distal intergenic peaks (82%; 30.2% distal intergenic / 37.2% total intergenic peaks), indicating that many intergenic regulatory elements could not have been identified using previous approaches. The average distance of intergenic peaks was 73,351 bp away from the nearest gene (Fig. S4B). Nearly a fifth of regulatory elements (18.2%) were also located on contigs that did not contain genes, and were classified as 'unknown'.

As our peaks represent the merged combination of peaks across individual time points, we also evaluated the proportion of peak types from each time point (Fig. 3H). Peaks analyzed in this way could reveal differences in the spatial distribution of peaks across the genome over development. After normalizing for the number of 'unknown' peaks, the proportion of peaks belonging to intergenic and promoter regions declined slightly as development progressed, whereas the proportion of gene body peaks increased (Fig. S4E). There appeared to be slightly more intergenic peaks than gene body peaks at all developmental time points, and the ratio of intergenic peaks to gene body peaks declined very slightly over time (Fig. S4F). These data indicate that intergenic and gene body peaks have different enrichment trajectories over time. As development progresses, gene body peaks are increasingly enriched, and may have a greater impact on gene regulation.

Inferring nucleosome positioning and transcription factor footprints from Omni-ATAC-seq

ATAC-seq identifies regions of open chromatin. However, with sufficient depth of sequencing, one can infer additional information from the pattern of insertions of ATAC-seq adapters into the genome. Using our Omni-ATAC data, we predicted the temporal positioning of nucleosomes near Omni-ATAC peaks using NucleoATAC and inferred the footprints of transcription factors bound to accessible peaks using HINT-ATAC.

NucleoATAC enables the prediction of nucleosome positioning from ATAC-seq data (Schep et al., 2015). We applied NucleoATAC
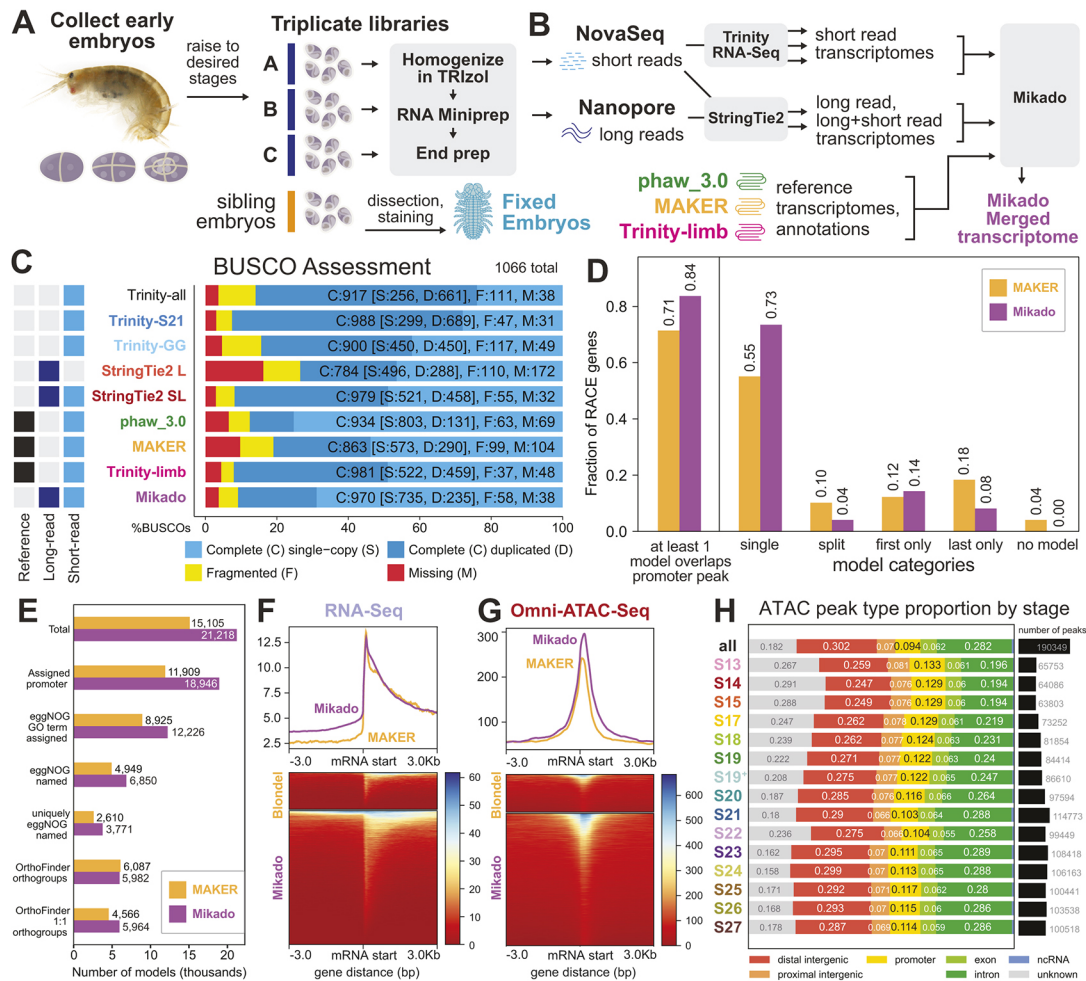
Fig. 3. Time-course RNA-seq of *Parhyale* embryos. (A) RNA-seq protocol overview. Triplicate libraries were generated for four developmental stages. (B) RNA-seq transcriptome assembly and merging pipeline. Short-read, long-read and long-+short-read transcriptomes were generated for pooled developmental stages matched to Omni-ATAC time points (S13, S19, S21, S23) using Trinity and StringTie2. Transcriptomes were also assembled using Trinity for short reads from a limb developmental time course (see Materials and Methods). Assembled transcriptomes were then merged with additional transcripts from previous publications into a Mikado transcriptome. (C) BUSCO scores for representative datasets and Mikado transcriptome. Overall, the StringTie2 short-+long-read transcriptome (StringTie2 SL) performed comparably with the best Trinity transcriptomes. The Mikado transcriptome appeared to have a comparable BUSCO complement to the other datasets with high BUSCO completeness. (D) Gene model completeness for Mikado transcriptome versus MAKER genome annotation based on a dataset of 49 RACE genes. The Mikado transcriptome appears to have more complete gene models and less fragmentation. (E) Comparison of number of gene models of different functional annotation categories between MAKER genome annotation and Mikado transcriptome. The Mikado transcriptome appears to produce more annotated genes with a higher annotation quality. (F) Plot comparing RNA-seq signal pileups between MAKER mRNA starts and Mikado mRNA starts. Axis represents mean signal at mRNA start sites for each dataset. (G) Plot comparing Omni-ATAC-seq signal pileups between MAKER mRNA starts and Mikado mRNA starts. (H) Categorization of Omni-ATAC-seq peaks in each stage-specific ATAC-seq library by their position relative to Mikado gene models. Distal intergenic peaks are defined as those >10 kb away from the nearest gene, whereas unknown peaks were peaks that were located on a contig that did not have a gene model.

to a window of ±500 bp around our Omni-ATAC-seq peaks at each developmental stage. NucleoATAC depends on a stereotypical signal of Tn5 insertion around nucleosomes referred to as a 'V-plot', which resembles the fragmentation pattern around chromatin subjected to chemical fragmentation. We observed across developmental time points a clear V-plot signal generated by NucleoATAC (example of V-plot from S21 libraries shown in Fig. S8A).

A well-established signal found across eukaryotes is the presence of strongly positioned +1 and −1 nucleosomes around gene promoters, and nucleosome depletion at RNA-Pol II binding sites (referred to as nucleosome-free region; NFR) (Brogaard et al., 2012; Radman-Livaja and Rando, 2010). We visualized NucleoATAC signal and nucleosome occupancy around mRNA starts genome-wide, and observed strong signals associated with +1 and −1 nucleosomes (Fig. 4A,B; Fig. S8B-D). We also observed a depletion of nucleosomes and a decrease in NucleoATAC signal between the +1 and −1 nucleosomes, which may represent NFRs. This signal was observed consistently in each of our stage-specific libraries (Fig. S8E,F). These data indicate that we were able to identify the stereotypical nucleosome signal at promoters, suggesting that NucleoATAC is able to infer nucleosome positions within our dataset genome-wide.

Deeply sequenced ATAC-seq data also enables transcription factor binding inference. We used HINT-ATAC to identify transcription factor footprints in Omni-ATAC-seq peaks from each developmental stage (Li et al., 2019) and evaluated the enrichment of these transcription factor footprints across developmental time. We used the JASPAR database of position frequency matrices (PFMs) to comprehensively identify possible
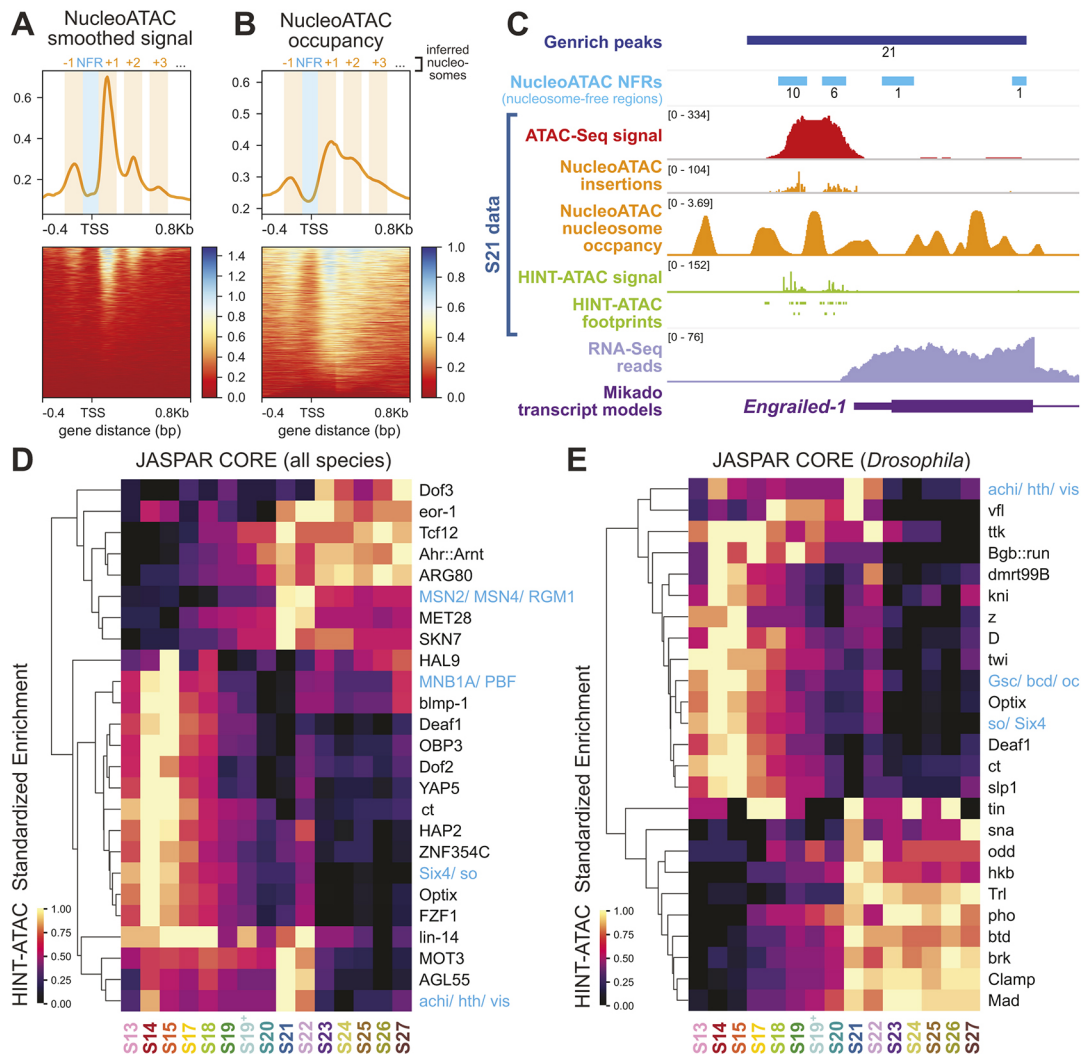
Fig. 4. Inference of nucleosome positioning and transcription factor binding from Omni-ATAC-seq data. (A) NucleoATAC smoothed signal at mRNA starts genome-wide. A clear pattern of peaks in the signal can be observed, suggestive of nucleosome positioning. Axis reflects mean signal across all mRNA starts at stage S21. (B) Inferred nucleosome occupancy at mRNA starts genome-wide. Axis reflects mean signal across all mRNA starts at stage S21. Inferred nucleosome annotations on A and B added based on observed signal. (C) Visualization of NucleoATAC, HINT-ATAC, Omni-ATAC and RNA-seq data at the inferred *Engrailed-1* promoter at stage S21. Deeper inference on Omni-ATAC-seq data allows insight into the broader chromatin landscape at developmental time points included in this dataset. (D) Summary of standardized enrichment of JASPAR CORE position frequency matrices (PFMs) among HINT-ATAC identified footprints across developmental stages. The top 25 most-enriched PFMs are displayed. Color is based on standardized enrichment, where minimum and maximum enrichment within each row are set to 0 and 1, respectively. PFMs marked in blue had identical enrichment ratios within each group at all time points, likely due to having highly similar PFMs. Overall, PFMs appeared to be enriched at early, late or middle developmental stages. (E) Summary of top 25 most-enriched *Drosophila* PFMs from the JASPAR CORE database. Overall, PFMs appeared enriched at early, late or middle developmental stages.

binding motifs (Fornes et al., 2020). We observed that HINT-ATAC-enriched transcription factor footprints appeared to form three groups based on the timing of highest enrichment: early, S21/S22 and late (Fig. 4D,E). These results are consistent with the general trends observed in PCA, which suggest that the early, middle and late time points in development have the greatest variation.

The data generated from these analyses enable the visualization of chromatin accessibility, predicted nucleosome position, inferred transcription factor binding and RNA-seq expression at individual loci across around half of *Parhyale* development. The strengths of these data are illustrated in Fig. 4C at the predicted promoter of the *Parhyale Engrailed-1* locus for stage S21. Thus, for any genomic region of interest, one can develop a comprehensive prediction of the local chromatin environment at each of the 15 time points in our dataset.

Identifying gene regulatory programs from Omni-ATAC-seq using fuzzy clustering

To assess the potential for our dataset to provide new biological insights, we attempted to identify distinct gene regulatory programs based on differential accessibility. We used the Mfuzz package to perform fuzzy c-means clustering on the matrix of read counts generated from our Omni-ATAC-seq peaks (Kumar and E Futschik, 2007). We determined that our peaks could be optimally partitioned into nine clusters (see Materials and Methods).

We evaluated the differences between clusters through a variety of approaches (Figs 5 and 6). Fig. 5D shows the standardized raw accessibility scores of peaks in each of the nine clusters over time. Our clusters appeared to show substantial differences in the timing of their maximum accessibility, the GO terms associated with genes located near peaks in each cluster and also the enrichment of binding footprints for different transcription factor families. For example,
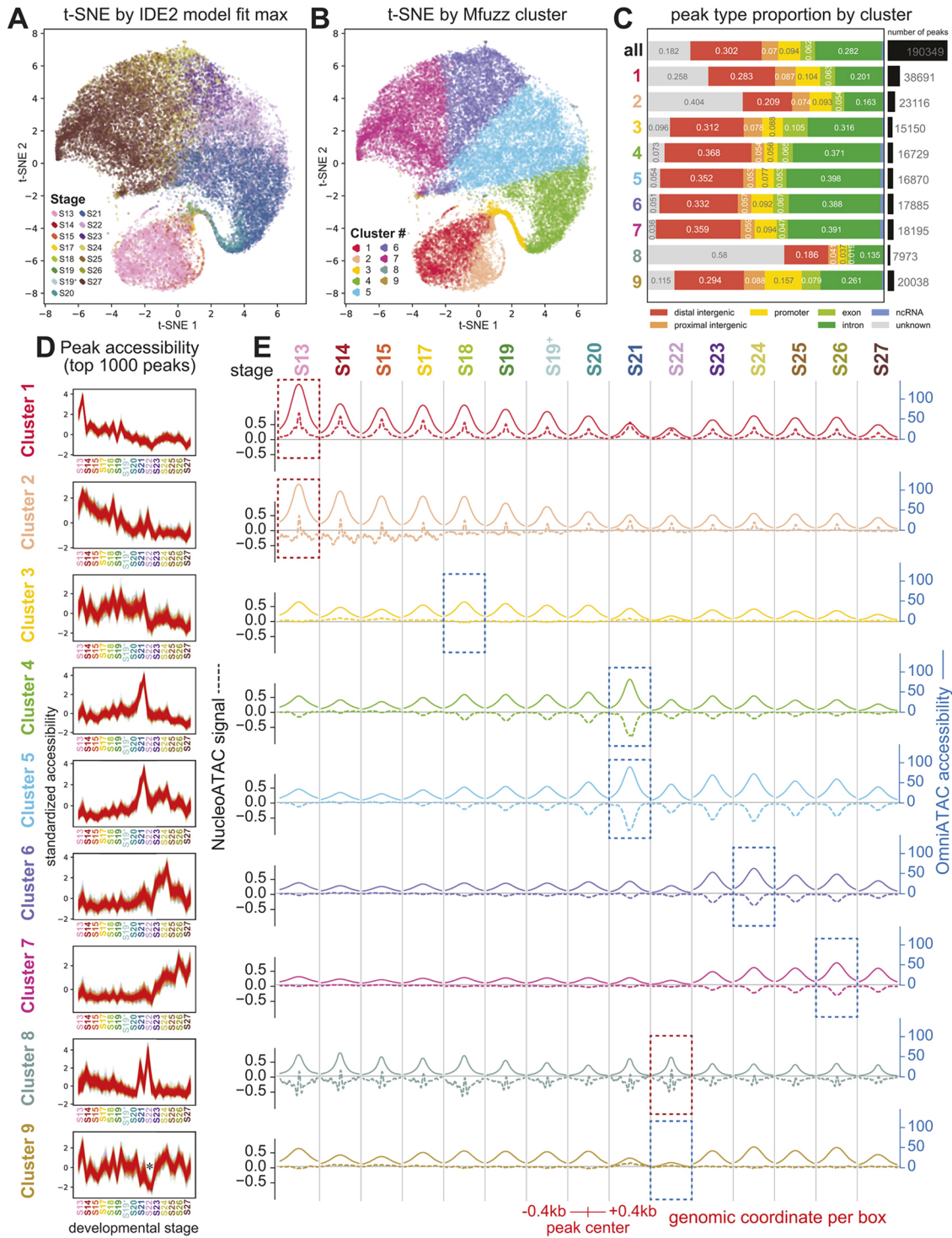
Fig. 5. Identification and classification of regulatory element clusters. (A) t-SNE plot of the top 10,000 most statistically significant peaks identified by IDE2, colored by IDE2 max fit. (B) t-SNE plot of the same points from A colored by Mfuzz cluster. (C) Distribution of peak position categories by Mfuzz cluster. Clusters 1, 2 and 8 appeared to be enriched for 'unknown' peaks, whereas other clusters appeared to be enriched for distal intergenic peaks. (D) Standardized accessibility plot for top 1000 peaks with strongest membership in each of the nine clusters. Each library is plotted as a separate point along the line plot. Line color indicates the cluster membership value for each peak in each plot, with dark red indicating strong cluster membership and blue indicating weaker cluster membership. An asterisk in the Cluster 9 panel marks the decrease in accessibility observed commonly in Cluster 9 peaks, more clearly shown in the IDE2 model fits in Fig. S9B. (E) Line plots showing Omni-ATAC accessibility (solid line, blue axis) and NucleoATAC signal (dashed line, black axis) for peaks in each cluster across time. Each box contains a line plot summarizing each signal at all peaks within that cluster with respect to a given developmental Omni-ATAC-seq library. The center of each line plot reflects the average signal at the center of all Omni-ATAC-seq peaks in that cluster at that developmental stage. Signal is visualized at 0.4 kb upstream and downstream of peak centers. Within each signal type, the axes across line plots for each cluster are identical. Dotted lines around subsections of the plot highlight regions of interest. Dotted blue lines indicate the time point during which peaks in a given cluster achieved both maximum accessibility and decreased nucleosome occupancy. Dotted red lines indicate the time point during which peaks in a given cluster achieved both maximum accessibility and increased nucleosome occupancy.
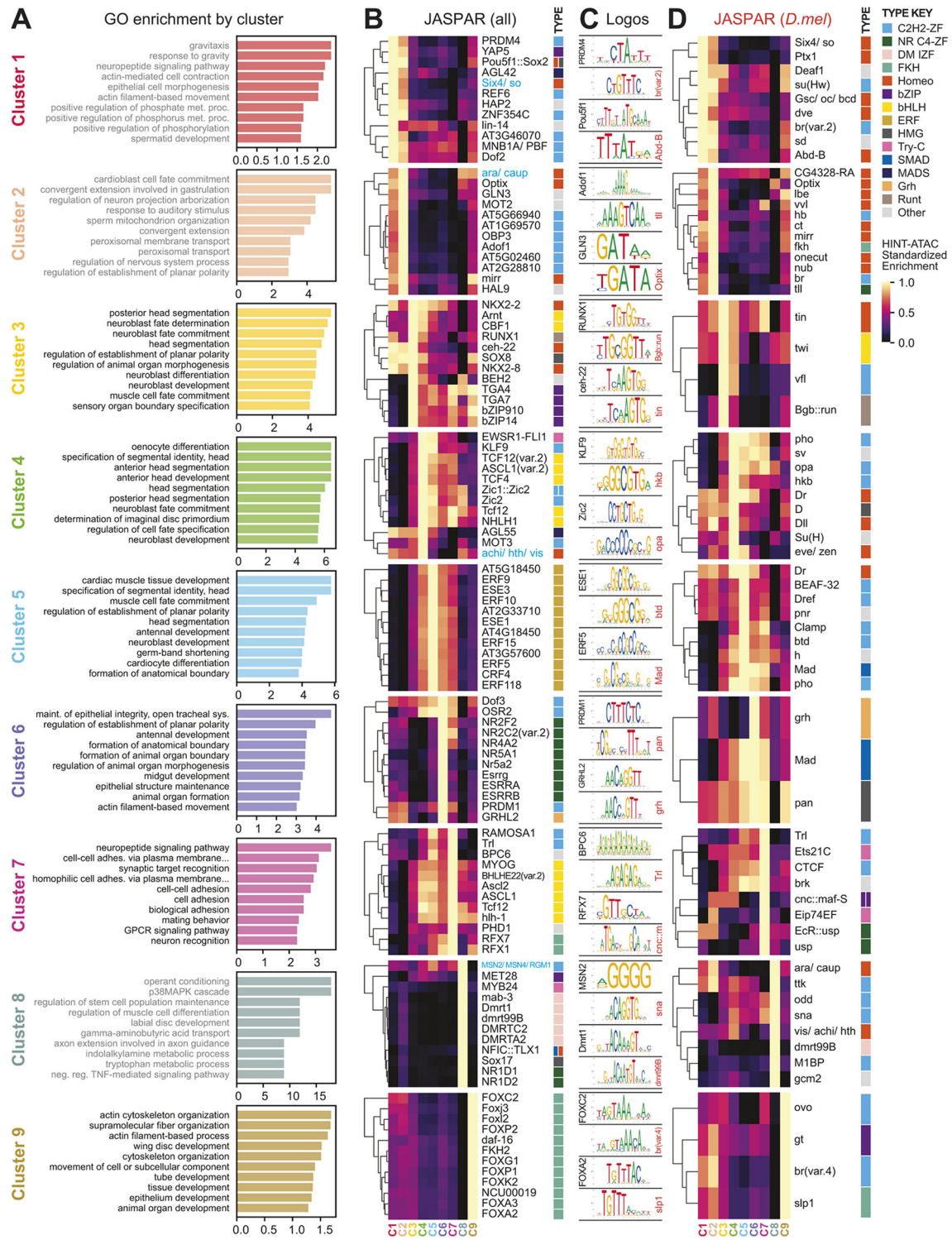
Fig. 6. Functional annotation of clusters using GO enrichment and transcription factor binding prediction. (A) GO enrichment per cluster based on the nearest gene for each peak. For each peak in each cluster, we identified the nearest gene and extracted the *Drosophila* OrthoFinder gene name for that gene. We performed GO enrichment using *Drosophila* GO terms, using the list of all OrthoFinder gene names found in our genome as the background for the analysis. Bar charts display fold enrichment for each GO term. Gray text represents GO terms that did not have a significant *P*-value (*P*<0.05) after false discovery rate (FDR) correction. (B) HINT-ATAC enrichment per cluster for JASPAR CORE PFMs. For each cluster, we identified all unique transcription factor footprints across all 15 time points found within peaks of that cluster and performed enrichment relative to a randomly-selected background. We show the top 12 PFMs with the greatest enrichment for each cluster that also had the highest enrichment in that cluster relative to other clusters. Colored boxes represent the transcription factor family of each transcription factor in the clustermap. (C) JASPAR motif logos for select pairs of transcription factors between non-*Drosophila* and *Drosophila* transcription factor with similar motifs as determined by STAMP alignment. For each cluster, two pairs of transcription factors are shown: a non-*Drosophila* transcription factor (labeled in black) and a *Drosophila* transcription factor (labeled in red). In some cases, a strong similarity is observed between the non-*Drosophila* and *Drosophila* PWMs. (D) HINT-ATAC enrichment per cluster for *Drosophila* JASPAR CORE PFMs. Some clusters had fewer than 12 PFMs for which that cluster had the highest enrichment.

Cluster 9 appeared to be associated with peaks that decreased in accessibility in the middle of development (Fig. 5D, asterisk), and genes located near peaks in this cluster appeared to show GO enrichment for cytoskeletal terms. For a fuller description of the differences between clusters, see supplementary Materials and Methods.

We also evaluated transcription factor footprint enrichment for each cluster. Given that many JASPAR PFMs come from non-arthropod sources, we used STAMP to align non-*Drosophila* to *Drosophila* motif sequences based on similarity (Mahony and Benos, 2007) (see Fig. S11). Selected pairs of motifs are shown in Fig. 6C for each cluster (non-*Drosophila* labeled in black, *Drosophila* labeled in red). For some pairs, we were able to identify candidate *Drosophila* transcription factors that appeared highly similar to non-*Drosophila* sequences. For example, in Cluster 9, we observed strong enrichment for FKH family transcription factors, the PFMs of which matched the br(var.4) and slp1 PFMs. This suggests that peaks found in Cluster 9 may be regulated by br(var.4) and slp1 binding in *Parhyale*.

Altogether, our results indicate that clustering Omni-ATAC data using accessibility can identify groups of peaks with similar accessibility trajectories, which can be further analyzed using nucleosome occupancy, GO enrichment and transcription factor footprint enrichment to understand possible biological functions and genetic mechanisms behind differential accessibility.

Concordant and discordant expression and accessibility dynamics appear across development

To understand the relationship between peak accessibility and gene expression, we examined the relationship between the change in accessibility of individual peaks and the change in expression of nearby genes (Fig. 7). We used DESeq2 to evaluate the log_2-fold change of both expression (RNA-seq) and accessibility (ATAC-seq) for all genes and peaks. Among our differential accessibility and expression analyses, we observed both 'concordant' and 'discordant' relationships. Peak-gene pairs with concordant accessibility and gene expression were those in which the sign of the log_2-fold change in expression and accessibility were in agreement – for example, peaks for which an increase in accessibility was observed concurrently with an increase in expression. Meanwhile, peak-gene pairs with discordant accessibility and gene expression were those for which an increase in accessibility was observed concurrently with a decrease in expression, or vice versa.

We observed both classes of peak-gene pairs among all peaks, but also among promoter peaks. Over time, the number of concordant peak-gene pairs increased from 1888 at the S19 versus S13 comparison, to 5738 at the S21 versus S19 comparison, and then decreased to 5240 at the S23 versus S21 comparison (Fig. S12D). Meanwhile, the number of discordant peaks gradually increased over time (761 to 2231 to 3115 peak-gene pairs) (Fig. S12D). These trends were also observed for the promoter-only peaks. The gradual increase in discordant peaks may indicate an increase in repressive gene regulation as gene expression becomes refined over the course of differentiation.

Overall, the classification of peaks as concordant or discordant in accessibility and gene expression may provide downstream users of this data with hypotheses about CRE function. Given that CREs have recently been shown to function as both activating elements and silencers, depending on tissue context (Gisselbrecht et al., 2020; Halfon, 2020), we cannot directly map concordant peaks to enhancers or discordant peaks to silencers. Without information about higher-order chromosome contacts, it is also difficult to precisely assign a given regulatory element's function to a particular gene of interest.

However, information about concordance and discordance of peaks could be useful for researchers deciding which among many peaks surrounding a gene of interest could be most fruitful for reporter construction. To facilitate analysis of accessibility and expression for peak-gene pairs of interest, we have included functions enabling visualization of these two factors for arbitrary peak-gene pairs, illustrated in Fig. S13. The visualization tool displays accessibility and gene expression for user-selected Omni-ATAC peaks and Mikado genes over RNA-seq stages (S13, S19, S21, S23), fold change in accessibility and expression between adjacent time points, significance of fold change as evaluated by DESeq2 and concordance/discordance assignment at each time point.

Omni-ATAC-seq re-identifies known *Parhyale* regulatory elements

To assess the practical usefulness of our Omni-ATAC-seq dataset, we compared our Omni-ATAC-seq peaks to known regulatory elements in *Parhyale*. A very limited set of regulatory elements have been described in *Parhyale*: a muscle reporter, PhMS (Pavlopoulos and Averof, 2005); a heat shock element, HS2a (Pavlopoulos et al., 2009); an embryonic ubiquitous reporter, PEB (Liubicich, 2007); and two reporter constructs for *Parhyale Opsin-1* and *Opsin-2* (Ramos et al., 2019). We were unable to locate any Omni-ATAC-seq peaks for the *Parhyale Opsin* genes, likely owing to their late expression (outside of our developmental time course) and very low cell number (a handful of cells per eye). However, strong Omni-ATAC-seq enrichment was observed at the PhMS, HS2a and PEB elements, as illustrated in Fig. S14.

Each of these reporter constructs consists of the first exon and intron of a gene. In each case, we observed an Omni-ATAC peak overlapping the 5′ end of the gene. For the PhMS reporter (Fig. S14B), previous work had identified a cluster of putative basic helix-loop-helix (bHLH) transcription factor binding sites. These sites were captured within our Omni-ATAC peak. The HS2a reporter has been reported to contain two binding sites for heat shock factor (HSF) proteins upstream of a minimal promoter (PhHsp70), both of which are captured within our Omni-ATAC peak (Fig. S14D). Finally, the PEB element appears to contain two Omni-ATAC peaks. These results suggest Omni-ATAC is able to identify the position of functional CREs, and has been able to capture those regions important for CRE function.

Minos transposase reporter assays can reveal novel promoters and distal enhancers

To assess the function of candidate regulatory elements identified by Omni-ATAC-seq, we employed a *Minos* transposase reporter assay (Fig. 8A). In this assay, we injected *Minos* transposase mRNA along with a transposon donor plasmid containing a reporter gene construct that expressed DsRed to one- and two-cell embryos. Once per day, from 3 to 10 days postfertilization (dpf), we screened for DsRed expression on a Zeiss LSM780 confocal microscope. We tested a variety of different reporters, summarized in Table 1 (see Materials and Methods for a discussion of candidate reporter selection approaches and see supplementary Materials and Methods and Fig. S16 and Fig. S17 for description of spontaneous expression patterns observed).

Among the novel reporter constructs we tested, two showed robust expression: pMi(Hsc70-4) and pMi(ne1)-Hsp70-p2. The pMi(Hsc70-4) construct was constructed using the putative
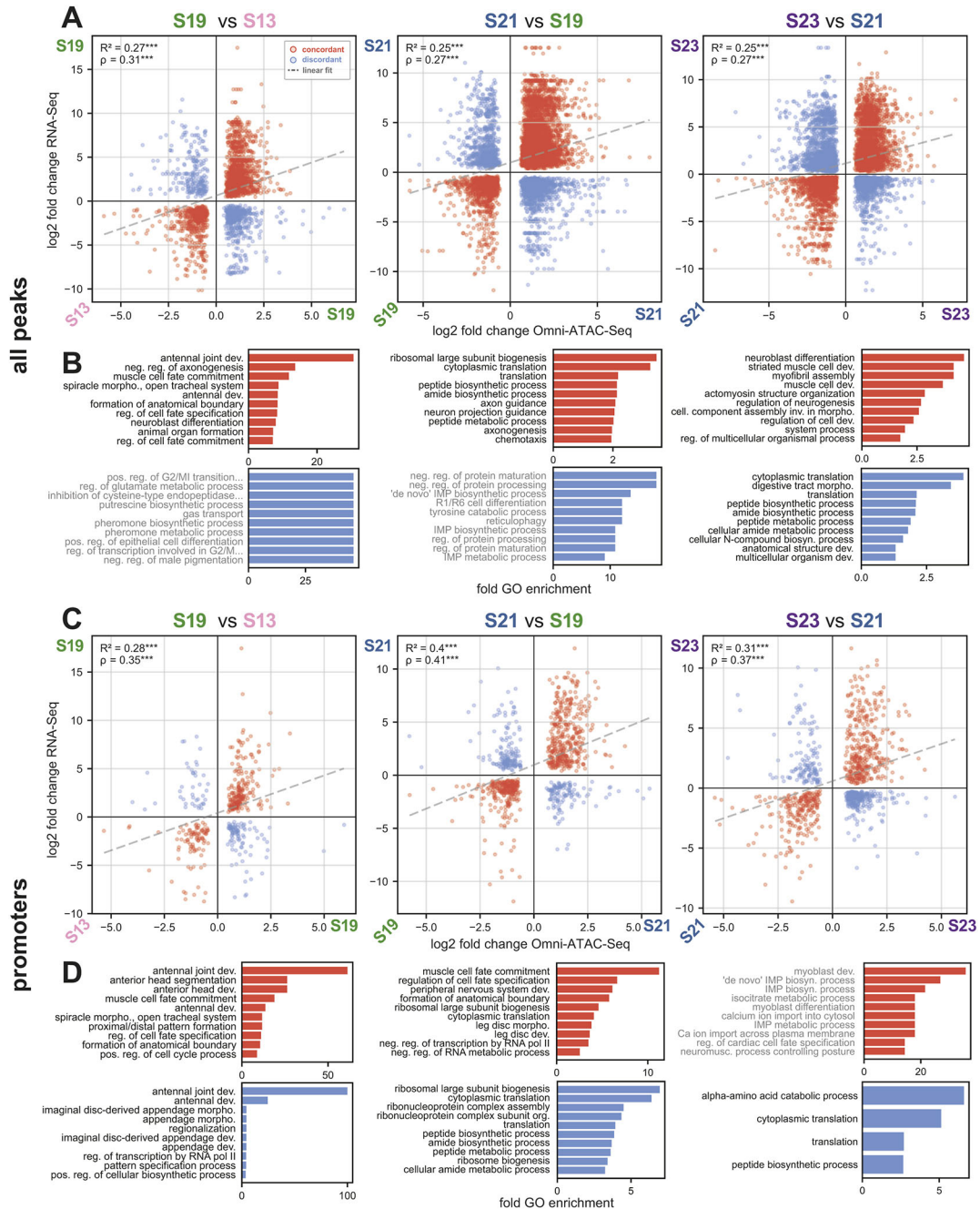
Fig. 7. Correlation between RNA-seq and Omni-ATAC-seq. (A) Correlation between log$_2$-fold change RNA-seq and log$_2$-fold change Omni-ATAC-seq at peak-gene pairs in sequential pairwise comparisons of developmental stages. Only peak-gene pairs with a significant difference in expression and accessibility between the two time points are plotted. Positive axis reflects higher expression or accessibility at the later developmental stage for each plot. Points colored in red reflect concordant relationships between log$_2$-fold change in RNA-seq and Omni-ATAC-seq; blue peaks reflect discordant relationships. Dotted line represents a linear fit to all data in the plot. Pearson correlation R^2 and Spearman correlation ρ for all points are displayed in each plot. ***P<0.001. (B) GO-term enrichment for gene-peak pairs with concordant or discordant expression and accessibility log$_2$-fold change for all peaks. Gene lists were extracted based on the nearest Mikado gene to each peak. (C) Correlation between log$_2$-fold change RNA-seq and log$_2$-fold change Omni-ATAC-seq at all peaks in each developmental stage. Positive axis reflects higher expression at the later developmental stage for each plot. Pearson correlation R^2 and Spearman correlation ρ for all points are displayed in each plot. ***P<0.001. (D) GO-term enrichment for gene-peak pairs with concordant or discordant expression and accessibility log$_2$-fold change, for promoter peaks.

promoter peak of the Hsc70-4 gene (Fig. 8B,D). This reporter showed strong expression in a cluster of neurons associated with the eye beginning at ~9 dpf (Fig. 8C). The expression pattern of these neurons appears to be distinct from those observed using a synthetic 3XP3 enhancer, which has previously been used as a positive marker of transformation *Parhyale* (Pavlopoulos and Averof, 2005).

Thus, it is possible to build new reporters from our Omni-ATAC peaks using putative promoter peaks.

The second reporter that showed robust expression was the pMi(ne1)-Hsp70-p2 construct, which contains the PhHsp70 minimal promoter along with a strong peak extracted from the Hsp70 cluster, about 30 kb away from the location of the minimal
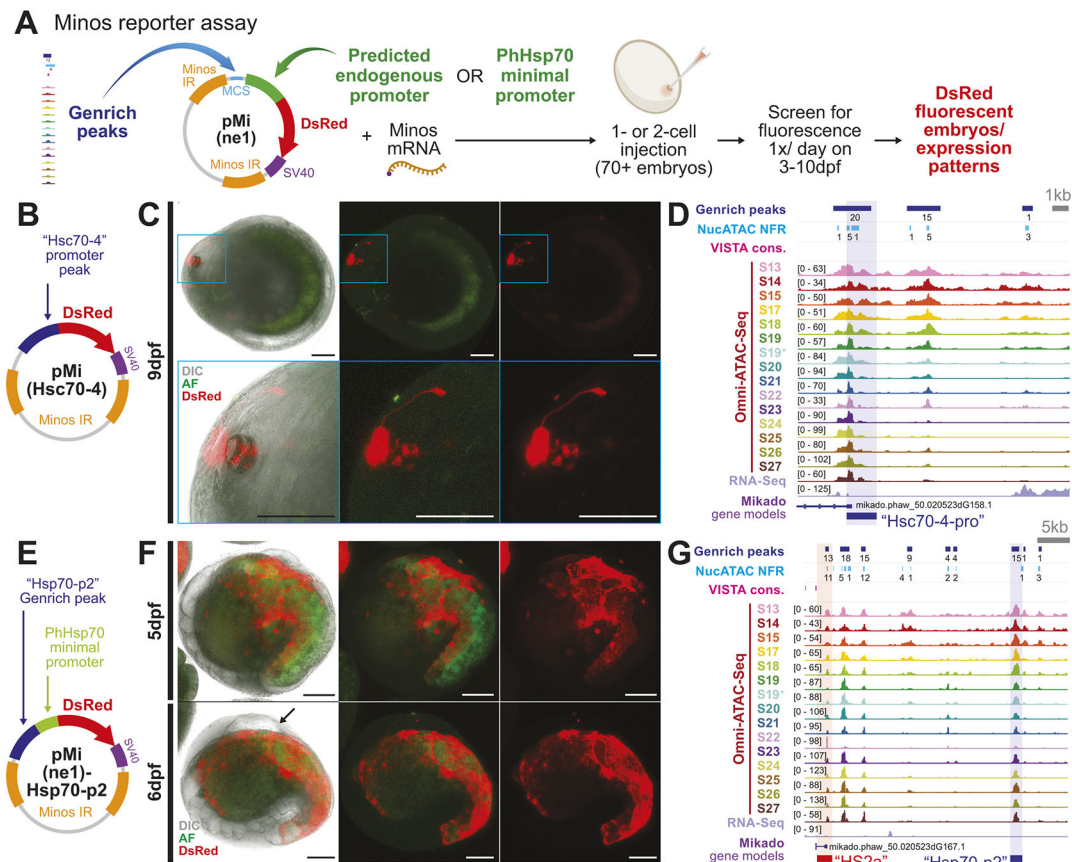
Fig. 8. Minos transposase reporter assay identification of novel regulatory elements. (A) Minos transposase reporter assay and screening approach. (B) Plasmid schematic for the pMi(Hsc70-4) reporter. (C) Expression of the pMi(Hsc70-4) reporter in neurons associated with the eye of a 9 days postfertilization embryo. Bottom row shows a magnified image of the head region. A clear projection of the neurons can be seen. (D) The genomic region of the Hsc70-4 promoter peak. (E) Plasmid schematic for the pMi(ne1)-Hsp70-p2 reporter. (F) Expression of pMi(ne1)-Hsp70-p2 reporter at 5 and 6 dpf showing strong expression in the yolk. Arrow marks aberrant morphology at the dorsal side of the embryo observed consistently in all embryos with fluorescent expression. (G) The genomic region of the Hsp70-p2 peak relative to the HS2a element. The PhHsp70 minimal promoter used in this construct is located within the HS2a element. Scale bars: 100 µm.

promoter (Fig. 8E,G). Hsp70-p2 drives strong expression in cells of the embryonic yolk (Fig. 8F). Some yolk cells labeled by this reporter moved dynamically around the yolk, whereas others appeared static and showed spongiform morphology (Movie 1). We observed defects in the development of the dorsal portion of the embryo in mid-stage (∼6-7 dpf) embryos injected with this construct, which stereotypically displayed a separation between the embryo proper and the eggshell (Fig. 8F, bottom panel, marked with arrow). Although we were unable to establish a genetic line using this construct, we observed strong and reproducible expression in many embryos. This reporter is the first indication that distal enhancers exist and can be identified in *Parhyale*.

Although it remains challenging to identify novel CREs in *Parhyale*, our results suggest that functional regulatory elements can be identified from our Omni-ATAC data. Among the numerous peaks identified in our dataset, we expect that many will reveal novel gene regulatory dynamics. Future work to identify such elements will require the optimization of reporter expression and transgenic strategies to increase the throughput and sensitivity of reporter gene assays.

DISCUSSION
New genomic data enable deeper understanding of underlying biology. This is especially true in emerging research organisms, for which data and resources are limited. This work provides a

wealth of genomic resources for the amphipod crustacean *P. hawaiensis*. In addition to providing Omni-ATAC-seq and RNA-seq data for a broad developmental time course, our work generates multiple new transcriptomes, an updated genome annotation, a catalog of dynamically accessible chromatin regions and predictions of nucleosome occupancy and transcription factor binding. Moreover, our thorough analysis of regulatory element dynamics through clustering, GO enrichment, predicted transcription factor binding and correlations with RNA-seq present numerous hypotheses for CRE function. Such data will support researchers in the growing *Parhyale* community in efforts to identify and characterize developmental regulatory elements, and provide the foundations for more advanced approaches, including single-cell sequencing techniques, which rely on high-quality reference datasets.

The *Parhyale* genome contains many distant and dynamic regulatory elements
From our data, we were able to glean new information about the global dynamics of gene regulation during *Parhyale* development, as well as the composition of the *Parhyale* genome. We observed that the vast majority of our peaks showed dynamic accessibility over developmental time, and that peaks could have a variety of different temporal dynamics, including transient increases and decreases in accessibility alongside more absolute increases and decreases. These results indicate that much of the *Parhyale* genome

DEVELOPMENT

11

Table 1. Plasmids tested using Minos transgenesis

Plasmid	Type	Gene – *Drosophila* expression	Number injected	Number survived	Number of DsRed or GFP+	Phenotype
pMi(gm)	Promoter	N/A – *Parhyale* muscle reporter PhMS	109	39	24	Green muscle expression
pMi(retn)	Promoter	*retained* – expressed in some neurons	240	104	5	No consistent phenotype
pMi(SVP)	Promoter	*seven up* – expressed in ventral neuroblasts	122	31	0	None
pMi(su-s-2)	Pro+peak	*suppressor of sable* – embryonic ubiquitous expression	127	57	7	No consistent phenotype
pMi(Rx)	Pro+peak	*retinal homeobox* – expressed in various neural tissues	74	28	0	None
pMi(Hsc70-4)	Promoter	*Heat shock 70 complex protein 4* – embryonic ubiquitous	212	99	22	Neurons associated with the eye
pMi(En1)	Promoter	*engrailed* – parasegment stripes	71	33	0	None
pMi(En1)-p1	Pro+peak	''	67	46	0	None
pMi(En1)-p1.5	Pro+peak	''	70	51	0	None
pMi(En1)-p5	Pro+peak	''	68	58	1	Faint expression domain
pMi(Sp69)	Promoter	*Sp69* – leg gap gene	74	56	0	None
pMi(Sp69)-p1	Pro+peak	''	50	30	0	None
pMi(Sp69)-p1.5	Pro+peak	''	86	32	3	No consistent phenotype
pMi(Sp69)-p2.5	Pro+peak	''	71	35	0	None
pMi(Sp69)-p3	Pro+peak	''	241	121	8	Head, leg, gill
pMi(ne1)-Hsp70-p1	Pro+peak	Hsp70 complex	45	26	0	None
pMi(ne1)-Hsp70-p2	Pro+peak	Hsp70 complex	310	81	61	Yolk cells

N/A, not applicable; '', same as previous row.

undergoes dynamic changes in accessibility over developmental time.

With an improved genome annotation, we were also able to determine that many (∼37.2%) regulatory elements in *Parhyale* are located in intergenic regions, and that among all elements at all time points, elements located >10 kb away from the nearest gene (distal intergenic peaks) were the largest group, followed by intronic peaks. These results are notable, given that many of the currently studied protostome genomes are small (for example, *Drosophila melanogaster* is 180 Mb, *Tribolium castaneum* is 200 Mb and *Bombyx mori* is 530 Mb). In such organisms, regulatory elements tend to occur within short distances from promoters, and it is common in such systems to attempt to build reporter genes using a short window upstream of the promoter of a gene of interest.

Parhyale is an example of an arthropod with a large genome (3.6 Gb, or ∼10% larger than *Homo sapiens* at 3.2 Gb). Given the large genome size, and the large proportion of elements located distant to genes, we expect that regulatory elements within *Parhyale* may be located more distantly from gene promoters than in other arthropods. Future attempts at building reporter genes in this organism will need to account for the large number of distant regulatory elements, which may be crucial to proper reporter expression.

Our data are among the first to identify regulatory elements genome-wide in a non-insect arthropod (Gatzmann et al., 2018; Kissane et al., 2021), and also examine one of the largest sequenced arthropod genomes available to date (Kao et al., 2016). These data will enable other researchers to examine the relationship between genome size and regulatory element composition across more diverse taxa, and will be useful in developing a deeper understanding of how genome organization affects gene expression across the metazoans.

Combined short- and long-read transcriptomes can improve genome annotations

Our work can also serve as a guide for other researchers working in emerging research organisms. First, our work demonstrates the utility of combining short- and long-read sequencing in generating more accurate genome annotations for emerging research organisms. Even with a small number of sequences (∼1.2 million reads) generated using inexpensive Nanopore sequencing, we were able to assemble a transcriptome with a moderate BUSCO score (71%). Combining short reads and long reads (StringTie2 SL), we were able to construct a transcriptome with both a high BUSCO score (91.8%) and low rate of gene fragmentation.

Moreover, the completeness of the StringTie2 SL transcriptome met or exceeded that of several of our Trinity-derived transcriptomes. Examination of individual genomic regions in the *Parhyale* genome indicates that Trinity generates numerous spurious transcripts, which can be confounding for whole-genome analysis (see Fig. S6 for examples). Thus, for other researchers with access to a genome assembly, we would strongly recommend performing both short- and long-read sequencing, and assembling using multiple different assembly strategies to generate higher quality genome annotations. For those without access to a high-quality genome, we would caution against relying strictly on transcripts generated from software such as Trinity, which, in our system, generated many transcripts that did not appear to match with previous RACE data. Filtering transcripts by expression values, as well as employing transcriptome-merging pipelines such as Mikado (in the case of a high-quality genome) and EvidentialGene (in the absence of a genome), would likely help remove such spurious transcripts.

Deep analysis of ATAC-seq data can reveal more than just dynamic accessibility

Although the most direct analysis of ATAC-seq data can enable the identification of dynamic regulatory elements, our work uses additional tools to infer nucleosome positioning and transcription factor binding from our ATAC-seq data. Using NucleoATAC, we were able to recover clear signals of nucleosome positioning at promoters, a quality observed in numerous organisms. Using HINT-ATAC, we were able to predict transcription factor binding across developmental stages. Such analyses enable a glimpse of the possible chromatin landscape of the genome, particularly in the absence of more direct and specific, but also more time-consuming and expensive, approaches such as ChIP-seq or CUT&RUN.

DEVELOPMENT

In addition to performing inference, we were able to identify clusters of peaks with potentially distinct biological functions using fuzzy clustering. These clusters of peaks appeared to differ in their accessibility dynamics, nucleosome positioning and enrichment of transcription factor footprints, indicating that meaningful biological differences can be gleaned from deeper analysis of ATAC-seq data. For example, we observed strong enrichment in Cluster 9 for FKH transcription factor footprints matching the PFMs for *Drosophila* br(var.4) and slp1. The strongest peaks in this cluster appeared to share a common decrease in accessibility during stages S21-S22, and an enrichment for genes with GO terms related to cytoskeletal function. This group of peaks appeared to show a decrease in accessibility at a time point during which the embryo undergoes a dramatic morphogenetic event in which it splits along the midline (Browne et al., 2005). We hypothesize that FKH domain-containing transcription factors may play a role in that important morphogenetic event. Using our analyses as a starting point, other *Parhyale* researchers may be able to make hypotheses about their own developmental processes of interest.

Omni-ATAC-seq identifies old and new regulatory elements, including enhancers

Our data were able to capture most of the previously-identified *Parhyale* regulatory elements. In addition, we were able to demonstrate that our data contain new regulatory elements. Notably, we demonstrate the first identification of a distal regulatory element, Hsp70-p2, located about 30 kb from the PhHsp70 minimal promoter. These results are the first identification of an enhancer separated by a large distance from a minimal promoter element in this organism.

Our results demonstrate that Omni-ATAC-seq is able to identify novel regulatory elements, and we expect that numerous new reporter genes will be built from these data. Previous work relied on examining candidate regulatory elements within individual genomic regions, or attempting to build new reporters through random integration events. With this dataset, it is now possible to identify candidate regulatory elements for any gene of interest and to screen for expression patterns using the Minos reporter system. Moreover, by evaluating concordance and discordance of gene expression and accessibility using this dataset, it is possible to further refine the list of candidates for reporter construction based on the potential of individual peaks to drive or repress gene expression.

We were unable to identify clear reporters from some of the developmental genes we tested, such as *Sp69*; however, many peaks near these genes remain to be tested. Future work to build new reporters in *Parhyale* may require the development of more efficient transgenesis and screening strategies. For example, improving the efficiency of CRISPR-mediated homologous recombination or CRISPR-mediated non-homologous end-joining transgenesis could enable researchers to insert a reporter near a candidate regulatory region of interest. Moreover, future approaches will also need to account for the numerous distant regulatory elements in the *Parhyale* genome, which may be important for gene regulation. Examining local DNA interactions using approaches such as Hi-C or other chromatin conformation capture approaches will be instrumental in identifying distant regulatory regions.

Together, these approaches have enabled the exploration of chromatin dynamics and transcription factor binding in an emerging model organism. Our work illustrates how a single, easily adapted protocol can yield data amenable to deep analysis. We would recommend that other researchers using ATAC-seq or similar assays in emerging model organisms also take advantage of the additional information that could be gleaned from deep analysis of their data, which could provide further insights into how both local and global changes to nucleosome occupancy and transcription factor binding influence their biological processes of interest.

Conclusion

By combining Omni-ATAC-seq with RNA-seq across a broad developmental time course, our work is able to identify and classify numerous candidate CREs in the genome of the amphipod crustacean *P. hawaiensis*. We demonstrate how deep analysis of Omni-ATAC data can facilitate the identification of peaks with distinct accessibility, nucleosome occupancy and transcription factor footprint enrichment. We further classify peaks as concordant or discordant regulatory elements by integrating differential accessibility and differential expression, revealing potential relationships between regulatory elements and nearby genes. Moreover, we show the potential to identify novel reporter genes using candidate promoters and enhancers from our data.

This work provides a substantial resource to the *Parhyale* community, and should accelerate the study of gene regulation in this emerging research organism. In addition, our work can serve as a framework for other researchers deploying ATAC-seq and RNA-seq approaches in emerging research organisms. Through deep analysis of ATAC-seq data, combined short- and long-read sequencing, and integration of accessibility and gene expression, such researchers can identify regulatory elements with distinct biological functions and advance the study of gene regulation in their organisms of interest.

MATERIALS AND METHODS

Crustacean cultures

Parhyale hawaiensis of the Chicago-F isolate were raised at 25°C and fed a diet of carrots, shrimp pellets and Spirulina flakes in Tropic Marin artificial seawater with a salinity of 31-35 ppm in plastic tanks.

Embryo cultures

Embryos were collected using previously described protocols (Rehm et al., 2009), staged using the Browne et al. (2005) staging guide, and raised at 27°C in a humidity-controlled incubator in filter-sterilized artificial seawater. For ATAC-seq and RNA-seq experiments, clutches of 15 or more embryos were collected and staged between S1 and S6, as these time points are among the shortest and most morphologically identifiable during early development.

Embryo staging

Two groups of five embryos each from a single clutch were used in the ATAC-seq experiments, and three groups of five embryos each from a single clutch were used in the RNA-seq experiments. Embryos of the correct stage were selected based on morphological characteristics as described in the Browne et al. (2005) staging guide, and any embryos with abnormal or asynchronous morphology were discarded. Morphologically representative embryos from the same clutch as used for the ATAC-seq were photographed on ventral and lateral positions within 10 min of beginning the tagmentation procedure. Any remaining embryos were boiled briefly and fixed using the protocol described in Rehm et al. (2009) and DAPI stained to further confirm staging. The process of staging embryos, imaging embryos, boiling and mashing embryos for Omni-ATAC tagmentation were all performed within a 30 min time interval for each developmental stage.

RNA-seq

Embryos for RNA isolation were homogenized in TriZol using a DWK Life Sciences (Kimble) Biomasher II Closed System Disposable Tissue Homogenizer, and RNA was isolated using the Zymo Direct-Zol Miniprep Plus kit. RNA quality was assessed based on fragment analysis using an Aligent 2100 Bioanalyzer. cDNA was generated from RNA using

the TaKaRa SMART-seq v4 Ultra Low-Input kit. cDNA was then sequenced using the Illumina NovaSeq to generate short reads. Using the Nanopore Direct cDNA barcoding kit (SQK-LSK109), cDNA was also sequenced on a Nanopore MinION flow cell to generate long reads.

Limb development RNA-seq methods
A pool of embryos consisting of embryonic stages 19, 20, 22, 23, 25 and 28 were homogenized with DWK Life Sciences Kimble Kontes Pellet Pestle Cordless Motor in Trizol and extracted using Trizol. PolyA+ libraries were prepared with the Truseq V1 kit (Illumina), starting with 0.6-3.5 mg of total mRNA, and sequenced on the Illumina HiSeq 2000 as paired-end 100 base reads, at the QB3 Vincent J. Coates Genomics Sequencing Laboratory.

Limb development RNA-seq de novo transcriptome assembly
For the Trinity-limb (old) assembly, transcripts were assembled de novo using Trinity r2013_08_14 (Grabherr et al., 2011) with parameters –JM 170G –CPU 10 –inchworm_cpu 2 –min_kmer_cov 2 –min_contig_length 49 –group_pairs_distance 700. Input RNA-seq reads were treated as paired-ends. Output transcript assemblies shorter than 200 bp were discarded. The remaining assemblies were screened for contaminants with BLASTX (BLAST+ v2.2.26; parameters: -num_descriptions 50 -num_alignments 50 -evalue 1e-5 -lcase_masking -soft_masking true -seg yes) (Camacho et al., 2009) against a database of all bacterial proteins downloaded from NCBI (retrieved 2013-08-31), then against Swiss-Prot UniProt human sequences (retrieved 2013-08-31) (The UniProt Consortium, 2019). All hits were further required to cover a minimum of 40% of each assembled transcript sequence. When filtering for human contaminants, a 98% identity threshold was also required. For the Trinity-limb (new) assembly, transcripts were assembled using Trinity v2.5.1 using standard settings.

Transcriptome assemblies
Using the Trinity (Haas et al., 2013) pipeline to assemble the short-read sequences, we generated four developmental stage-specific de novo transcriptomes (Trinity-S13 to Trinity-S23) and two merged transcriptomes, one de novo (Trinity-all) and one genome-guided (Trinity-GG). In addition, we generated two transcriptomes with reads from several developmental stages covering the time course of limb development [Trinity-limb (new) and Trinity-limb (old)]. Although the merged transcriptomes were generated from all four developmental stages (Trinity-all, Trinity-GG), they scored lower compared with stage-specific transcriptomes when coding DNA sequence completeness was assessed using BUSCO (Fig. 3D). One possible explanation for this could be an increase in transcript assembly fragmentation from de Bruijn graph assembly caused by the high heterozygosity previously described of the Parhyale genome (Kao et al., 2016). For long-read sequencing, we mapped the reads to the phaw_5.0 genome using minimap2 and assembled a transcriptome using StringTie2 (Kovaka et al., 2019; Li, 2018). We also used StringTie2 to generate a combined transcriptome containing both short and long reads. Although the long-read transcriptome (StringTie2 L) yielded a low BUSCO score, the short-+long-read transcriptome (StringTie2 SL) scored comparably with our other transcriptomes (Fig. 3D).

Short-read RNA-seq data was used to generate both de novo and genome-guided transcriptomes using Trinity. For genome-guided assembly, reads were mapped using HISAT2. Long-read RNA-seq data was mapped to the most recent Parhyale genome (phaw_5.0) using minimap2 and assembled using StringTie2. A combined transcriptome using both HISAT2-mapped short reads and minimap2-mapped long reads was also generated using StringTie2. See supplementary Materials and Methods ('Trinity transcriptome assembly parameters' and 'StringTie2 transcriptome assembly parameters') for additional information about assembly parameters.

Mikado transcriptome
Short-read Trinity transcriptomes and the transcriptome from Kao et al. (2016) were mapped to the phaw_5.0 genome using GMAP. Short-read transcriptomes and long-read StringTie2 transcriptomes were merged using the Mikado software along with a previous genome annotation generated by

Leo Blondel (Harvard University, MA, USA) (using MAKER). See supplementary Materials and Methods for additional information about Mikado parameters.

Gene model completeness analysis
To quantify the level of gene model fragmentation in our dataset, we generated a series of manual gene annotations based on RACE sequences. Among 143 previously-generated RACE transcripts, we selected 49 multi-exonic transcripts that appeared to have a single promoter based on the Omni-ATAC-seq data (Fig. S5D). For each of these RACE transcripts, we manually annotated the extent of the first and last exon by comparing RNA-seq read pileups to RACE data and the current genome annotation. We also identified the Omni-ATAC-seq peak most likely to capture the gene promoter, based on the strength of the peak (as evaluated by the number of time points over which we observed a statistically significant peak), as well as overlap to RNA-seq read data. We used these manual annotations (see Table S2) to evaluate the gene models from each of the different transcript sources (transcriptomes or gene annotations).

For each of the 49 genes in our dataset, we evaluated whether any models in each of the transcript sources in our dataset overlapped the promoter peak (Fig. S5E). The ability to unambiguously identify promoters is essential for downstream analyses, including building reporter constructs or targeting CRISPR guides to the 5′-most end of genes. Among the transcript sources, the Trinity de novo transcriptomes had the highest proportion (0.96) of genes for which at least one model overlapped with the promoter peak. However, the Trinity models we observed often (0.98, 48/49 RACE genes examined for the Trinity-limb transcriptome) contained numerous spurious transcript fragments in introns, exons and 3′ untranslated region (UTR) (see Fig. S6 for summary statistics and examples). Among the remaining transcript sources, the Mikado transcriptome had the highest fraction of genes for which at least one model overlapped with the promoter peak (0.84).

In addition, we assessed the degree of fragmentation of gene models across the 49 RACE transcripts (Fig. S5D). For each gene, we determined first whether any single gene model spanned the first and last exon, representing a complete 'single' transcript. If there was not a single model, we next assessed whether two separate models overlapped with the first and last exon, or if all models only overlapped either the first or last exon. These results together formed a measure of transcript completeness. Comparing the different transcript sources, the StringTie2 SL transcriptome had the highest fraction (0.76) of complete 'single' transcripts, with the Mikado transcriptome having a slightly lower fraction of 'single' transcripts (0.73) (Fig. S5E).

Functional annotation of transcripts
To assess the quality of the two genome annotations, we performed automated functional annotations to assign gene names and functions. We used two approaches: eggNOG and OrthoFinder (Emms and Kelly, 2019; Huerta-Cepas et al., 2019). eggNOG is a rapid and lightweight genome annotation software that assigns gene names, KEGG pathway information and GO terms, among numerous other metrics, to gene models. OrthoFinder facilitates the identification of orthogroups between provided peptide libraries, enabling automated comparisons between gene lists in species of interest.

We observed that the Mikado transcriptome included a greater number of GO-term assigned genes, eggNOG named genes and eggNOG uniquely-named genes than were found in the MAKER annotation (Fig. 3E). We used OrthoFinder to assign orthogroups between the Mikado or MAKER transcriptomes and the list of all D. melanogaster peptides from the UNIPROT-SWISSPROT database. The Mikado and MAKER transcriptomes produced similar numbers of orthogroups; however, the Mikado transcriptome produced a greater number of 1:1 orthogroups, which proved useful for downstream GO term enrichment analysis (summarized in Fig. 3E; see Fig. S7 for further explanation of orthogroup size comparisons). See supplementary Materials and Methods ('eggNOG annotation' and 'OrthoFinder annotation') for additional information about functional annotation.

DEVELOPMENT

ATAC-seq data

We performed conventional ATAC-seq as per Buenrostro et al. (2015) for stages S13-S22. Data were used for benchmarking in comparison to Omni-ATAC data, but were not used for downstream analyses.

Omni-ATAC-seq tagmentation

Tagmentation was performed using the reagents described in Corces et al. (2017) Omni-ATAC-seq paper, with the following modifications. Instead of the Illumina Nextera TD buffer, 2× Tagmentation Buffer from Wang et al. (2013) was used. Homemade Tn5 enzymes were purified and received as a gift from Jase Gehring (University of California, Berkeley, CA, USA; California Institute of Technology, CA, USA) and assembled into Tn5 transposomes as per Picelli et al. (2014), using adapters purchased from IDT with 5′ phosphorylation (Picelli et al., 2014). Before adding RSB+D, embryos were washed 3× using 1× PBS. Embryos were mashed into a near-uniform solution using the tip of a low-retention p10 pipette, and the pipette was visually inspected for any remaining debris. We used the Qiagen MinElute kit for purification and concentration of DNA.

Omni-ATAC-seq library preparation

Saturation PCR conditions for Omni-ATAC-seq libraries were performed using a Roche Lightcycler 480 as per Corces et al. (2017). Optimal conditions for additional amplification after pre-amplification were determined based on the number of additional PCR cycles corresponding to one-third of the maximum value, rounded up. Barcodes were added to samples using primers supplied by QB3-Berkeley using PCR as described in Corces et al. (2017). Following amplification, adapters were removed from libraries using Ampure bead purification and analyzed using a Bioanalyzer or Fragment Analyzer machine to assess library quality. Final libraries were pooled together to have relatively equal proportions based on additional qPCR quantification and size-selected using a Pippin Prep to remove fragments greater than 1.5 kb in size. All libraries were sequenced on one lane of each of Illumina NovaSeq SP 150PE and NovaSeq S1 150PE. Adapters used for libraries are listed in Table S1.

Omni-ATAC-seq quality control

We employed a battery of standard tests used to assess the quality of ATAC-seq data (Fig. S2A). Omni-ATAC-seq libraries were also compared with previous ATAC-seq libraries, and had greater library size, higher numbers of non-duplicated reads and lower amounts of mitochondrial DNA contamination (Fig. S2B,C). Fragment size analysis of Omni-ATAC-seq reads after mapping revealed distinct 1-nucleosome and 2-nucleosome peaks, indicating that tagmentation was efficient (Fig. S2E,F).

To determine whether our Omni-ATAC-seq data represented genuine enrichment in accessible chromatin regions, we examined promoter accessibility genome-wide. A hallmark of successful ATAC-seq experiments is strong enrichment of reads in gene promoters (Yan et al., 2020). We evaluated our ATAC-seq using the most recent phaw_5.0 gene annotations (obtained from Leo Blondel, hereafter referred to as the 'MAKER annotation') at the start of annotated mRNAs and across mRNA lengths, and compared our results with RNA-seq read pileups (Fig. 2A,B). As expected, Omni-ATAC-seq signal is enriched symmetrically at mRNA starts, whereas RNA-seq reads are enriched 3′ of mRNA starts. In addition, Omni-ATAC-seq data show greater enrichment at promoters than over annotated mRNA regions, and decreased enrichment outside of annotated mRNAs. These data suggest that our Omni-ATAC-seq performed as expected in identifying promoter regions genome-wide.

Adapter trimming was performed using Trim Galore, which leverages Cutadapt and FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/; https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/; Martin, 2011). FASTQC was also performed before and after adapter trimming to confirm removal of sequencing and Tn5 adapters. Percentage of reads remaining after deduplication was estimated based on FASTQC metrics. Library size was estimated by multiplying raw read count by percentage of reads remaining after deduplication for each lane, and then summing the two lanes. Reads were aligned to the phaw_5.0 genome as well as the *Parhyale* chrM using Bowtie2, and percentage read mapping was determined using Bowtie2 output (Langmead and Salzberg, 2012). After peak calling, merged peaks were used to evaluate correlation between replicates. Replicates were highly correlated, with a mean Pearson correlation of 0.988 and a mean Spearman correlation of 0.936 (Fig. S2H,I).

Omni-ATAC peak calling

Aligned reads were quality-filtered with a *q*-value of 10 in samtools and used for Genrich analysis. Genrich analysis was run on both duplicate libraries simultaneously; Genrich performs peak calling on each peak individually, and then merges the *P*-values of the replicates using Fisher's method to generate a *q*-value, obviating the need to calculate an irreproducible discovery rate. Output .bedgraph-like files from Genrich were reformatted using a custom Python script to be standard .bedgraph files, which were converted to bigWig files using the bedGraphToBigWig executable from the University of California, Santa Cruz Genome Tools software bundle. Output .narrowPeak files from Genrich were converted into .bed files for ease of visualization in Integrative Genomics Viewer using a custom Python script (https://github.com/mezarque/Parhyale_Genome_Resources).

IDE2 concept and analysis

IDE2 differs from other software for differential expression analysis in that it allows the investigation of trajectories of dynamic expression over large numbers of time points. It does so by modeling a gene expression trajectory as an 'impulse' function that is the product of two sigmoid functions (Chechik and Koller, 2009; Yosef and Regev, 2011). This approach enables the modeling of a trajectory of gene expression in three parts: an initial value, a peak value and a steady state value, thus summarizing an expression trajectory using a fixed number of parameters. With the ability to capture the differences between early, middle and late expression values for each gene in a dataset, IDE2 also enables the detection of transient changes in gene expression or accessibility during a time course. Identifying differential expression over large numbers of time points is difficult for more categorical differential expression software such as edgeR and DESeq2, which generally use pairwise comparisons between time points to assess change over time (Love et al., 2014; Robinson et al., 2010).

Assigning spatial categories to Omni-ATAC peaks

We first assigned the nearest peak within 5 kb of the first 200 bp of each gene (mRNA and ncRNA) as a promoter peak using bedtools closest. We then assigned the remaining peaks into categories based on their position relative to mRNA and ncRNA annotations. Peaks that overlapped with mRNA and ncRNA annotations were assigned as exonic or intronic regulatory elements. The remaining peaks – those which had not been classified as promoters, and which did not overlap with genes – were classified as intergenic peaks. The intergenic peaks were divided into two categories: proximal and distal intergenic peaks. Proximal peaks were those <10 kb away from the nearest gene, whereas distal intergenic peaks were those >10 kb away from the nearest gene. We established this cutoff with the rationale that peaks beyond this distance would be considerably more difficult to isolate as single fragments combined with a promoter peak using PCR, agnostic of their orientation with respect to the promoter element. Such peaks could not have been easily identified using previous approaches, and thus differ from 'proximal' peaks by their necessary identification using genomic techniques.

Fuzzy clustering concept and analysis using Mfuzz

Fuzzy clustering differs from categorical clustering approaches, such as hierarchical clustering and k-means clustering, in that it assigns a probability for each element in the dataset to fall into each of the identified clusters. For example, rather than a given peak falling in either cluster A or B exclusively, a given peak might have a 95% chance of falling in cluster A and a 5% chance of falling in cluster B.

The Mfuzz package allows the specification of two parameters: c, the number of clusters to create from the data, and m, the stringency of the

clustering. To determine the optimal number of clusters in our fuzzy c-means clustering, we varied the c-value from 3 to 13, and evaluated the quality of clustering using the cluster overlap metric provided by Mfuzz, as well as the silhouette score metric (Fig. S10) (Kumar and E Futschik, 2007). We determined that nine clusters yielded the highest mean silhouette score, while avoiding clusters below the silhouette score average and minimizing size variability between clusters. Fig. 5A plots with t-stochastic neighbor embedding (t-SNE; implemented in scikit-learn) (Hinton and Roweis, 2003; Van der Maaten and Hinton, 2008; Pedregosa et al., 2011) the top 10,000 most significant peaks called by IDE2, where each point is colored by its IDE2 model max fit, while Fig. 5B illustrates the same clusters, colored by their Mfuzz cluster. Fig. 5C indicates the spatial categories of peaks in each cluster.

Differential expression and accessibility analyses

We performed differential expression analysis using IDE2 and DESeq2 using standard settings. To generate the Omni-ATAC-seq read count matrix, we used bedtools multicov, using the merged Genrich peaks as our regions of interest. For our IDE2 analyses using the Omni-ATAC-seq data, IDE2 model fits were extracted from the IDE2 output and used to visualize model expression using a custom Python script (see supplementary Materials and Methods; 'DESeq2 parameters' and 'ImpulseDE2 analysis'). To generate the RNA-seq read count matrix for DESeq2, we generated a gene_trans_map file for the Mikado transcriptome, as would be available for the Trinity RNA-seq analysis pipeline, and used the built-in Trinity differential expression pipeline (align_and_estimate_abundance.pl, abundance_estimates_to_matrix.pl) with Kallisto to generate a matrix of read counts. To comply with the requirement for integer counts in DESeq2 analysis, we rounded each value to the nearest whole number.

NucleoATAC nucleosome predictions

Quality-filtered reads from each biological duplicate were merged and analyzed using NucleoATAC, with genomic regions set as ±500 bp windows around Genrich peaks.

HINT-ATAC transcription factor footprinting

Quality-filtered reads from each biological duplicate were merged and analyzed using rgt-hint footprinting. We used the JASPAR2020 database and converted the position weight matrices from JASPAR format into a simple matrix format expected by RGT-HINT using the R package 'universalmotif' and generated a '.mtf' file to store database information (see supplementary Materials and Methods). For enrichment analyses, we used bedtools random to generate 13 million random 20 bp sequences, as this was the average footprint size of genuine footprints detected by RGT-HINT in our data. This set of random sequences was used as background for our enrichment analyses. For cluster-specific enrichment analyses, we collated all unique transcription factor footprints from all developmental stages for each cluster (e.g. all footprints across S13, S14, etc. for all peaks in a given cluster) and compared enrichment levels with our randomly generated background.

Candidate reporter selection approaches

Careful examination of developmentally important genes revealed that many are surrounded by large numbers of peaks (>10) spread over large genomic distances. To further filter our Omni-ATAC peaks, we identified regions of sequence conservation to another amphipod crustacean, *Hyalella azteca* (Poynton et al., 2018), using the VISTA sequence alignment software (Ratnere and Dubchak, 2009) (Fig. S15A). *Hyalella* serves as a useful comparison to *Parhyale*, as its genome size is smaller (1.05 Gb) but, as it is also an amphipod crustacean, we expect that key developmental regulatory elements might have some level of sequence conservation.

Among our Omni-ATAC peaks, we took two strategies to identify candidate reporter elements. In the first approach, we identified all peaks within 5 kb of mRNA starts. In doing so, we were able to identify a handful of genomic regions in which we were able to locate both a putative promoter and a candidate proximal enhancer. We also identified several putative promoters, which we tested in isolation. In the second approach, we examined the genomic regions around important developmental genes of

interest. Many of the genes we examined showed very large numbers of strong peaks, and were therefore intractable to thorough analysis. For the purposes of this study, we focused on three regions: the region around the *Engrailed-1* gene, the region around the *Sp-69* gene and the region around the *Heat shock protein 70* complex, where two previous cis-regulatory elements (PhMS and HS2a) had been identified.

Minos transposon cloning

Minos transposon reporter plasmids were cloned using Gibson homology-mediated cloning approaches and the New England Biolabs Gibson Assembly or NEBuilder kits. As a base plasmid, we used the pMi(ne1) plasmid, which contains the Hsp70 minimal promoter, a DsRed protein sequence and an SV40 3′ UTR sequence, as well as two Minos inverted repeats. For plasmids containing the Hsp70 minimal promoter, the insert was integrated between the EcoRV and BglII restriction sites. For plasmids containing an endogenous promoter, the insert replaced the sequence between the EcoRV and NcoI restriction sites, thereby removing the Hsp70 minimal promoter. Cloned plasmids were Sanger sequenced to confirm a full DsRed open reading frame (ORF) and inclusion of desired genomic sequences.

Minos transposase assay

Minos transposase mRNA was generated using the Thermo Fisher Scientific mMESSAGE mMACHINE T7 or T7 ULTRA kit using NotI-digested pBlueSK-MimRNA (Addgene plasmid #102535). mRNA and concentrated DNA were mixed into a final concentration of 1 μg/μl in a solution of 0.1% phenol red in nuclease-free water. One- and two-cell *Parhyale* embryos were injected with ~3-5 pl of injection mix using a borosilicate glass capillary needle pulled using a Sutter P-80 or P-85 instrument. Embryos were raised until hatching and examined once per day from 3 dpf until 10 dpf using a Zeiss LSM780 confocal microscope to screen for DsRed fluorescence.

NucleoATAC signal, occupancy, insert and other miscellaneous data; HINT-ATAC transcription factor footprinting predictions and enrichment statistics; JASPAR2020 converted database and .mtf file for HINT-ATAC analyses; and differential accessibility analysis tables. Downstream analysis files for RNA-seq include BigWig files for visualizing RNA-seq read pileups; Nanopore alignment sequences; individual transcriptome GFF and FASTA files; eggNOG annotations and OrthoFinder annotations of transcripts; and RNA-seq differential expression analysis tables. In addition, the GEO database contains the updated Mikado genome annotation and assignment of peaks based on their position relative to genes, along with classification data for peaks as concordant or discordant in expression and accessibility. Code for analysis and visualizations is available at GitHub (https://github.com/mezarque/Parhyale_Genome_Resources). Software version numbers are cataloged in Table S3.

Peer review history
The peer review history is available online at https://journals.biologists.com/dev/article-lookup/doi/10.1242/dev.200793.

References

Borsari, B., Villegas-Mirón, P., Pérez-Lluch, S., Turpin, I., Laayouni, H., Segarra-Casas, A., Bertranpetit, J., Guigó, R. and Acosta, S. (2021). Enhancers with tissue-specific activity are enriched in intronic regions. *Genome Res.* **31**, 1325-1336. doi:10.1101/gr.270371.120

Brogaard, K., Xi, L., Wang, J.-P. and Widom, J. (2012). A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**, 496-501. doi:10.1038/nature11142

Browne, W. E., Price, A. L., Gerberding, M. and Patel, N. H. (2005). Stages of embryonic development in the amphipod crustacean, Parhyale hawaiensis. *Genesis* **42**, 124-149. doi:10.1002/gene.20145

Bruce, H. S. and Patel, N. H. (2020). Knockout of crustacean leg patterning genes suggests that insect wings and body walls evolved from ancient leg segments. *Nat. Ecol. Evol.* **4**, 1703-1712. doi:10.1038/s41559-020-01349-0

Buenrostro, J. D., Wu, B., Chang, H. Y. and Greenleaf, W. J. (2015). ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1-21.29.9. doi:10.1002/0471142727.mb2129s109

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421. doi:10.1186/1471-2105-10-421

Cazet, J. F., Cho, A. and Juliano, C. E. (2021). Generic injuries are sufficient to induce ectopic Wnt organizers in Hydra. *ELife* **10**, e60562. doi:10.7554/eLife.60562

Chechik, G. and Koller, D. (2009). Timing of gene expression responses to environmental changes. *J. Comput. Biol.* **16**, 279-290. doi:10.1089/cmb.2008.13TT

Christian, S.-E., Schultheis, D., Schwirz, J., Ströhlein, N., Troelenberg, N., Majumdar, U., Dao, V., Grossmann, D., Richter, T., Tech, M. et al. (2015). The iBeetle large-scale RNAi screen reveals gene functions for insect development and physiology. *Nat. Commun.* **6**, 7822. doi:10.1038/ncomms8822

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B. et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959-962. doi:10.1038/nmeth.4396

Crawford, K., Diaz Quiroz, J. F., Koenig, K. M., Ahuja, N., Albertin, C. B. and Rosenthal, J. J. C. (2020). Highly efficient knockout of a squid pigmentation gene. *Curr. Biol.* **30**, 3484-3490.e4. doi:10.1016/j.cub.2020.06.099

Emms, D. M. and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238. doi:10.1186/s13059-019-1832-y

Fischer, D. S., Theis, F. J. and Yosef, N. (2018). Impulse model-based differential expression analysis of time course sequencing data. *Nucleic Acids Res.* **46**, e119. doi:10.1093/nar/gky675

Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D. et al. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87-D92. doi:10.1093/nar/gkz1001

Gatzmann, F., Falckenhayn, C., Gutekunst, J., Hanna, K., Raddatz, G., Carneiro, V. C. and Lyko, F. (2018). The methylome of the marbled crayfish links gene body methylation to stable expression of poorly accessible genes. *Epigenet. Chromatin* **11**, 57. doi:10.1186/s13072-018-0229-6

Gehrke, A. R., Schneider, I., de la Calle-Mustienes, E., Tena, J. J., Gomez-Marin, C., Chandran, M., Nakamura, T., Braasch, I., Postlethwait, J. H., Gómez-Skarmeta, J. L. et al. (2015). Deep conservation of wrist and digit enhancers in fish. *Proc. Natl. Acad. Sci. USA* **112**, 803-808. doi:10.1073/pnas.1420208112

Gehrke, A. R., Neverett, E., Luo, Y.-J., Brandt, A., Ricci, L., Hulett, R. E., Gompers, A., Ruby, J. G., Rokhsar, D. S., Reddien, P. W. et al. (2019). Acoel genome reveals the regulatory landscape of whole-body regeneration. *Science* **363**, eaau6173. doi:10.1126/science.aau6173

Gilbert, D. G. (2019). Longest protein, longest transcript or most expression, for accurate gene reconstruction of transcriptomes? *bioRxiv*, 829184. doi:10.1101/829184

Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. and Lieb, J. D. (2007). FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17**, 877-885. doi:10.1101/gr.5533506

Gisselbrecht, S. S., Palagi, A., Kurland, J. V., Rogers, J. M., Ozadam, H., Zhan, Y., Dekker, J. and Bulyk, M. L. (2020). Transcriptional silencers in drosophila serve a dual role as transcriptional enhancers in alternate cellular contexts. *Mol. Cell* **77**, 324-337.e8. doi:10.1016/j.molcel.2019.10.004

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644-652. doi:10.1038/nbt.1883

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M. et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494-1512. doi:10.1038/nprot.2013.084

Halfon, M. S. (2020). Silencers, enhancers, and the multifunctional regulatory genome. *Trends Genet.* **36**, 149-151. doi:10.1016/j.tig.2019.12.005

Hinton, G. E. and Roweis, S. (2003). Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems* (ed. S. Becker, S. Thrun and K. Obermayer), pp. 857-864. MIT Press.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J. et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309-D314. doi:10.1093/nar/gky1085

Kao, D., Lai, A. G., Stamataki, E., Rosic, S., Konstantinides, N., Jarvis, E., Donfrancesco, A., Natalia, P.-S., Semon, M., Grillo, M. et al. (2016). The genome of the crustacean Parhyale hawaiensis, a model for animal development, regeneration, immunity and lignocellulose digestion. *Elife* **5**, 065789. doi:10.7554/elife.20062

Kissane, S., Dhandapani, V. and Orsini, L. (2021). Protocol for assay of transposase accessible chromatin sequencing in non-model species. *STAR Protoc.* **2**, 100341. doi:10.1016/j.xpro.2021.100341

Konstantinides, N. and Averof, M. (2014). A common cellular basis for muscle regeneration in arthropods and vertebrates. *Science* **343**, 788-791. doi:10.1126/science.1243529

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L. and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278. doi:10.1186/s13059-019-1910-1

Kumar, L. and E Futschik, M. (2007). Mfuzz: a software package for soft clustering of microarray data. *Bioinformation* **2**, 5-7. doi:10.6026/97320630002005

Lai, Y.-T., Deem, K. D., Borràs-Castells, F., Sambrani, N., Rudolf, H., Suryamohan, K., El-Sherif, E., Halfon, M. S., McKay, D. J. and Tomoyasu, Y. (2018). Enhancer identification and activity evaluation in the red flour beetle, Tribolium castaneum. *Development* **145**, dev160663. doi:10.1242/dev.160663

Langmead, B. and Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359. doi:10.1038/nmeth.1923

Legendre, P. and Legendre, L. (2012). *Numerical Ecology*. Elsevier.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100. doi:10.1093/bioinformatics/bty191

Li, Y., Chen, C., Kaye, A. M. and Wasserman, W. W. (2015). The identification of cis-regulatory elements: a review from a machine learning perspective. *Biosystems* **138**, 6-17. doi:10.1016/j.biosystems.2015.10.002

Li, Z., Schulz, M. H., Look, T., Begemann, M., Zenke, M. and Costa, I. G. (2019). Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* **20**, 45. doi:10.1186/s13059-019-1642-2

Liubicich, D. (2007). The role of Hox genes in crustacean development and appendage specialization. PhD thesis, University of California, Berkeley, Berkeley, CA.

Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. doi:10.1186/s13059-014-0550-8

Ludwig, M., Patel, N. and Kreitman, M. (1998). Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development* **125**, 949-958. doi:10.1242/dev.125.5.94

Mahony, S. and Benos, P. V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* **35**, W253-W258. doi:10.1093/nar/gkm272

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**, 10-12. doi:10.14806/ej.17.1.200

Martin, A., Serano, J. M., Jarvis, E., Bruce, H. S., Wang, J., Ray, S., Barker, C. A., O'Connell, L. C. and Patel, N. H. (2016). CRISPR/Cas9 mutagenesis reveals versatile roles of hox genes in crustacean limb specification and evolution. *Curr. Biol.* **26**, 14-26. doi:10.1016/j.cub.2015.11.021

Mito, T., Nakamura, T., Bando, T., Ohuchi, H. and Noji, S. (2011). The advent of RNA interference in entomology. *Entomol. Sci.* **14**, 1-8. doi:10.1111/j.1479-8298.2010.00408.x

DEVELOPMENT

Morton, J. T., Toran, L., Edlund, A., Metcalf, J. L., Lauber, C., Knight, R. and Jansson, J. K. (2017). Uncovering the horseshoe effect in microbial analyses. *mSystems* 2, e00166-16. doi:10.1128/mSystems.00166-16

Paris, M., Wolff, C., Patel N. H. and Averof, M. (2022). Chapter Eight - The crustacean model Parhyale hawaiensis. In *Current Topics in Developmental Biology* (ed. B. Goldstein and M. Srivastava), pp. 199-230. Academic Press.

Pavlopoulos, A. and Averof, M. (2005). Establishing genetic transformation for comparative developmental studies in the crustacean Parhyale hawaiensis. *Proc. Natl. Acad. Sci. USA* 102, 7888-7893. doi:10.1073/pnas.0501101102

Pavlopoulos, A., Kontarakis, Z., Liubicich, D. M., Serano, J. M., Akam, M., Patel, N. H. and Averof, M. (2009). Probing the evolution of appendage specialization by Hox gene misexpression in an emerging model crustacean. *Proc. Natl. Acad. Sci. USA* 106, 13897-13902. doi:10.1073/pnas.0902804106

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825-2830.

Pérez-Zamorano, B., Rosas-Madrigal, S., Lozano, O. A. M., Castillo Méndez, M. and Valverde-Garduño, V. (2017). Identification of cis-regulatory sequences reveals potential participation of lola and Deaf1 transcription factors in Anopheles gambiae innate immune response. *PLoS One* 12, e0186435. doi:10.1371/journal.pone.0186435

Picelli, S., Björklund, Å. K., Reinius, B., Sagasser, S., Winberg, G. and Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively-scaled sequencing projects. *Genome Res.* 24, 2033-2040. doi:10.1101/gr.177881.114

Podani, J. and Miklós, I. (2002). Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology* 83, 3331-3343. doi:10.1890/0012-9658(2002)083[3331:RCATHE]2.0.CO;2

Poynton, H. C., Hasenbein, S., Benoit, J. B., Sepulveda, M. S., Poelchau, M. F., Hughes, D. S. T., Murali, S. C., Chen, S., Glastad, K. M., Goodisman, M. A. D. et al. (2018). The toxicogenome of hyalella azteca: a model for sediment ecotoxicology and evolutionary toxicology. *Environ. Sci. Technol.* 52, 6009-6022. doi:10.1021/acs.est.8b00837

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841-842. doi:10.1093/bioinformatics/btq033

Radman-Livaja, M. and Rando, O. J. (2010). Nucleosome positioning: how is it established, and why does it matter? *Dev. Biol.* 339, 258-266. doi:10.1016/j.ydbio.2009.06.012

Ramos, A. P., Gustafsson, O., Labert, N., Salecker, I., Nilsson, D.-E. and Averof, M. (2019). Analysis of the genetically tractable crustacean Parhyale hawaiensis reveals the organisation of a sensory system for low-resolution vision. *BMC Biol.* 17, 67. doi:10.1186/s12915-019-0676-y

Rasys, A. M., Park, S., Ball, R. E., Alcala, A. J., Lauderdale, J. D. and Menke, D. B. (2019). CRISPR-Cas9 gene editing in lizards through microinjection of unfertilized oocytes. *Cell Rep.* 28, 2288-2292.e3. doi:10.1016/j.celrep.2019.07.089

Ratnere, I. and Dubchak, I. (2009). Obtaining comparative genomic data with the VISTA family of computational tools. *Curr. Protoc. Bioinform.* 26, 10.6.1-10.6.17. doi:10.1002/0471250953.bi1006s26

Rehm, E. J., Hannibal, R. L., Chaw, R. C., Vargas-Vila, M. A. and Patel, N. H. (2009). Fixation and dissection of parhyale hawaiensis embryos. *Cold Spring Harb. Protoc.* 2009, pdb.prot5127. doi:10.1101/pdb.prot5127

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

Rouhana, L., Weiss, J. A., Forsthoefel, D. J., Lee, H., King, R. S., Inoue, T., Shibata, N., Agata, K. and Newmark, P. A. (2013). RNA interference by feeding in vitro-synthesized double-stranded RNA to planarians: methodology and dynamics. *Dev. Dyn.* 242, 718-730. doi:10.1002/dvdy.23950

Schep, A. N., Buenrostro, J. D., Denny, S. K., Schwartz, K., Sherlock, G. and Greenleaf, W. J. (2015). Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.* 25, 1757-1770. doi:10.1101/gr.192294.115

Sharma, P. P., Schwager, E. E., Giribet, G., Jockusch, E. L. and Extavour, C. G. (2013). Distal-less and dachshund pattern both plesiomorphic and apomorphic structures in chelicerates: RNA interference in the harvestman Phalangium opilio (Opiliones). *Evol. Dev.* 15, 228-242. doi:10.1111/ede.12029

Srivastava, M., Mazza-Curll, K. L., van Wolfswinkel, J. C. and Reddien, P. W. (2014). Whole-body acoel regeneration is controlled by Wnt and Bmp-admp signaling. *Curr. Biol.* 24, 1107-1113. doi:10.1016/j.cub.2014.03.042

Stamataki, E. and Pavlopoulos, A. (2016). Non-insect crustacean models in developmental genetics including an encomium to Parhyale hawaiensis. *Curr. Opin. Genet. Dev.* 39, 149-156. doi:10.1016/j.gde.2016.07.004

Sun, D. A. and Patel, N. H. (2019). The amphipod crustacean Parhyale hawaiensis: An emerging comparative model of arthropod development, evolution, and regeneration. *Wiley Interdiscip. Rev. Dev. Biol.* 8, e355. doi:10.1002/wdev.355

The UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506-D515. doi:10.1093/nar/gky1049

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579-2605. doi:10.1118/1.3267037

Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L. and Swarbreck, D. (2018). Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience* 7, giy093. doi:10.1093/gigascience/giy093

Wang, Q., Gu, L., Adey, A., Radlwimmer, B., Wang, W., Hovestadt, V., Bähr, M., Wolf, S., Shendure, J., Eils, R. et al. (2013). Tagmentation-based whole-genome bisulfite sequencing. *Nat. Protoc.* 8, 2022-2032. doi:10.1038/nprot.2013.118

Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 5, 276-287. doi:10.1038/nrg1315

Yan, F., Powell, D. R., Curtis, D. J. and Wong, N. C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* 21, 22. doi:10.1186/s13059-020-1929-3

Yosef, N. and Regev, A. (2011). Impulse control: temporal dynamics in gene transcription. *Cell* 144, 886-896. doi:10.1016/j.cell.2011.02.015
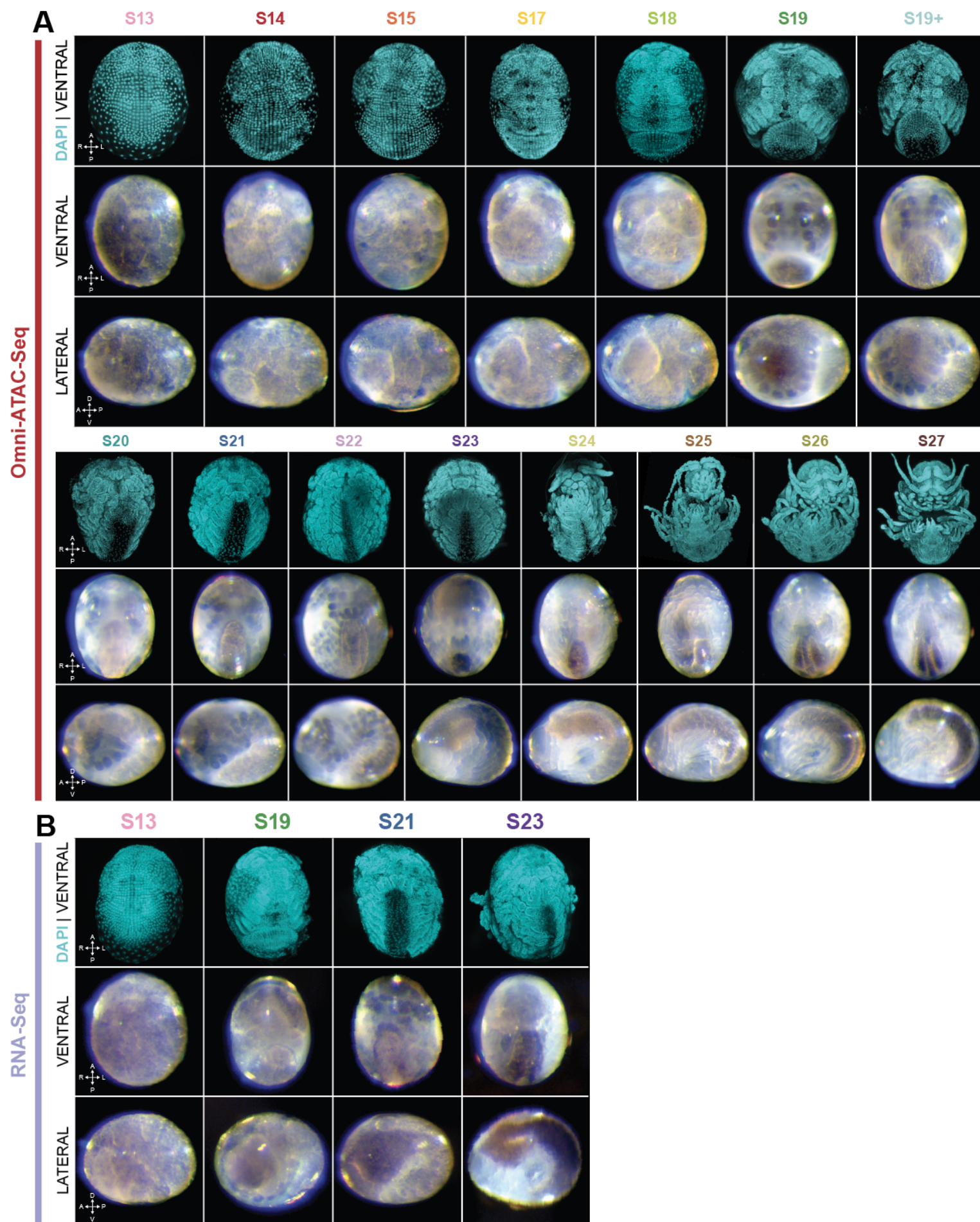
DEVELOPMENT

Fig. S1. Embryo images for time course Omni-ATAC-Seq and RNA-Seq

  A)   Representative embryo images from Omni-ATAC-Seq libraries. Top row shows ventral view of DAPI stained embryo. Bottom rows show ventral and lateral brightfield images of embryos shortly before tagmentation.

  B)   Representative embryo images from RNA-Seq libraries. Top row shows ventral view of DAPI stained embryo. Bottom rows show ventral and lateral brightfield images of embryos shortly before RNA extraction.

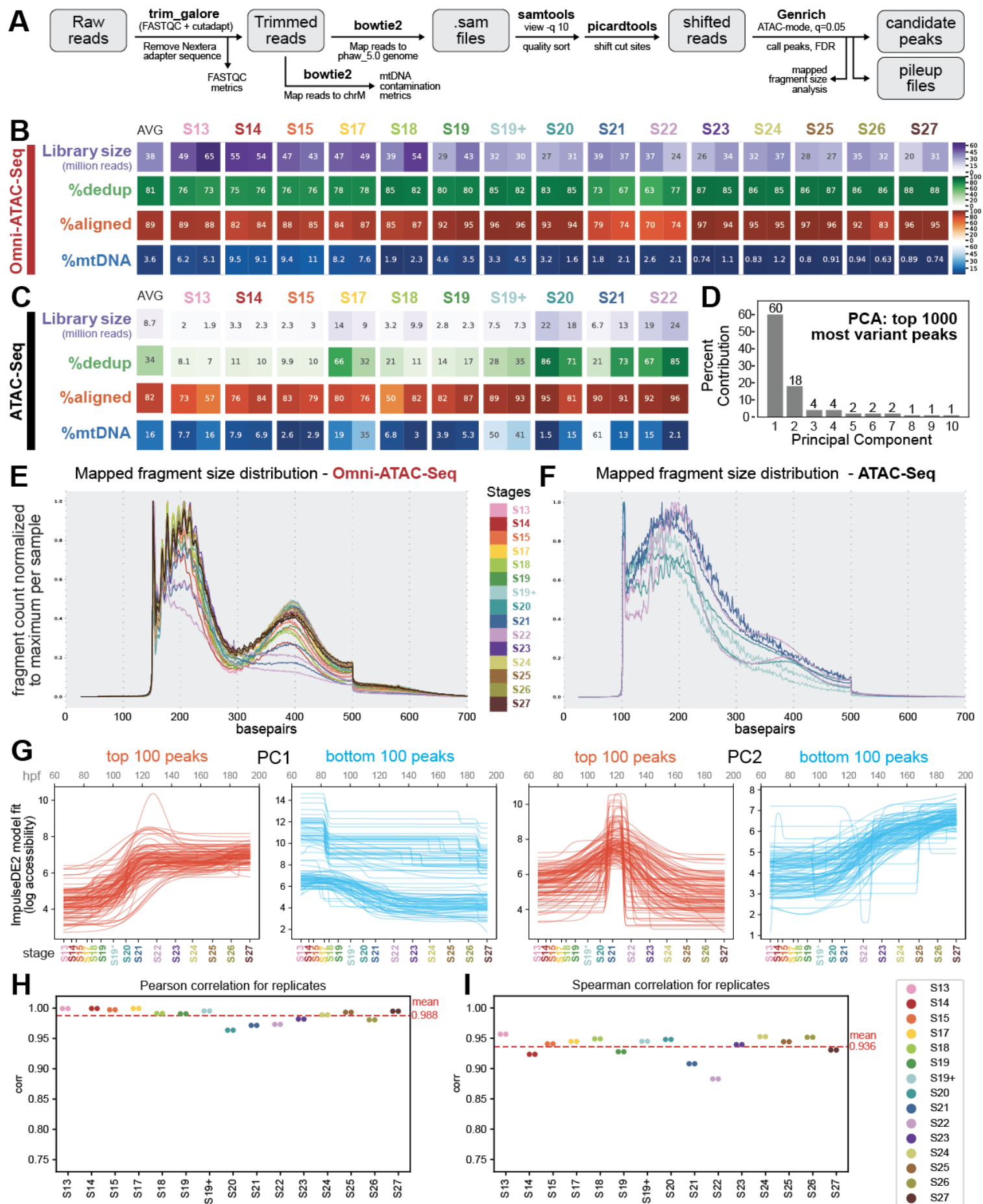(A: anterior, P: posterior, L: left, R: right; D: dorsal, V: ventral).

Fig. S2. Omni-ATAC-Seq analysis pipeline and quality control metrics

A) Analysis pipeline for generating Omni-ATAC peaks and QC metrics from raw reads. Reads were quality and adapter-trimmed using trim_galore, and trimmed reads were mapped to either the entire genome or to only the mitochondrial chromosome (ChrM) to generate a conservative estimate of mtDNA contamination. Mapped reads were further filtered using samtools (Q = 10) and read ends were shifted using Picardtools to reflect the mechanism of Tn5 insertion. Shifted reads were passed to Genrich in ATAC mode using a q-value cutoff of 0.05. Files produced by Genrich included stage-specific peaks and bedgraph-like read pileup formats.

B) Quality control metrics for Omni-ATAC-Seq libraries. Library size, percentage of reads remaining after deduplication (%dedup), percentage of aligned reads (%aligned), and percentage of reads mapping to the mitochondrial genome (%mtDNA) are displayed for individual Omni-ATAC libraries.

C) Metrics from B visualized for conventional ATAC-Seq libraries. Overall, Omni-ATAC seq shows similar or improved performance on all metrics.

D) Percentage contribution to variance of each of the top 10 principal components from PCA in Fig. 2C.

E) Mapped fragment size distribution of Omni-ATAC-Seq reads. Two clear peaks are seen in the fragment size distribution, reflecting proper sub-nucleosomal insertions.

F) Mapped fragment size distribution of conventional ATAC-Seq reads for 4 developmental stages: S19, S19+, S20, S21. Fragment size peaks are less distinct as compared to Omni-ATAC-Seq libraries.

G) Plots of ImpulseDE2 model fits to the top 100 and bottom 100 peaks loading for PC1 and PC2 in figure 1C. PC1 appears to be associated with developmental time, while PC2 appears to be associated with peaks that show an increase in accessibility in the middle of development.

H) Pearson correlation of mapped read counts in merged peaks between replicates for each developmental stage. Replicate libraries show a high average correlation value (0.988).

I) Spearman correlation of mapped read counts in merged peaks between replicates for each developmental stage. Replicate libraries show a high average correlation value (0.936). A lower correlation coefficient compared to Pearson may suggest that strong outliers contribute to the high Pearson correlation measure.
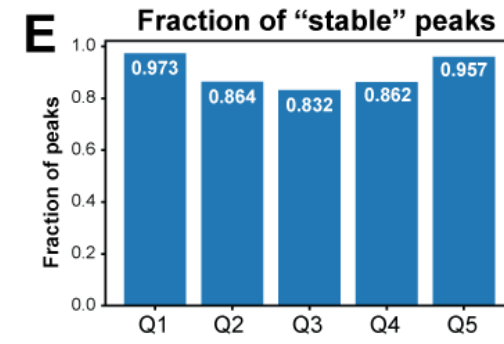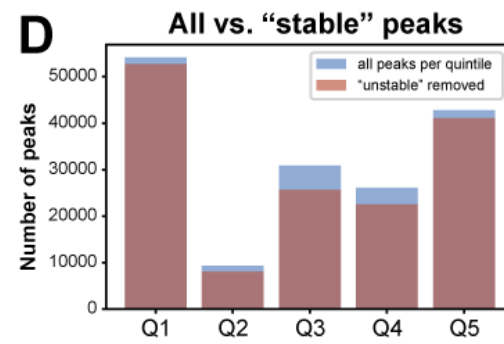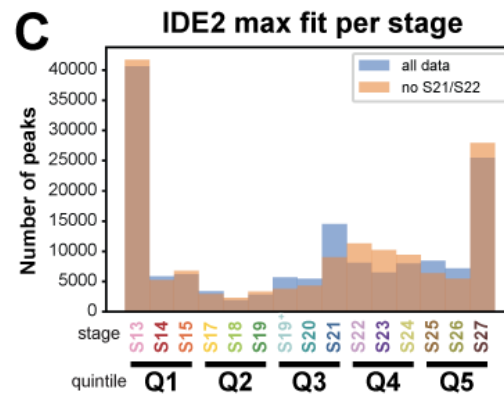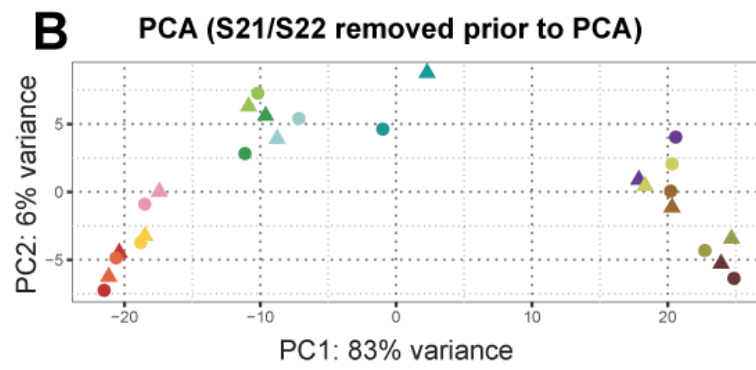
A PCA (all libraries)

B PCA (S21/S22 removed prior to PCA)

C IDE2 max fit per stage

D All vs. "stable" peaks
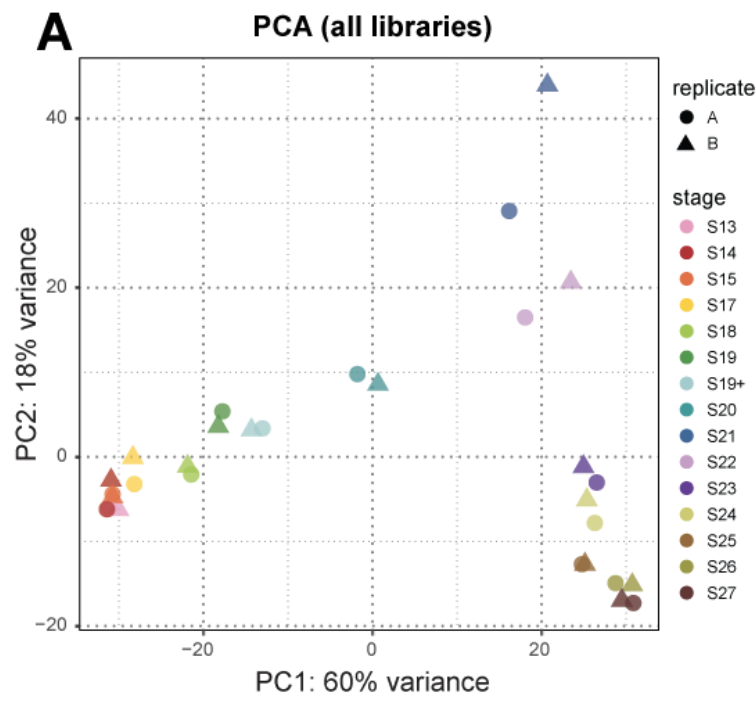
E Fraction of "stable" peaks

Fig. S3. Examining outlier effects of S21/S22 samples

A) Principal Components Analysis (PCA) of all libraries. Samples S21 and S22 appear highly separated in PC2 relative to other samples.

B) PCA of all libraries except S21 and S22 samples. The relative position of libraries in PC2 appears largely unchanged, with early and late time points showing negative loading in PC2, and middle timepoints showing higher loading in PC2. This suggests PC2 is not driven entirely by differences in the S21/ S22 samples.

C) ImpulseDE2 model max fit for all data (alldata, blue) or data excluding S21-S22 (noS21S22, orange). For each peak, the developmental time of maximum accessibility was estimated, and the results are summarized as a histogram. In the absence of S21-S22 samples, the same overall trend of early, middle, and late peak maximum fit is observed.

D) Peak stability analysis. The 15 developmental stages were broken into 5 quintiles of 3 stages each (label in C). For each peak, the IDE2 model max fit stage was calculated for alldata or noS21S22, and the shift in the model max fit was evaluated. Peaks that moved more than 3 developmental stages (equivalent to 1 quintile shift) were considered "unstable". The majority of peaks from each quintile were considered "stable" in the absence of S21/ S22 data.

E) Fraction of stable peaks from analysis in (D). For each quintile, >80% of peaks were considered "stable", suggesting that most peaks have the same trajectory regardless of the presence of S21/S22 data. This suggests that the remaining data are sufficient to reproduce the general accessibility trajectory of most peaks.

A phaw_50.283864a:3272910–3846006

B Peak-gene distance distribution

C En1 promoter : phaw_50.283864a_3320701_3323202

D En1-2_region (phaw_50.283864a:3,272,910-3,846,006)
peak strength:15 or higher

E Peak spatial categories over time

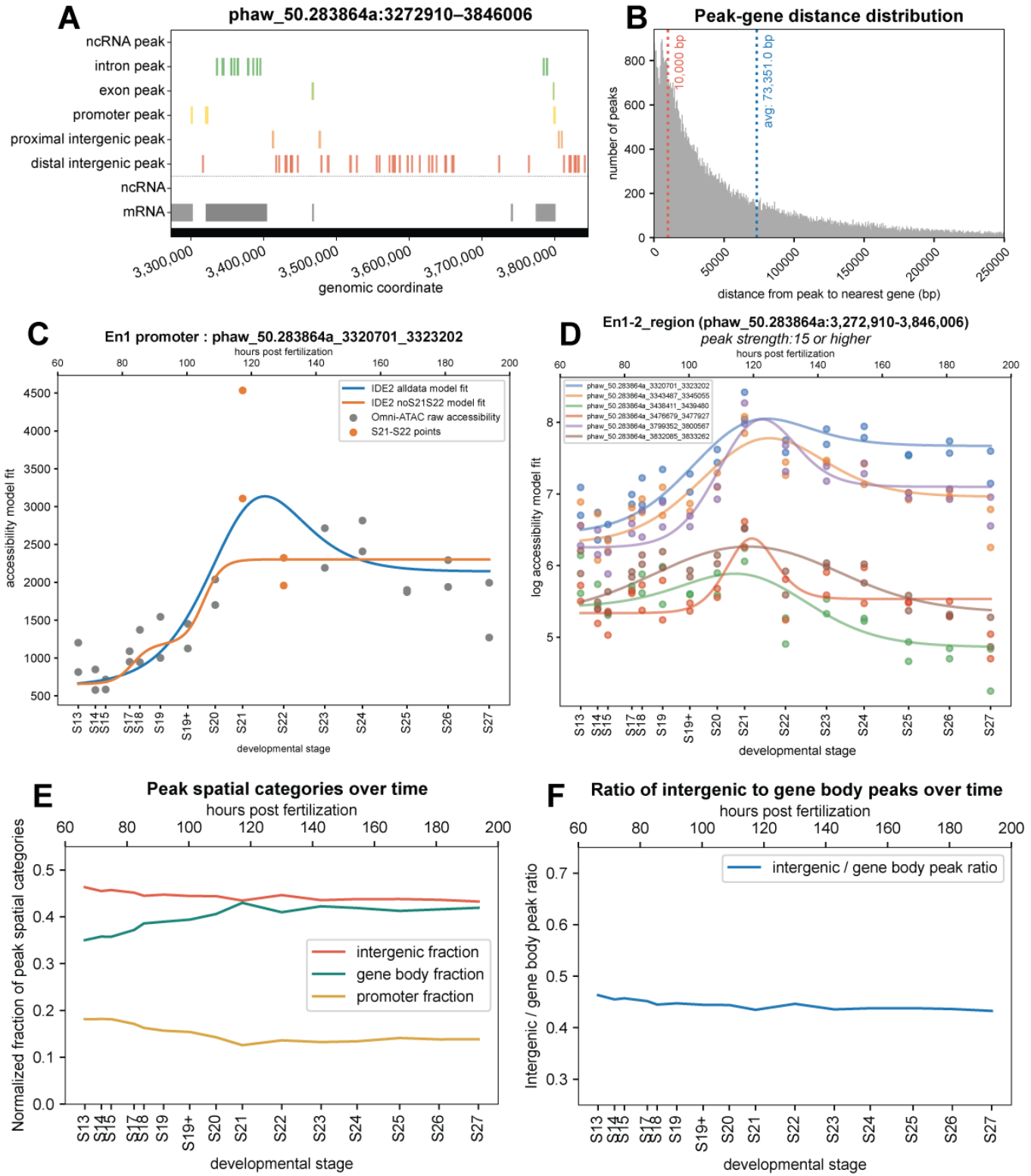**F** Ratio of intergenic to gene body peaks over time

Fig. S4. Peak category analyses and data visualization tools

A) Visualization of the genomic region around the *Parhyale Engrailed-1* and *Engrailed-2* loci using the Region Type Viewer found in the ParhyaleATACExplorer Jupyter notebook. Peak types are colored based on their identity. A 1000bp padding has been added to the 5' and 3' widths of each element, for ease of visualization.

B) Distribution of peak-nearest gene distances for intergenic peaks. Most peaks are "distal intergenic", or >10kb away (red dotted line) and could not have been identified easily using previous Minos reporter techniques. The average distance of intergenic peaks from the nearest gene is 73,351bp (blue dotted line).

C) Visualization of ImpulseDE2 model fits for all data (blue line) or noS21S22 data (orange line), plotted alongside raw accessibility values (points; stage S21 and S22 points are highlighted in orange), shown for the promoter peak of the *Engrailed-1* gene. Using the ImpulseDE2 Model Comparisons section of the ParhyaleATACExplorer Jupyter notebook, users can visualize the two model fits for arbitrary peaks of interest.

D) Visualization of ImpulseDE2 model fits for all data within a selected genomic region (same as shown in A), filtered by "peak strength" = 15, or the number of peaks found within a merged peak, summed over time. IDE2 model fits can be visualized for all peaks within an arbitrary genomic address, or a specific set of labeled peaks when provided as a Python dictionary.

E) Fraction of intergenic, gene body, and promoter peaks across developmental time, normalized for the number of peaks after removing "unknown" peaks. The fraction of intergenic and promoter peaks appears to slightly decline, while the fraction of gene body peaks increases over time.

F) Ratio of gene body peaks to intergenic peaks over time. The ratio appears to decline very slightly over time.

Fig. S5. Additional transcriptome and gene model evaluation metrics

A) Summarized completeness of RACE *Hox* gene models. Hox gene models for most genes are complete, including two isoforms each for *Ubx*, *abd-A*, and *Abd-B*. For *Hox3*, only a homeodomain sequence was isolated.

B) Summarized completeness of MAKER (phaw_5.0) *Hox* gene models. Many gene models show deviations from RACE sequences, including fragmentation, artefactual fusion of gene models, and extension of 5' UTRs.

C) Summarized completeness of *Hox* gene models from Mikado transcriptome. The Mikado transcriptome appears to have more complete *Hox* gene models than the MAKER genome annotation. In addition, a complete *Hox3* sequence was generated.

D) Model evaluation methods for gene model fragmentation. 49 RACE genes were selected based on the ability to unambiguously annotate a first exon, last exon, and promoter peak based on examination of all gene models, Omni-ATAC-Seq peaks, and Omni-ATAC-Seq and RNA-Seq read pileups. For each transcriptome or genome annotation, for each RACE gene, all models were compared to manual annotation windows using bedtools. If a single model from all models in a given transcriptome overlapped with the first and last exon of the manual annotation, that RACE gene was classified as a "single" model for that transcriptome. If no single model overlapped both first and last exons, but models existed that overlapped either, the RACE gene was classified as "split" for that transcriptome. If gene models only overlapped with manually annotated first or last exons, then the RACE gene was classified as "first" or "last". Finally, if no gene models overlapped, the RACE gene was classified as "no model".

E) Summary of gene model evaluations from D for select transcriptomes. Overall, the Trinity-S21 transcriptome had the highest number of models that overlapped the Omni-ATAC promoter peak, and the Mikado transcriptome had the second highest. The StringTie2 SL transcriptome had the highest number of single gene models, and the Mikado transcriptome had the second highest. For both metrics, the Mikado transcriptome outperformed the MAKER genome annotation.

F) BUSCO evaluation for additional transcriptomes included in the dataset.

G) Fraction of gene models attributed to each transcriptome in the final Mikado transcriptome. Overall, the StringTie2 SL transcriptome produced the greatest number of transcripts evaluated as "best" by the Mikado pipeline.
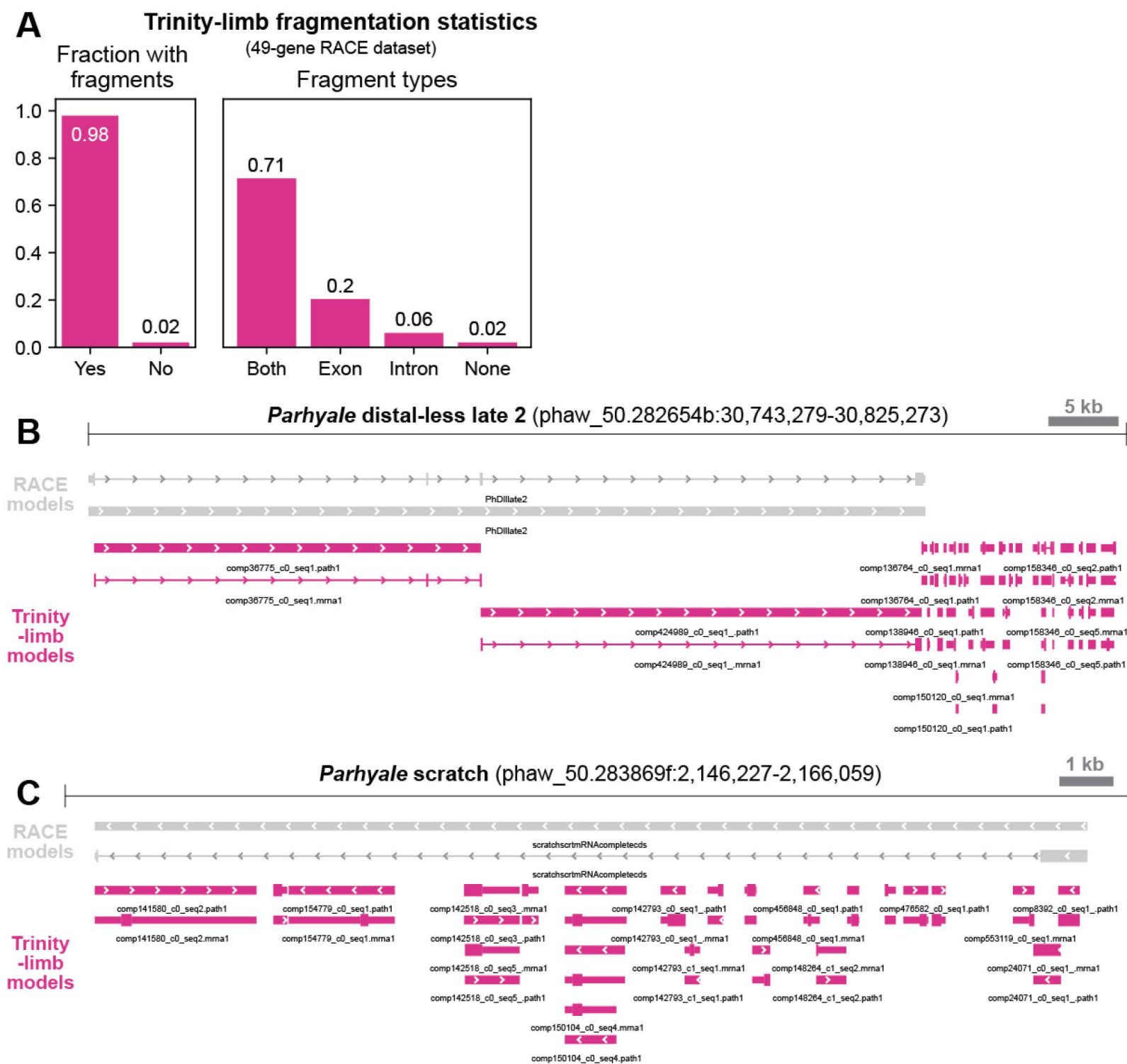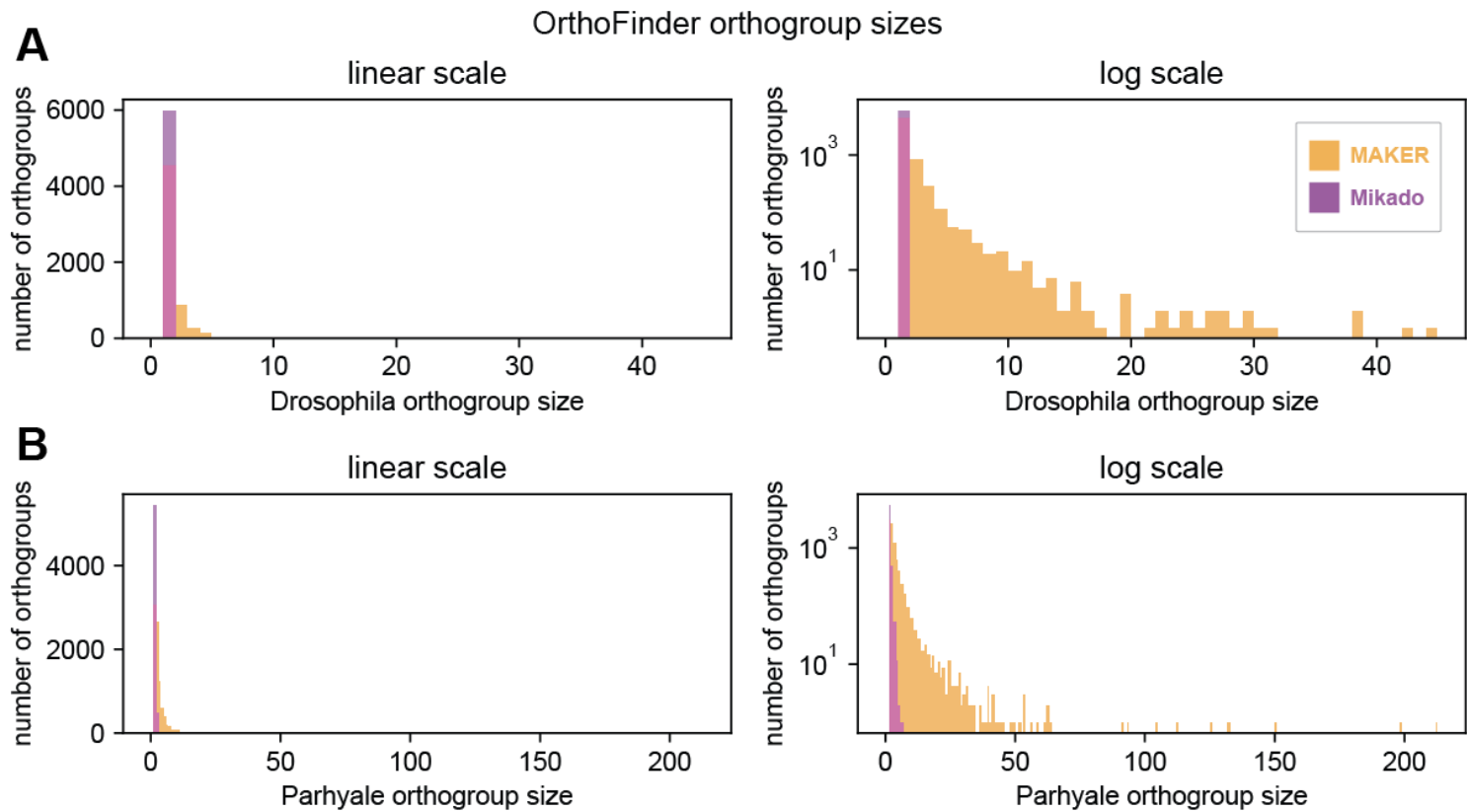
Fig. S6. Additional transcriptome and gene model evaluation metrics

A) Fraction of Trinity-limb (old) gene models with spurious transcript fragments when compared with RACE data. Fragments were observed overlapping with introns, exons, or both; some transcript fragments appeared to contain erroneous splice junctions joining exons and introns, while others were strictly intronic.

B) Example of transcript fragments compared to RACE data at the *Parhyale distal-less late 2* locus. This gene model had two transcript fragments overlapping with known exons, but no single gene model spanning the entire locus. Moreover, a large number of small 3'UTR-mapping fragments were observed, a signal that was frequently observed across genes in the RACE dataset.

C) Example of transcript fragments compared to RACE data at the *Parhyale scratch* locus. Transcript fragments that overlap with both intronic and exonic sequences are observed, as well as strictly intronic fragments. This gene model was a particularly extreme example of transcript fragmentation observed among the genes in the RACE dataset.

A OrthoFinder orthogroup sizes

linear scale / log scale — number of orthogroups vs Drosophila orthogroup size (MAKER, Mikado)

B linear scale / log scale — number of orthogroups vs Parhyale orthogroup size

C

Drosophila Orthologs	*Parhyale* Orthologs
MAKER OG0000000	
sp|Q9VNH6|EXOC4_DROME	augustus-phaw_50.000081a-processed-gene-10.10-mRNA-1.p1
	augustus-phaw_50.000135c-processed-gene-9.18-mRNA-1.p1
Exocyst complex component 4 (Sec8)	augustus-phaw_50.000135e-processed-gene-35.28-mRNA-1.p1
	augustus-phaw_50.000148b-processed-gene-9.12-mRNA-1.p1
	augustus-phaw_50.000203a-processed-gene-18.123-mRNA-1.p1
sp|Q9W5P1|MED21_DROME	augustus-phaw_50.000203b-processed-gene-55.179-mRNA-1.p1
	...
Mediator of RNA polymerase II transcription subunit 21 (MED21)	(total 213 genes)
	Numerous genes located on different contigs
Mikado OG0000000	
tr|Q9V436|Q9V436_DROME	mikado.phaw_50.282976aG39.1.p1
	mikado.phaw_50.282976aG39.2.p1
26S proteasome non-ATPase regulatory sub-unit 8 (Rpn12)	mikado.phaw_50.282976aG39.3.p1
	mikado.phaw_50.282976aG39.4.p1
	mikado.phaw_50.282976aG39.5.p1
	mikado.phaw_50.282976aG39.6.p1
	mikado.phaw_50.282976aG39.7.p1
	7 splice isoforms of 1 gene located on 1 contig

Fig. S7. Additional transcriptome and gene model evaluation metrics

A) Orthogroup sizes grouped by number of *Drosophila* orthologs in each orthogroup for the MAKER genome annotation and the Mikado transcriptome. The MAKER genome annotation, when annotated using OrthoFinder, produces many gene models that are grouped into orthogroups with large numbers of putative orthologs when compared to the results of using OrthoFinder on the Mikado transcriptome.

B) Orthogroup sizes grouped by number of *Parhyale* orthologs in each orthogroup for the MAKER genome annotation and the Mikado transcriptome. As with the *Drosophila* measure, the number of orthogroups with large numbers of orthologs is greater for the MAKER genome annotation.

C) Example of peptides spuriously classified as orthologs when comparing the MAKER genome annotation to the Mikado transcriptome. OG0000000 is the single largest orthogroup identified by OrthoFinder when comparing the *Drosophila* UNIPROT database to either the MAKER genome annotation or the Mikado transcriptome. In the MAKER annotation, the largest orthogroup contains two unrelated *Drosophila* proteins (Sec8 and MED21), which are grouped together along with 213 different *Parhyale* gene models scattered across the genome. Examination of other orthogroups revealed a similar result, in which unrelated *Drosophila* proteins were grouped along with numerous *Parhyale* gene models, impeding precise identification of orthologs between the species, and suggesting problems in the quality of the MAKER gene models. In the Mikado annotation, the largest orthogroup contains a single *Drosophila* protein (Rpn12) grouped along with 7 splice isoforms of a single *Parhyale* gene model located on a single contig. Examination of other orthogroups derived from the Mikado transcriptome revealed fewer obviously incorrect groupings than observed for the MAKER annotation.
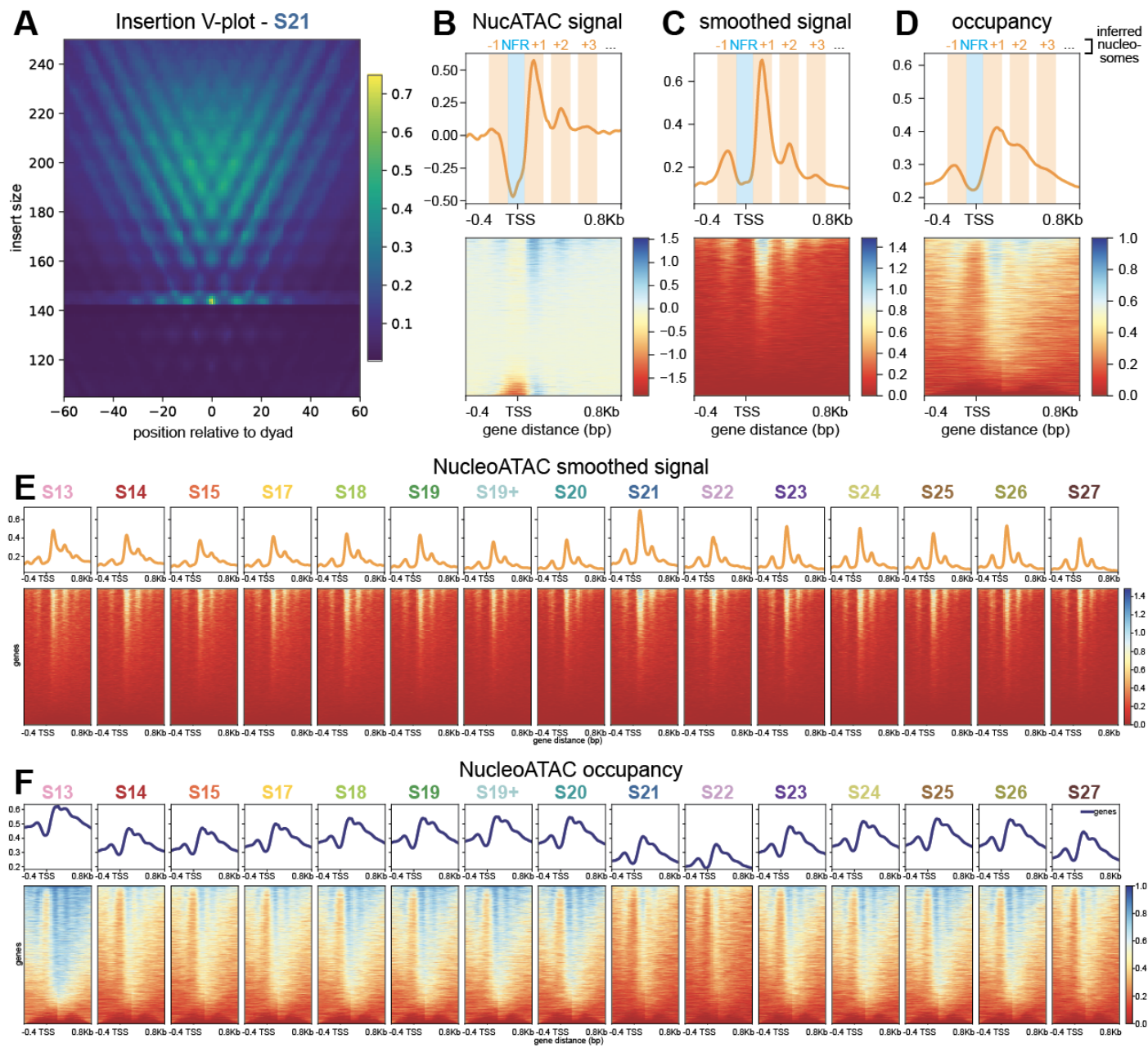
Fig. S8. NucleoATAC quality control and signal over time

A) Insertion V-plot generated from NucleoATAC using reads from S21 libraries. A clear V-shaped pattern is observed, as expected from read data fragmented along a nucleosomal distribution.

B) NucleoATAC raw signal at mRNA starts genome-wide. A strong +1 and -1 nucleosome position signal is observed. A strong negative signal is observed at the expected nucleosome-free region (NFR). Overlaid windows in B-D were drawn manually based on observed signal.

C) NucleoATAC smoothed signal at mRNA starts genome-wide. A strong +1 and -1 nucleosome position signal is observed.

D) NucleoATAC occupancy at mRNA starts genome-wide. A strong +1 and -1 nucleosome position signal is observed.

E) NucleoATAC smoothed signal at mRNA starts genome-wide for all developmental stages.

F) NucleoATAC occupancy signal at mRNA starts genome-wide for all developmental stages. The occupancy value calculated by NucleoATAC varies based on library size; in this plot, values are not normalized based on library size.

Fig. S9. Identification and classification of regulatory element clusters

A) Standardized accessibility plot for top 1000 peaks with strongest membership in each of the 9 clusters. Each library is plotted as a separate point along the line plot.

B) ImpulseDE2 model fits for the top 200 peaks with strongest membership in each of the 9 clusters.

C) Histogram of when all peaks in each cluster achieve their maximum accessibility as calculated by IDE2 max fit.

D) Line plots showing Omni-ATAC accessibility (dashed line, blue axis), NucleoATAC signal (dotted line, black axis), and relative NucleoATAC histone occupancy (solid line, pink axis) for peaks in each cluster across time. Each box contains a line plot summarizing each of the three signals at all peaks within that cluster with respect to a given developmental Omni-ATAC-Seq library. The center of each line plot reflects the average signal at the center of all Omni-ATAC-Seq peaks in that cluster at that developmental stage. Signal is visualized at 0.4kb upstream and downstream of peak centers. Within each signal type, the axes across line plots for each cluster are identical.
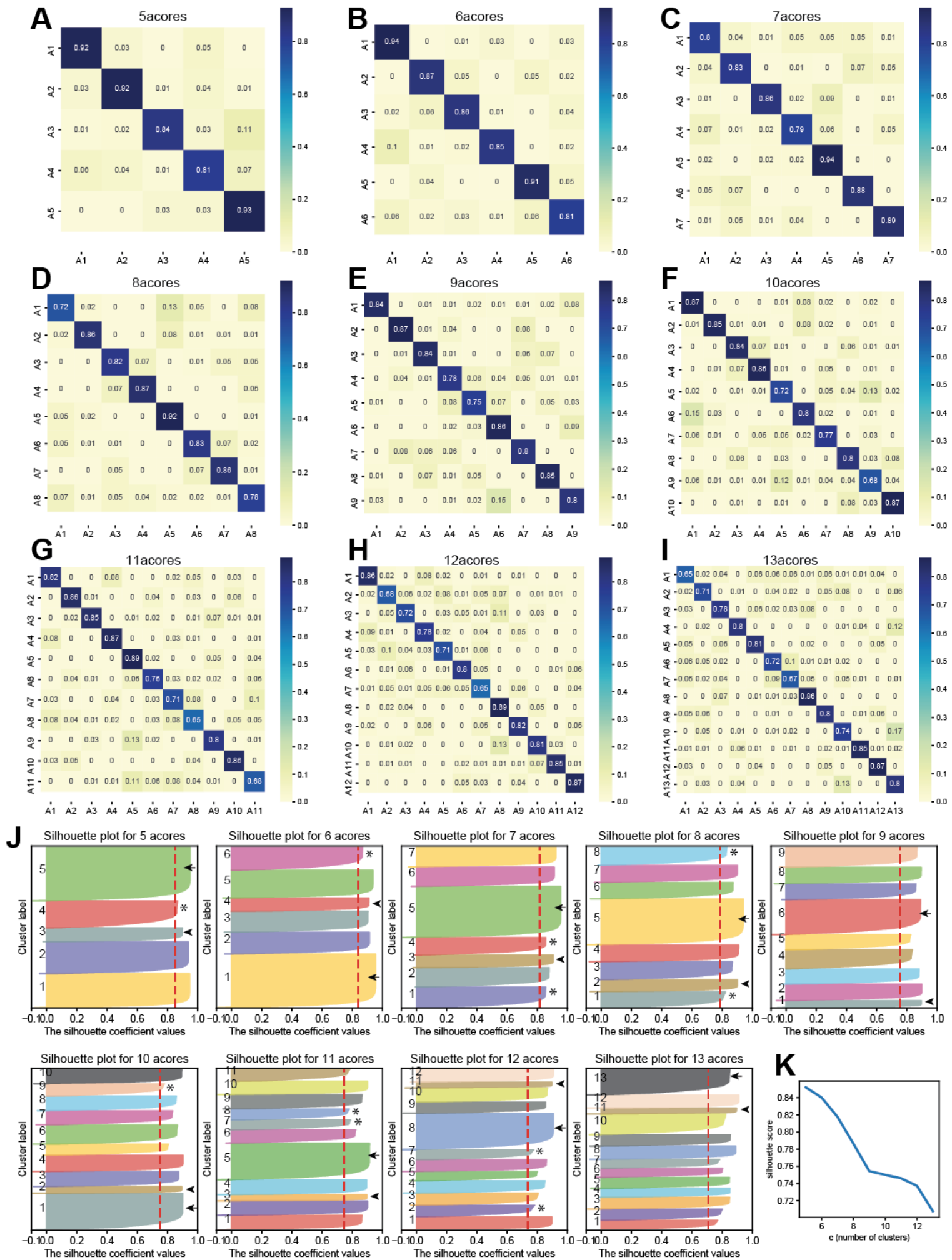
Fig. S10. Mfuzz clustering assessment

Note: cluster labels in this plot are for assessment purposes, and do not correspond to cluster labels used in other figures.

A–I) Cluster overlap matrices generated my Mfuzz indicating percentage overlap between clusters.

J) Silhouette plots for different numbers of clusters. Dashed red line marks the average silhouette score. Asterisks mark clusters with marginal silhouettes (silhouettes at or below the average). An arrowhead marks the thinnest silhouette, while an arrow marks the thickest silhouette within each plot. All cluster numbers other than 9 and 13 showed marginal clusters and large contrast between the thinnest and thickest silhouette widths.

K) Mean silhouette score by number of clusters.
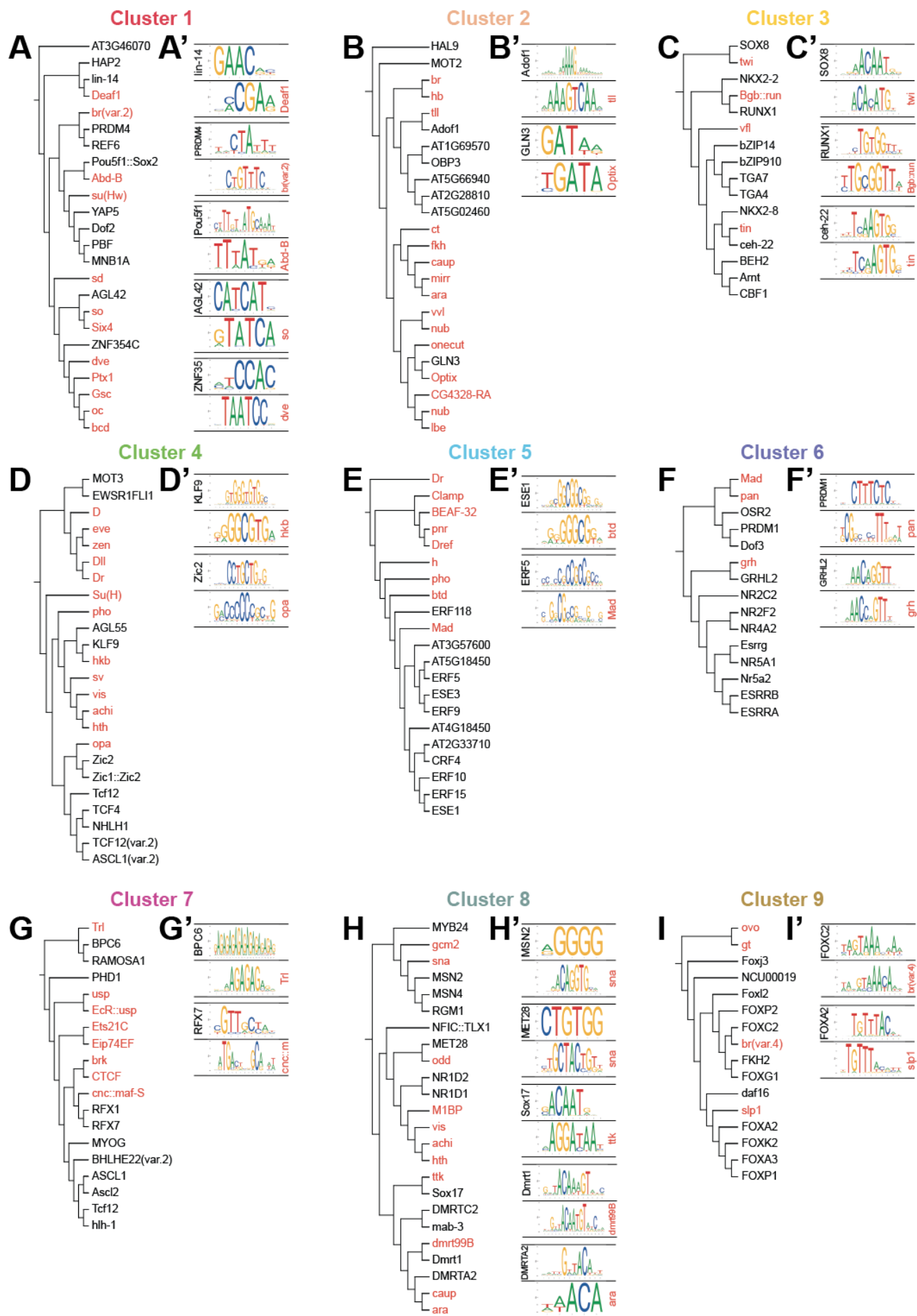
Fig. S11. Alignment of motifs in fuzzy clusters

A–I) Dendrograms generated from sequence alignment of JASPAR motifs using the online STAMP tool visualized using Interactive Tree of Life (iTOL). Top 12 most-enriched transcription factor binding sites from the entire JASPAR CORE database are in black for each dendrogram; up to top 12 most-enriched transcription factor binding sites from the *Drosophila* transcription factors in JASPAR CORE are in red for each dendrogram.

A'–I') Select pairs of transcription factor binding sites with apparently similar sequences based on STAMP alignment. All JASPAR CORE TFs are in black; *Drosophila* TFs are in red.

Fig. S12. Correlation between accessibility and gene expression at individual timepoints

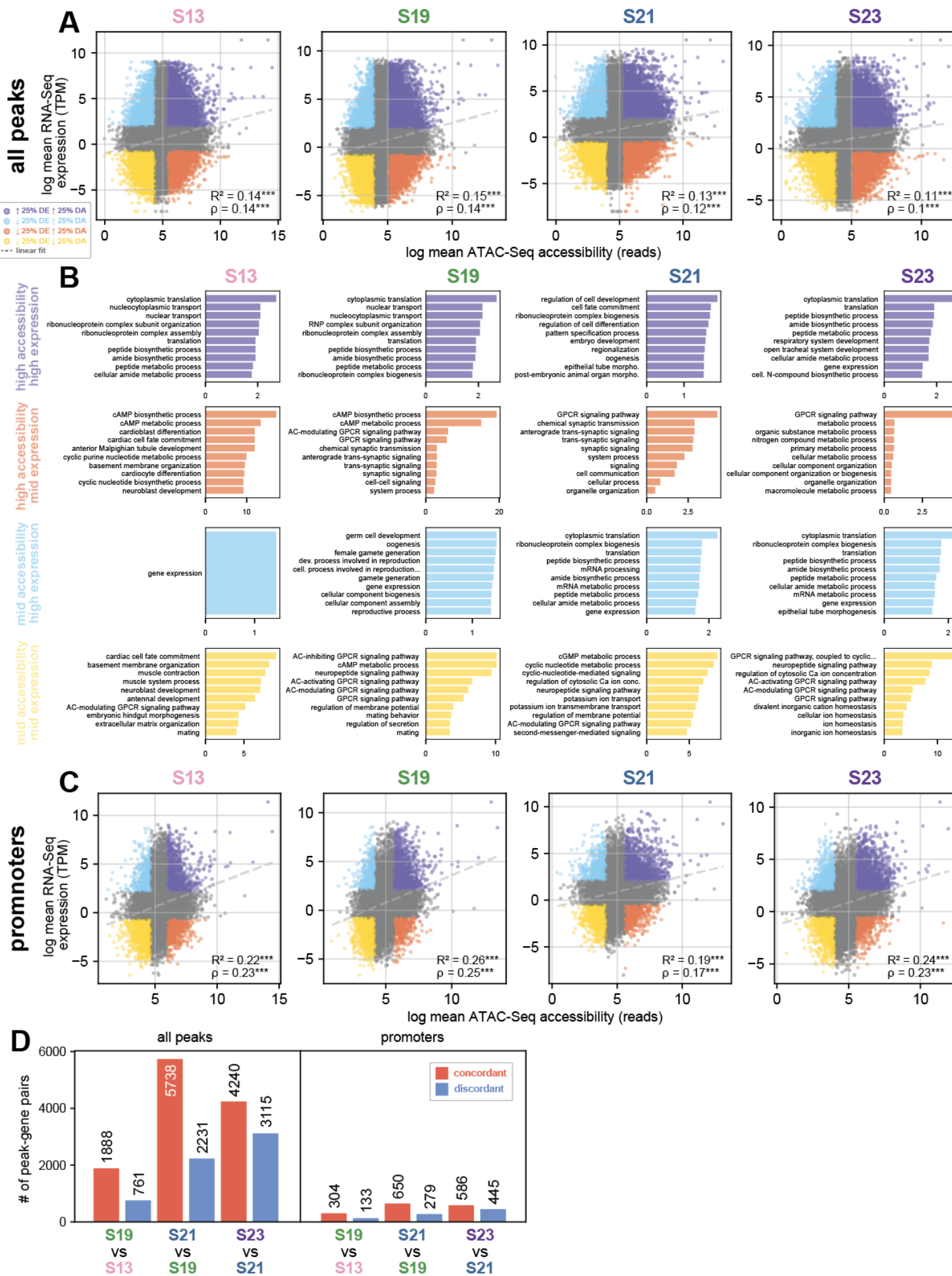A) Correlation between log mean RNA-Seq TPM and log mean Omni-ATAC-Seq accessibility at all peaks in each developmental stage. Peaks colored based on highest and lowest quartiles of expression and accessibility. Dotted line represents a linear fit to all data in the plot. Pearson correlation R^2 and Spearman correlation ρ for all points are displayed in each plot.

B) GO-term enrichment for gene-peak pairs with high accessibility and high expression (purple), high accessibility and medium expression (coral), medium accessibility and high expression (blue), and medium accessibility and medium expression (yellow), for all peaks. The threshold for "high" expression and accessibility were all values above the 75th percentile, while "mid" expression and accessibility were all values below the 25th percentile. Gene lists were extracted based on the nearest Mikado gene to each peak, and orthologous *Drosophila* names were assigned to genes using OrthoFinder.

C) Correlation between log mean RNA-Seq TPM and log mean Omni-ATAC-Seq accessibility at promoter peaks.

D) Quantification of number of significant peak-gene pairs at each time point comparison in Fig. 7. Left plot quantifies number of concordant and discordant peaks across all peak-gene pairs, while right plot quantifies peak-gene pairs that were assigned as promoters.
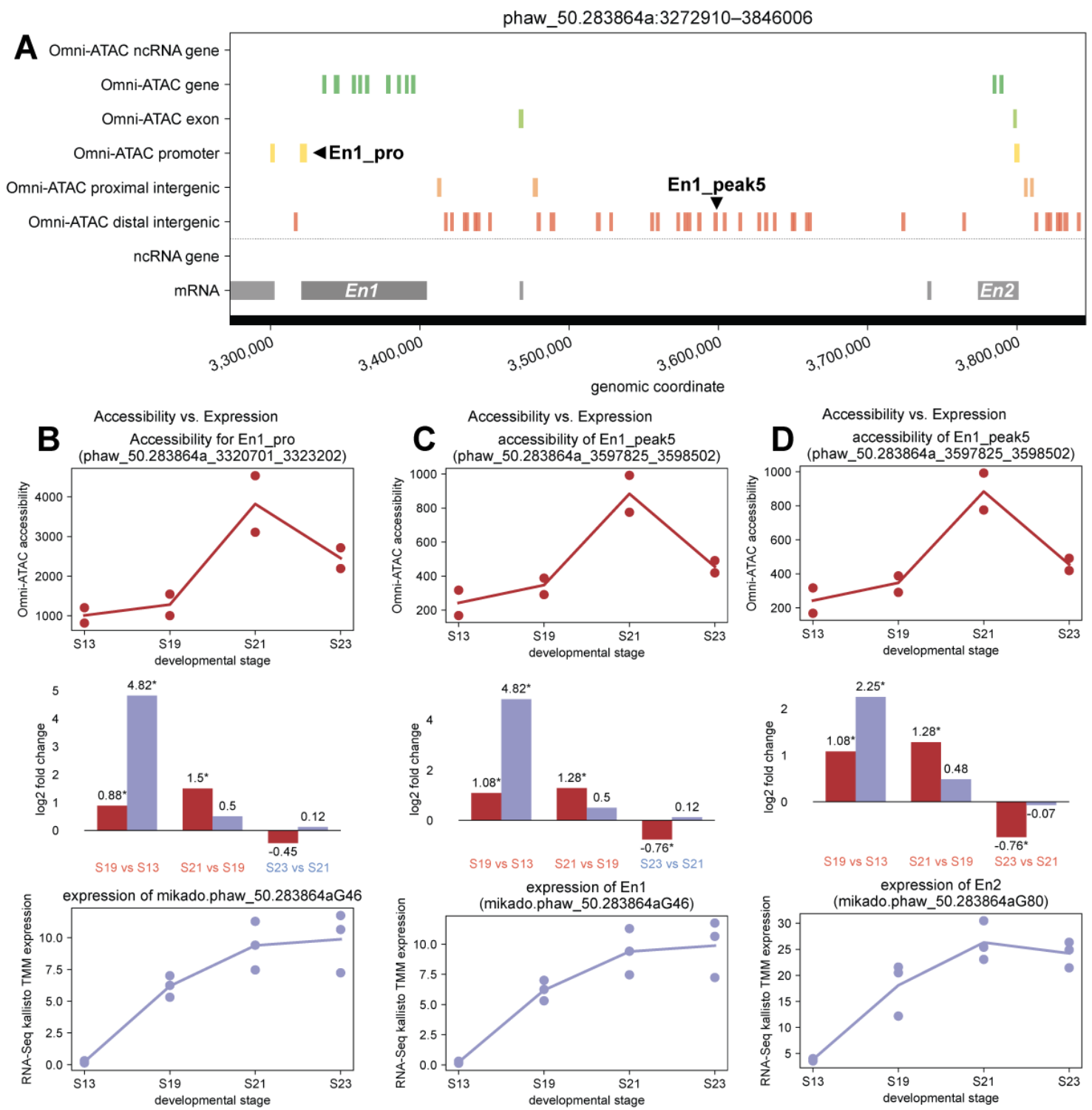
Fig. S13. Visualization of accessibility versus expression for arbitrary peak-gene pairs

A) Visualization of the genomic region around the Parhyale Engrailed-1 and Engrailed-2 loci. Text labels in plot region were added manually after plotting.

B) Visualization of the average accessibility (red line) and raw accessibility values (red dots) for the promoter peak of the *Engrailed-1* gene (top). Visualization of the average gene expression (lavender line) and raw expression values (lavender dots) of the nearest gene to the *Engrailed-1* promoter peak (in this case, *Engrailed-1*, or mikado.phaw_50.283864aG46)(bottom). Visualization of log2 fold change in accessibility (red bars) and gene expression (lavender bars) for each pair of timepoint comparisons (S19 vs. S13, S21 vs. S19, and S23 vs. S21) (middle). Asterisks in bar charts mark log2 fold changes called as statistically significant by DESeq2 (padj < 0.05). Color of comparison labels indicates concordant or discordant relationships between accessibility and gene expression (red for concordant, blue for discordant). When interpreting this visualization, it is important to note that some comparisons may show high log2 fold change that is non-significant; usually, this is when the measured values for a feature are very low. Additionally, statements about concordance and discordance should consider whether both accessibility and expression log2 fold changes are significant. In the case of this example, the comparisons at S19 vs. S13 would be considered truly "concordant" in our dataset, while the comparisons at S21 vs. S19 and S23 vs. S21 would not be considered concordant or discordant, as one or more features are not significantly changing over time.

C) Visualizations as described in (B) for the *Engrailed-1* gene and a distal intergenic peak located between *Engrailed-1* and *Engrailed-2*, nicknamed "En1_peak5", position labeled in (A).

D) Visualizations as described in (B) for the *Engrailed-2* gene and "En1_peak5".

A "PhMS" muscle reporter

B "PhMS" genomic region
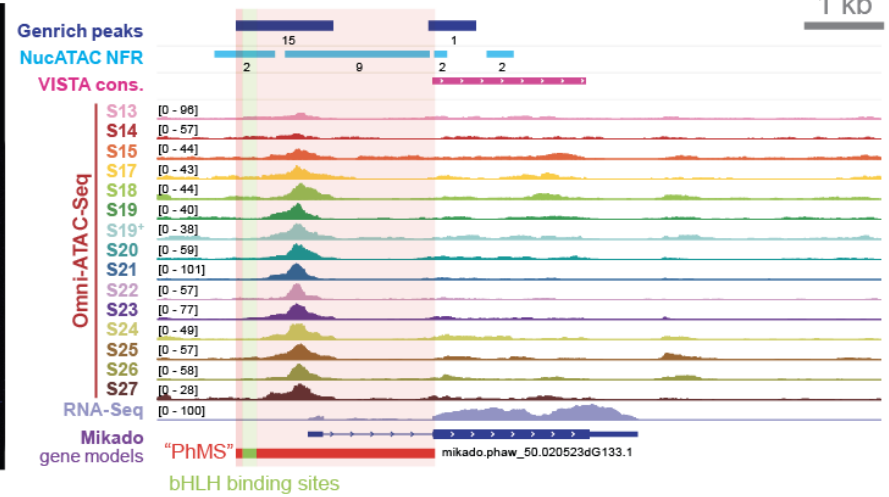
C "HS2a" heat shock element
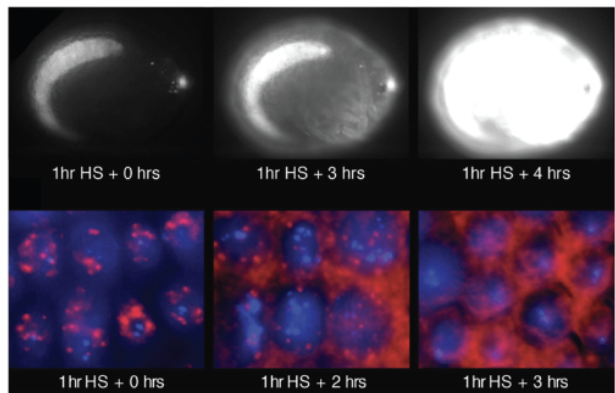
from Pavlopoulos et al. 2009

D "HS2a" genomic region
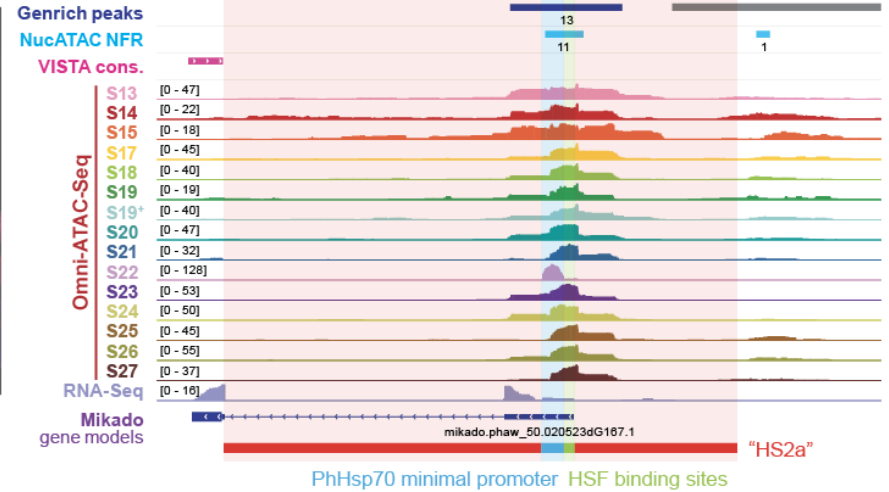
E "PEB"/ EF1a ubiquitous reporter
injected unilaterally; episomal expression
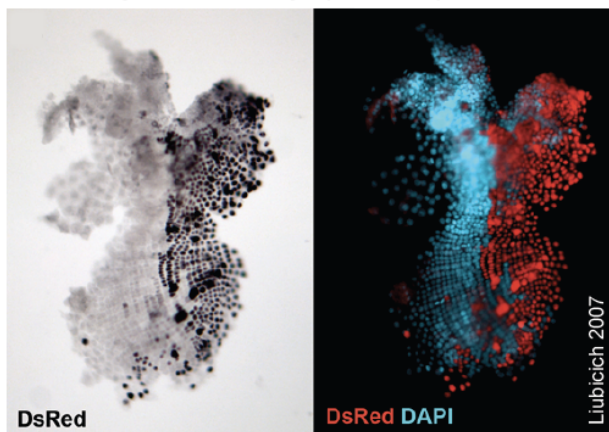
F "PEB" genomic region
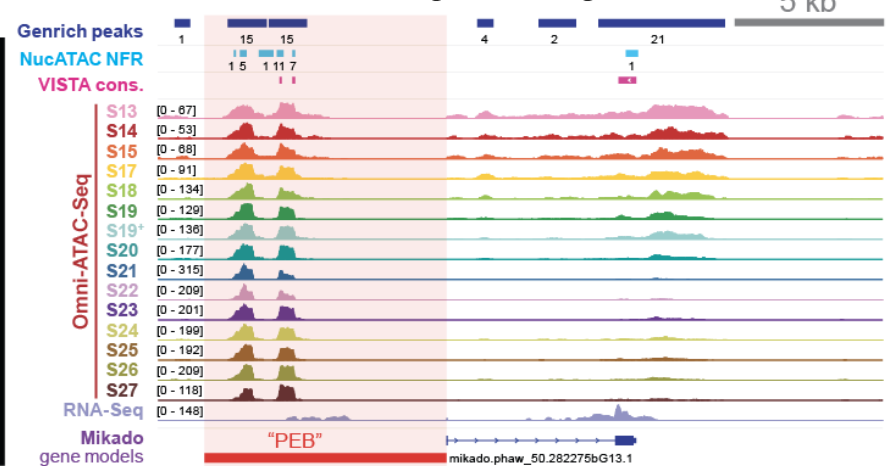
Fig. S14. Omni-ATAC-Seq recovers previously identified regulatory elements

A) Expression of the PhMS reporter visualized using GFP in a juvenile *Parhyale*.
B) The PhMS genomic region. A strong Genrich peak overlaps with a predicted cluster of b-HLH binding sites.
C) Heat shock expression images from a transgenic *Parhyale* embryo carrying the HS2a element.
D) The HS2a genomic region. A peak covers the PhHsp70 minimal promoter and heat shock factor binding sites.
E) Expression of the PEB reporter, visualized in a dissected embryo using a DsRed antibody. Injected embryos showed nearly full expression of this reporter in each germ layer (Liubicich 2007 PhD thesis).
F) The PEB genomic region. Two strong peaks overlap with the 5' end of this reporter.
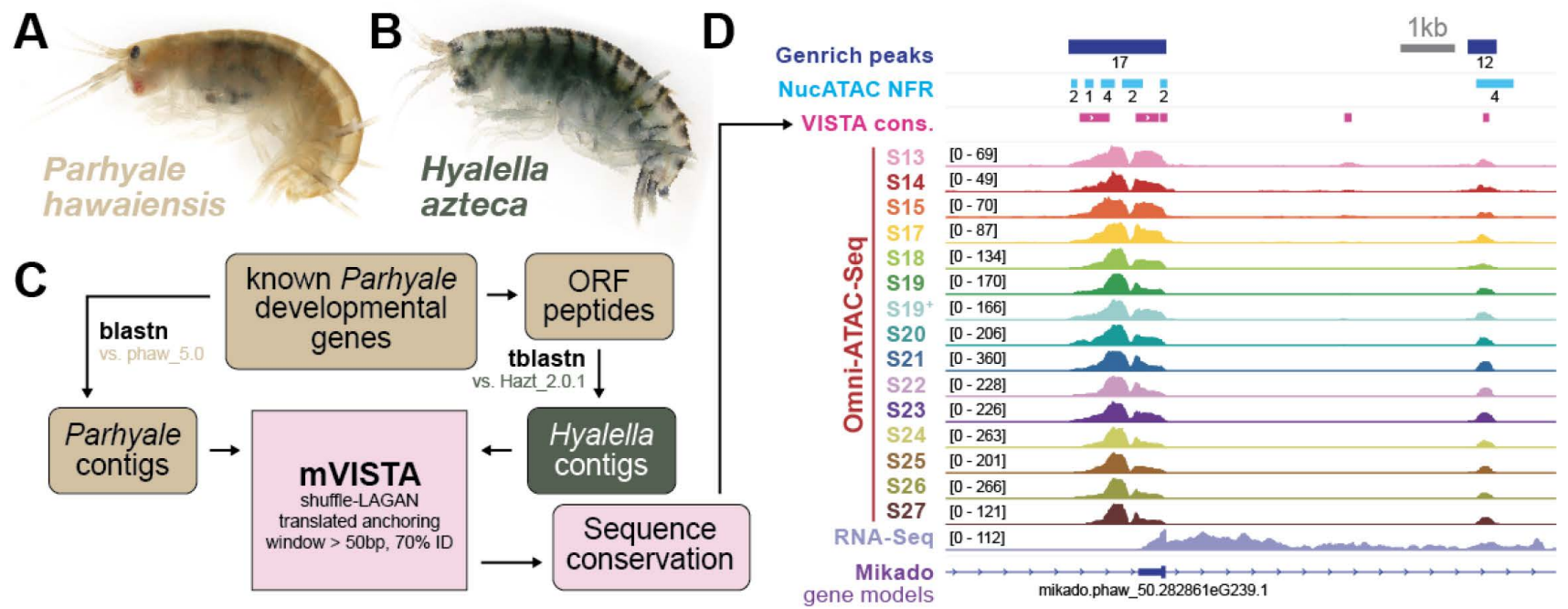
Fig. S15. VISTA homology identification between *Parhyale* and *Hyalella*

A) Adult female *Parhyale hawaiensis*.
B) Adult female *Hyalella azteca*.
C) Pipeline for identification of orthologous genomic regions between *Hyalella* and *Parhyale* for VISTA sequence conservation analysis.
D) Visualization of Genrich peaks, NucleoATAC nucleosome-free regions (NFRs), VISTA conservation, Omni-ATAC-Seq signal, RNA-Seq signal, and Mikado gene models at the putative Sp69 promoter peak.
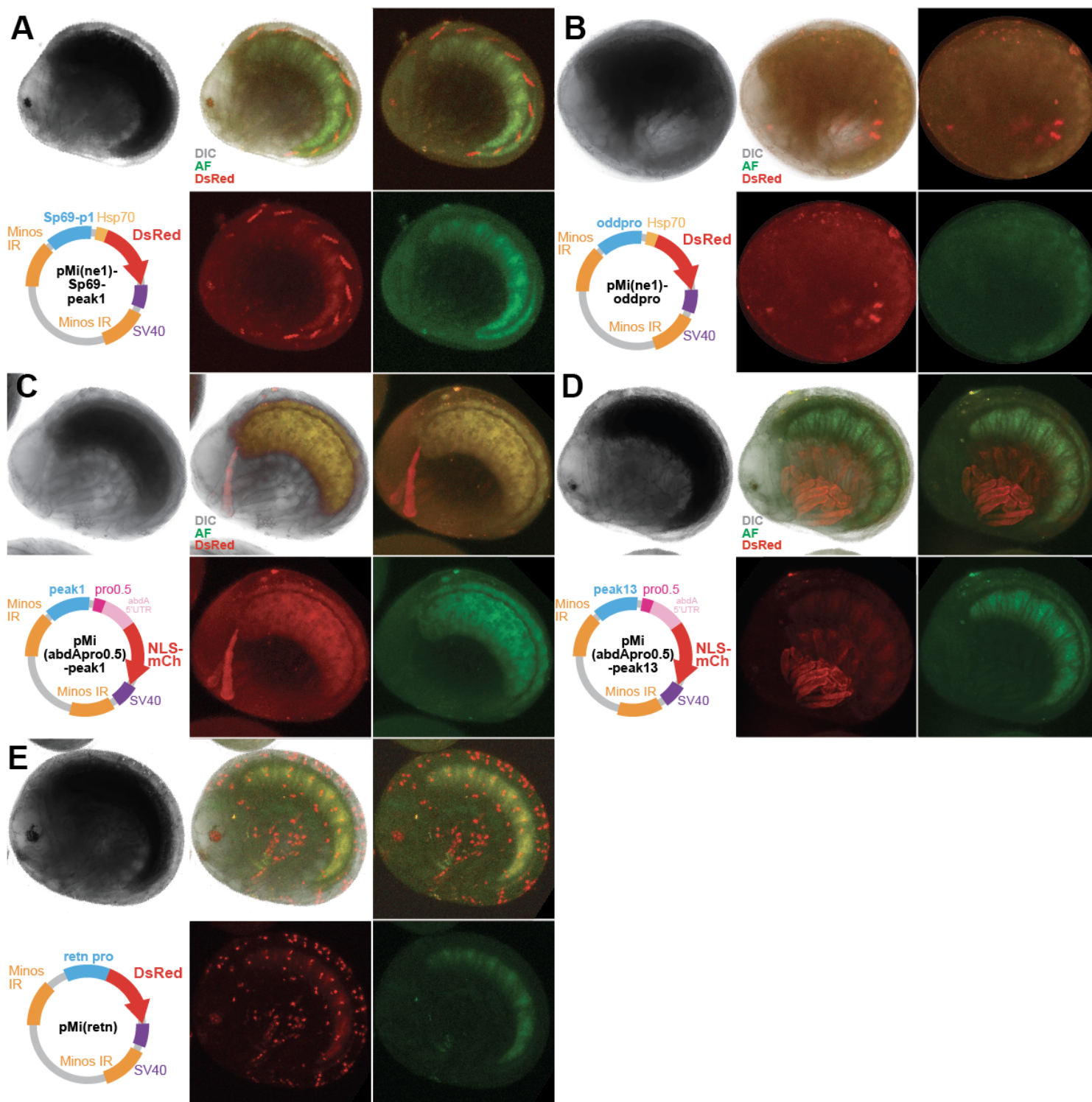
Fig. S16. Spontaneous, non-reproducible expression patterns observed in Minos assays

A) Expression observed in one embryo injected with a plasmid containing the PhHsp70 minimal promoter and the Sp69-p1 peak.
B) Expression observed in one embryo injected with a plasmid containing proximal promoter peak for the odd-skipped gene along with an Hsp80 minimal promoter.
C) Expression observed in one embryo injected with a plasmid containing a putative minimal promoter for the abd-A gene and an upstream peak (abd-A-peak1).
D) Expression observed in one embryo injected with a plasmid containing a putative minimal promoter for the abd-A gene and an upstream peak (abd-A-peak13).
E) Expression observed in one embryo injected with a plasmid containing the putative *Parhyale-retn* promoter.
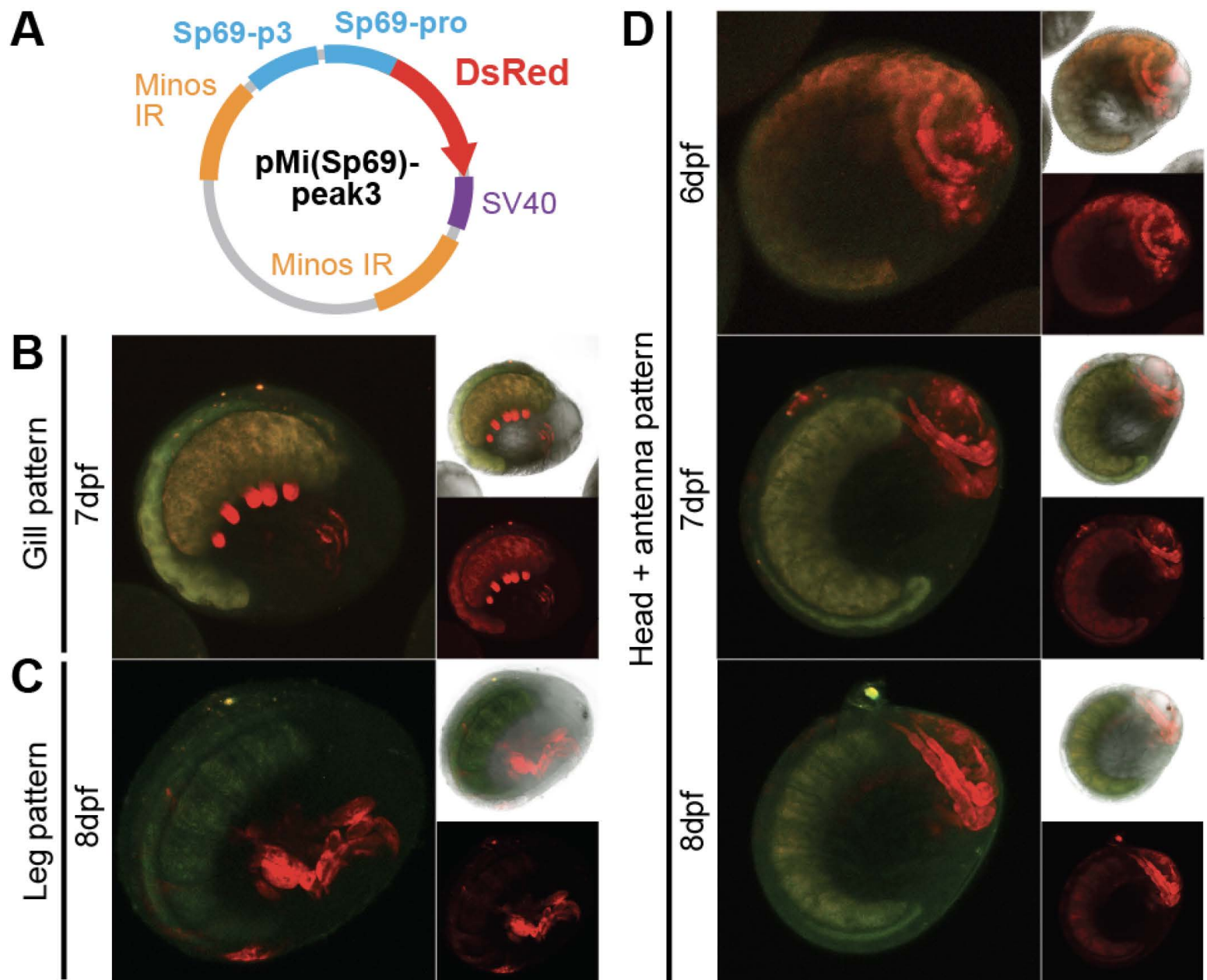
Fig. S17. Expression observed from Sp69-p3 reporter

A) Plasmid map of the pMi(Sp69)-peak3 construct. The construct contains both the putative promoter peak of Sp69, along with a distal peak (peak3).

B) Gill and claw expression observed in one embryo injected with this plasmid.

C) Leg expression observed in one embryo injected with this plasmid.

D) Head and antennal expression patterns observed in 5 embryos injected with this plasmid; embryo with strongest expression pattern shown.

Table S1. Illumina sequencing indices for Omni-ATAC libraries

Sample ID	i7 (4000)	i5 (4000) PE	i5 (4000) SR Flowcell/MiSeq
S13A	TACAGAGC	CTTGGATG	CATCCAAG
S13B	AAGCGTTC	GGAAGGAT	ATCCTTCC
S14A	AGTGACCT	TTCTCTCG	CGAGAGAA
S14B	ACCATAGG	AAGTCCGT	ACGGACTT
S15A	CTCGAACA	ATGGTCCA	TGGACCAT
S15B	TGCGTAAC	TGTCCAGA	TCTGGACA
S17A	AACAGTCC	GATTGCTC	GAGCAATC
S17B	CTAAGACC	GTTGTAGC	GCTACAAC
S18A	TGTTCCGT	GTGAATCC	GGATTCAC
S18B	ACTCAACG	ACTCCATC	GATGGAGT
S19A	GCCTTCTT	TGAGGTGT	ACACCTCA
S19B	TGCTCTAC	TCTTGACG	CGTCAAGA
S19plusA	GTACCACA	GATACTGG	CCAGTATC
S19plusB	GCATTGGT	ATCTTCGG	CCGAAGAT
S20A	CTGTGGTA	ACATTGCG	CGCAATGT
S20B	TTACCGAC	GTTGTTCG	CGAACAAC
S21A	TGACCGTT	CATGGAAC	GTTCCATG
S21B	GCTAAGGA	AACGTGGA	TCCACGTT
S22A	TAGCCATG	CTGCACTT	AAGTGCAG
S22B	ACTTGGCT	GTAGCATC	GATGCTAC
S23A	TGTCACAC	GTGCTTAC	GTAAGCAC
S23B	GACACAGT	CAACCTAG	CTAGGTTG
S24A	GTGGTATG	ATCTCGCT	AGCGAGAT
S24B	ATGCGCTT	GGACCTAT	ATAGGTCC
S25A	AGGAACAC	ACGTTACC	GGTAACGT
S25B	TGGAAGCA	TACTGCGT	ACGCAGTA
S26A	GTCGTTAC	GGTACTAC	GTAGTACC
S26B	GAACCTTC	AGCTTGAG	CTCAAGCT
S27A	TATGACCG	TGTGAAGC	GCTTCACA
S27B	AACGCACA	CGTTATGC	GCATAACG

Table S2. Manual annotation coordinates of 49 selected RACE genes

gene_name	IGV_address	Strand	pro_peak_address	exon1_address	exon-1_address	Trinity fragments?	Fragment
AbdominalBisoform1cloneJ7b11	phaw_50.282695a:24623698-24688129	+	phaw_50.282695a:24622599-24624450	phaw_50.282695a:24,623,624-24,624,479	phaw_50.282695a:24,687,725-24,692,898	Yes	Both
abdomminalAisoform2cloneK313	phaw_50.282695a:25629814-25736948	+	phaw_50.282695a:25628608-25630296	phaw_50.282695a:25,627,088-25,630,363	phaw_50.282695a:25,736,469-25,741,275	Yes	Both
Antennapedia5racecloneL18	phaw_50.282695a:26532519-26613138	+	phaw_50.282695a:26532243-26532820	phaw_50.282695a:26,532,505-26,533,152	phaw_50.282695a:26,611,100-26,613,267	Yes	Both
ash26F	phaw_50.283875a:11478705-11507717	+	phaw_50.283875a:11477924-11479069	phaw_50.283875a:11,478,677-11,480,112	phaw_50.283875a:11,507,570-11,508,560	Yes	Both
betacateninmRNAcompletecds	phaw_50.282639b:6484601-6538431	+	phaw_50.282639b:6451035-6452617	phaw_50.282639b:6,451,925-6,452,777	phaw_50.282639b:6,537,263-6,541,062	Yes	Both
col	phaw_50.283865b:130292-293921	+	phaw_50.283865b:127661-128803	phaw_50.283865b:130,178-130,591	phaw_50.283865b:296,755-299,392	Yes	Both
DeformedcDNAcloneFLb2	phaw_50.282695a:27062438-27152191	+	phaw_50.282695a:27061893-27063246	phaw_50.282695a:27,062,404-27,063,447	phaw_50.282695a:27,151,053-27,152,222	Yes	Exon
deltaproteinmRNAcompletecds	phaw_50.015400b:11871425-11889172	+	phaw_50.015400b:11870888-11871925	phaw_50.015400b:11,871,436-11,871,564	phaw_50.015400b:11,890,317-11,891,420	Yes	Intron
distallessEarlymRNAcompletecds	phaw_50.282654b:30641259-30681459	+	phaw_50.282654b:30640553-30642307	phaw_50.282654b:30,641,130-30,641,834	phaw_50.282654b:30,685,403-30,687,854	Yes	Both
dpp	phaw_50.015400a:5773544-5790898	+	phaw_50.015400a:5772703-5774732	phaw_50.015400a:5,773,547-5,774,013	phaw_50.015400a:5,789,594-5,799,808	Yes	Both
engrailed1	phaw_50.283864a:3322422-3404220	+	phaw_50.283864a:3320702-3323202	phaw_50.283864a:3,321,545-3,323,280	phaw_50.283864a:3,403,484-3,411,112	Yes	Both
engrailed2	phaw_50.283864a:3774988-3800052	-	phaw_50.283864a:3799353-3800567	phaw_50.283864a:3,798,531-3,800,059	phaw_50.283864a:3,757,132-3,758,228	Yes	Intron
eve2	phaw_50.283028b:9254180-9262623	-	phaw_50.283028b:9262111-9262983	phaw_50.283028b:9,262,071-9,262,616	phaw_50.283028b:9,250,289-9,256,607	Yes	Exon
extradenticleproteinexdgene	phaw_50.283468a:8063095-8076889	+	phaw_50.283468a:7799208-7800212	phaw_50.283468a:7,799,472-7,800,056	phaw_50.283468a:8,097,516-8,101,622	Yes	Both
forkheadORF	phaw_50.000135a:2643865-2644971	+	phaw_50.000135a:2649799-2650196	phaw_50.000135a:2,649,559-2,649,662	phaw_50.000135a:2,642,441-2,645,164	Yes	Exon
hes4	phaw_50.282654b:31719607-31734841	+	phaw_50.282654b:31719410-31720007	phaw_50.282654b:31,719,607-31,719,748	phaw_50.282654b:31,732,919-31,736,371	Yes	Both
homothoraxproteinhthgene	phaw_50.283815b:16274653-16574076	+	phaw_50.283815b:16604672-16607172	phaw_50.283815b:16,605,602-16,606,372	phaw_50.283815b:16,168,073-16,170,044	Yes	Both
KNIRPS1kni1mRNAcompletecds	phaw_50.283866:9026088-9035313	+	phaw_50.283866:9025431-9026438	phaw_50.283866:9,026,082-9,026,813	phaw_50.283866:9,034,022-9,037,291	Yes	Both
KNIRPS2kni2mRNAcompletecds	phaw_50.283866:9982901-10018092	-	phaw_50.283866:10018846-10019492	phaw_50.283866:10,017,879-10,018,098	phaw_50.283866:9,982,950-9,984,746	Yes	Exon
Kruppel5raceclone5	phaw_50.004430:12895048-13017154	+	phaw_50.004430:12893620-12894503	phaw_50.004430:12,893,830-12,895,369	phaw_50.004430:13,014,676-13,017,950	Yes	Both
notchproteinmRNAcompletecds	phaw_50.283815c:31626720-31812600	-	phaw_50.283815c:31812314-31813325	phaw_50.283815c:31,812,341-31,812,602	phaw_50.283815c:31,625,065-31,629,214	No	None
odd1	phaw_50.000289b:10469078-10470818	-	phaw_50.000289b:10470131-10471455	phaw_50.000289b:10,470,089-10,470,823	phaw_50.000289b:10,469,064-10,470,057	Yes	Exon
odd3	phaw_50.000289b:9790462-9792439	+	phaw_50.000289b:9789912-9790964	phaw_50.000289b:9,790,460-9,790,970	phaw_50.000289b:9,791,530-9,795,846	Yes	Exon
odd5	phaw_50.000289b:10003623-10007578	+	phaw_50.000289b:10007157-10007887	phaw_50.000289b:10,007,151-10,007,894	phaw_50.000289b:10,001,868-10,005,067	Yes	Both
opa1consensus	phaw_50.000135f:28969563-29076765	+	phaw_50.000135f:28968467-28970375	phaw_50.000135f:28,969,527-28,970,470	phaw_50.000135f:29,072,172-29,076,827	Yes	Both
optixmRNAcompletecds	phaw_50.283817f:8745320-8792652	+	phaw_50.283817f:8744687-8746467	phaw_50.283817f:8,745,412-8,746,075	phaw_50.283817f:8,788,522-8,794,388	Yes	Both
Par6	phaw_50.283823c:3497885-3548486	-	phaw_50.283823c:3548118-3548740	phaw_50.283823c:3,548,186-3,548,556	phaw_50.283823c:3,492,657-3,499,147	Yes	Both
Pax371proteinmRNAcompletecds	phaw_50.000214e:5044320-5078499	+	phaw_50.000214e:5078053-5078983	phaw_50.000214e:5,078,120-5,078,455	phaw_50.000214e:5,042,664-5,047,700	Yes	Both
pdm	phaw_50.283869d:4505837-4506191	+	phaw_50.283869d:4420193-4421142	phaw_50.283869d:4,420,800-4,421,678	phaw_50.283869d:4,506,146-4,509,682	Yes	Both
PhDlllate1	phaw_50.282654b:30431942-30563945	+	phaw_50.282654b:30429539-30431011	phaw_50.282654b:30,430,577-30,432,821	phaw_50.282654b:30,563,597-30,569,176	Yes	Both
PhDlllate2	phaw_50.282654b:30746554-30806154	+	phaw_50.282654b:30745371-30747232	phaw_50.282654b:30,746,377-30,747,054	phaw_50.282654b:30,805,338-30,813,233	Yes	Both
pnt3RACE	phaw_50.282861b:41123311-41125508	-	phaw_50.282861b:41129249-41130109	phaw_50.282861b:41,128,992-41,129,723	phaw_50.282861b:41,112,188-41,123,884	Yes	Exon
proboscipedia5racecloneI192	phaw_50.282695a:27659540-27817538	+	phaw_50.282695a:27657466-27658309	phaw_50.282695a:27,657,840-27,659,994	phaw_50.282695a:27,818,947-27,821,498	Yes	Both
prosperoproteinmRNApartialcds	phaw_50.283875a:6013042-6035299	-	phaw_50.283875a:5998087-5999786	phaw_50.283875a:5,998,945-6,001,387	phaw_50.283875a:6,034,691-6,038,907	Yes	Intron
rho5RACEvariant1	phaw_50.000203b:25395477-25496923	+	phaw_50.000203b:25395456-25395780	phaw_50.000203b:25,395,472-25,395,662	phaw_50.000203b:25,496,406-25,503,322	Yes	Both
rho5RACEvariant2	phaw_50.000203b:25265811-25496923	+	phaw_50.000203b:25265169-25265874	phaw_50.000203b:25,264,762-25,265,904	phaw_50.000203b:25,496,405-25,503,316	Yes	Both
runt2	phaw_50.283028b:1783456-1810535	-	phaw_50.283028b:1809649-1811397	phaw_50.283028b:1,809,475-1,810,690	phaw_50.283028b:1,776,166-1,784,663	Yes	Exon
scallopedproteinsdgene	phaw_50.007301a:763757-786495	-	phaw_50.007301a:920698-921986	phaw_50.007301a:921,178-921,537	phaw_50.007301a:762,031-762,973	Yes	Both
scratchscrtmRNAcompletecds	phaw_50.283869f:2146785-2165201	-	phaw_50.283869f:2270072-2271735	phaw_50.283869f:2,270,060-2,271,025	phaw_50.283869f:2,146,761-2,150,484	Yes	Both
Sexcombsreduced5racecloneE17c1	phaw_50.282695a:26769068-26862063	+	phaw_50.282695a:26768486-26770008	phaw_50.282695a:26,861,434-26,862,081	phaw_50.282695a:26,768,947-26,770,207	Yes	Both
shortgastrulationproteinmRNAcompletecds	phaw_50.015400b:7523582-7867524	+	phaw_50.015400b:7522832-7524249	phaw_50.015400b:7,523,573-7,525,453	phaw_50.015400b:7,866,925-7,867,874	Yes	Both
slp1	phaw_50.283811:7718478-7720485	+	phaw_50.283811:7718255-7718898	phaw_50.283811:7,718,433-7,720,788	phaw_50.283811:7,725,774-7,731,670	Yes	Both
snail2proteinmRNApartialcds	phaw_50.283864:718097-721779	+	phaw_50.283864:713312-713863	phaw_50.283864a:713,632-713,751	phaw_50.283864b:717,219-721,810	Yes	Both
snail3proteinmRNApartialcds	phaw_50.283864b:30526-33704	-	phaw_50.283864b:25948-26939	phaw_50.283864b:26,453-26,990	phaw_50.283864b:29,154-33,680	Yes	Both
Sp69protein	phaw_50.282861e:14546239-14663576	+	phaw_50.282861e:14544573-14546375	phaw_50.282861e:14,545,862-14,546,358	phaw_50.282861e:14,667,910-14,675,477	Yes	Both
spi5RACElargefragment3RACE	phaw_50.000081a:2326730-2344027	+	phaw_50.000081a:2243466-2244527	phaw_50.000081a:2,243,861-2,244,179	phaw_50.000081a:2,347,070-2,349,001	Yes	Both
SuH	phaw_50.283826:2280509-2340463	+	phaw_50.283826:2416373-2417483	phaw_50.283826:2,416,066-2,416,895	phaw_50.283826:2,270,773-2,276,731	Yes	Exon
unc4NPhsequence	phaw_50.000135e:7542170-7621463	+	phaw_50.000135e:7541550-7542771	phaw_50.000135e:7,542,144-7,542,554	phaw_50.000135e:7,644,646-7,649,912	Yes	Exon
wnty	phaw_50.283866:2773623-3090108	-	phaw_50.283866:3089459-3090541	phaw_50.283866:3,089,558-3,090,096	phaw_50.283866:2773623-3090108	Yes	Both

Table S3. Software version numbers

Software name	Version	Usage
bamtools	2.5.1	Bam file manipulations
bedtools	2.28.0 or 2.30.0	Bed file manipulations
Bowtie2	2.3.0 or 2.3.4.1	Read alignment
BUSCO	3.0.2	Transcriptome and genome completion evaluation
cutadapt	2.4	Removing sequencing adapters from reads
deeptools	3.3.1	Visualization of data genome-wide
DESeq2 (R)	1.34.0	Differential accessibility/ expression analyses in pairwise comparisons
eggNOG-mapper	2.0.5	Automated gene function assignment
FASTQC	0.11.7	Library quality assessment
Genrich	0.6	Omni-ATAC peak calling
GMAP	2020-11-20	Aligning transcriptomes to genome for Mikado analyses
HISAT2	2.1.0	Aligning RNA_Seq reads to genome
igvtools	2.3.98	File format conversions for viewing in IGV
ImpulseDE2 (R)	0.99.10	Performing IDE2 analyses
JASPAR2020 (R)	0.99.10	Extracting PWMs from JASPAR
kallisto	0.43.1	Read abundance estimation for RNA-Seq
Mfuzz	2.54.0	Fuzzy clustering of peak accessibility
Mikado	2.3.0	Mikado transcriptome merging
minimap2	2.18-r1052-dirty	Aligning long-read Nanopore sequences to *Parhyale* genome
NucleoATAC	0.2.1	Prediction of nucleosome positions using Omni-ATAC data
OrthoFinder	2.5.4	Orthology assignment between *Parhyale* and *Drosophila*
PicardTools	2.9.0	Deduplication of reads
Portcullis	1.1.2	Identifying valid splice junctions from RNA-Seq read data
Python	3.7 or 3.8.8	Interfacing with Jupyter notebooks and running Python scripts on a SLURM scheduler
R	4.1.1 or 4.0.3 or 3.8	Performing IDE2, DESeq2 analyses; converting JASPAR files to RGT-HINT format
RGT-HINT	0.13.1	Inference of transcription factor footprints
samtools	1.8	Quality filtering (q=10) and sorting of bowtie2-mapped reads
StringTie	2.1.7	Assembly of StringTie2 L and StringTie2 SL transcriptomes
Transdecoder	5.0.2	Generating peptide sequences for OrthoFinder, Mikado, and other analyses
trim_galore	0.4.4	Trimming reads used in all analyses
Trinity	r2013_08_14	Assembling Trinity-limb (old) transcriptome
Trinity	2.5.1	Assembling Trinity-all, Trinity-gg, Trinity-S13, Trinity-S19, Trinity-S21, Trinity-S23, Trinity-limb (new) transcriptomes
universalmotif (R)	1.12.1	Converting JASPAR PWMs to RGT-HINT compatible format
VISTA	web VISTA	Identifying sequence homology between *Parhyale* and *Hyalella*

Supplementary Materials and Methods

S21-S22 sample outlier properties

For several metrics, the data we generated for samples at stages S21 and S22 appeared to show outlier effects, such as strong separation in PC2 of PCA analysis (Fig. 2C), a lower percentage of mapped reads (Supp. Fig. 2.1B), and lower inter-sample Spearman correlation (Supp. Fig. 2.1I). These data could be a reflection of variability in library quality, or could represent genuine biological differences with these samples. The shape of the PCA curve in Fig. 2C is consistent with a previously-described "horseshoe effect", "Guttman effect", or "arch effect" that is observed in microbial ecology datasets where a single strong gradient drives separation between samples in the dataset (Legendre and Legendre, 2012; Morton James T. et al.; Podani and Miklós, 2002). In the case of our dataset, we expect to find a single strong gradient – developmental time – which should separate out our samples.

To assess the degree to which PC2 in our data is driven by separation of S21/S22 samples, we removed the S21/S22 samples from our dataset and performed PCA. We observed that the relative positions of the remaining libraries remains largely similar (Supp. Fig. 2.2A, B), with time points closer to the middle of development showing a positive loading in PC2, and time points closer to the beginning and end of development showing a negative loading. This suggests that the second major axis of variation in our dataset would remain a contrast between middle vs. terminal timepoints, even without the S21/S22 data, and that the relative positioning of the remaining data within PC-space is not entirely driven by S21/S22.

To further assess the degree of the S21/S22 samples' outlying effects, we also performed ImpulseDE2 analysis to generate model fits without S21/S22 data. Doing so allowed us to determine to what degree the S21/S22 stages are necessary for driving the accessibility trajectory of individual peaks, and of the data more broadly. We performed IDE2 with either all data, or the S21/S22 data removed prior to input into IDE2. This generated two sets of model fits to the "cloud" of accessibility vs. time measurements: one that included the S21/S22 data, and one without. We evaluated, for each peak in our dataset, the time point at which the IDE2 model achieved maximum accessibility (the "IDE2 max fit"), and plotted both the "all" and "noS21S22" data as a histogram (Supp. Fig. 2.2C). The presence of peaks that achieve predicted maximum accessibility in the S21/S22 stages in the "no S21/S22" data is a result of how we calculate "max fit", which does not require that there is a known accessibility value at a given timepoint; only that the time point during which the model fit is maximum is closest to the timing of that developmental stage. Overall, we still observed early, middle, and late enrichment of IDE2 max fit even when the S21/S22 data are removed. We do see a rightward shift in the middle timepoint histogram in the direction of later stages, although this may be expected given the absence of concrete accessibility values at S21/S22 in the "no S21/S22" data. This indicates that our data globally retain the general trends of early, middle, and late enrichment of accessibility in the absence of the S21/S22 data. Moreover, this suggests that, even without the S21/S22 data, the remaining data from early and late stages result in a model fit that still predicts maximum accessibility at middle developmental stages for many peaks.

To further measure the influence of the S21/S22 data in IDE2 model fit, we also evaluated the degree of change in the general accessibility of a peak when the S21/S22 stages were removed. This analysis aimed to assess whether removing S21/S22 data resulted in an IDE2 model with the same general trajectory as with all data, as opposed to the more stringent requirement of evaluating whether the exact developmental stage of the peak was changed. To perform this analysis, we grouped developmental stages into five quintiles, each representing three stages of development. We asked, for each peak in our dataset, whether that peak's IDE2 max fit was "stable" when the S21/S22 data were removed; that is, if the quintile of the IDE2 max fit was altered when the S21/S22 data were removed (i.e. if a peak moved more than 3 developmental stages away from its original position), a peak was considered "unstable" (Supp. Fig. 2.2D, E). We observed that over 80% of peaks in each quintile remained "stable" after removing the S21/S22 data, suggesting that the vast majority peaks show the same general trajectory of accessibility even without the S21/S22 data. Peaks within the middle time points appeared to be more unstable than peaks at the terminal timepoints, which could be expected given that the S21/S22 timepoints constituted the middle-most timepoints in our dataset.

It is difficult to determine whether a change in accessibility trajectory for a given peak caused by the removal of S21/S22 data is indicative of technical differences in sample preparation, such as batch effects; biological variation, such as a potentially unknown mutant or sick embryo; or due to genuine wildtype biological processes that occur at the

S21/S22 stages. These caveats acknowledged, a comparative analysis of the data in the absence of the S21/S22 stages suggests that much of the global picture of development remains the same.

To facilitate future users of this dataset, we have included the model parameters calculated from IDE2 using both the full dataset and the data with S21/S22 removed, as well as a Jupyter notebook (ParhyaleATACExplorer.ipynb) that allows users to plot the raw accessibility data and IDE2 model fits with respect to time, so that downstream experiments can consider the potential differences with the S21/S22 samples.

Trinity transcriptome assembly parameters

Trinity *de novo* transcriptome (all stages) was assembled with --max_memory 1500G and --CPU 32. Illumina NovaSeq short reads were supplied as a sample file, where each sample had two read files for read1 and read2. Read files were generated by first trimming transcripts using trim_galore, and then combining the val_1 reads with unpaired_1, and val_2 reads with unpaired_2 for each stage using cat. The same merged read files were used for stage-specific assemblies. For the genome-guided assembly, *in silico* normalized reads generated by Trinity for read1 and read2 across all stages were aligned to the genome using hisat2 --dta and then sorted using samtools. Trinity was run as Trinity --genome_guided_bam <bam file> --genome_guided_max_intron 300000.

StringTie2 transcriptome assembly parameters

Nanopore reads, when generated, are partitioned into chunks in different folders in the output. Reads were compiled by barcode. All reads were combined together into a single file and mapped to the genome using minimap2 -ax splice using the phaw_5.0.fa genome file as a target. StringTie2 L assembly was run using stringtie -L <nanopore_reads.bam>, while StringTie2 SL assembly was run using stringtie -mix <illumina_reads.bam> <nanopore_reads.bam>. Both StringTie2 assemblies used default settings.

Mikado setup parameters

Trinity transcriptomes were aligned to the phaw_5.0.fa genome using GMAP with --max-intronlength-ends=300000 --format=gff3_gene. All HISAT2 aligned read BAM files from Illumina RNA-Seq were fed into Portcullis using portcullis full --force --copy --verbose -t 32. Due to yet-to-be-resolved errors in the Portcullis package, we manually converted some assignments of the portcullis output using the following commands:

```
$ cat portcullis_out/2-junc/portcullis_all.junctions.tab |awk '!(($11=="?" && $14=="NA") || ($11=="?" &&
    $13=="NA"))' > portcullis_out/2-junc/portcullis_all.junctions.cleaned.tab
$ portcullis filter --verbose portcullis_out/1-prep portcullis_out/2-junc/portcullis_all.junctions.cleaned.tab
```

The scoring parameters and other variables for the Mikado configure file are found in the "20210630_list.txt" file. We applied penalties to the Trinity-based transcriptomes and bonuses to the StringTie2 transcriptomes based on our observations that Trinity transcripts, when aligned to the genome, showed many apparently spurious short transcripts at the RACE genes we examined.

For mikado configure, we used --mode permissive --scoring mammalian.yaml --copy-scoring mammalian.yaml -bt uniprot_sprot.fasta. We reasoned that the large genome size of *Parhyale* would be appropriate to analyze using the mammalian settings as opposed to the other available settings.

For the Transdecoder step, we ran the analysis with default settings. We omitted BLAST from our Mikado pipeline, as the extremely large file size of our mikado_prepared.fasta file proved intractable to straightforward BLASTX analysis. We estimated that it would take around 5 months to run using our current computing approaches. An updated Mikado transcriptome could be generated by breaking up the mikado_prepared.fasta file into many smaller BLASTX analyses, which could run much faster. The required files for further Mikado analyses are available in the GEO accession data.

eggNOG annotation

We generated a .pep file of the final Mikado transcriptome or the MAKER genome annotation using Transdecoder, and used emapper.py with default settings and -m diamond.

OrthoFinder annotation

Using a .pep file generated from the final Mikado transcriptome or the MAKER genome annotation using Transdecoder, we ran OrthoFinder to compare each transcript source to the database of all *Drosophila melanogaster* proteins found in the UNIPROT database.

DESeq2 parameters

For differential expression and accessibility analyses, we performed pairwise DESeq2 runs for adjacent pairs of developmental stages (e.g. S19 vs. S13, S21 vs. S19). Analysis was performed using standard settings. Peaks or genes were considered differentially expressed when padj < 0.05.

ImpulseDE2 analysis

We performed ImpulseDE2 analysis using standard settings, using the matrix of reads generated from bedtools multicov and used for DESeq2 analysis. For the Time parameter, we used the number of hours for each developmental stage, as indicated in the Browne et al. *Parhyale* staging guide (Browne et al., 2005). The time points were: 72, 77, 80, 87, 90, 96, 104, 112, 120, 132, 144, 155, 168, 180, and 192 hours, representing stages 'S13', 'S14', 'S15', 'S17', 'S18', 'S19', 'S19+', 'S20', 'S21', 'S22', 'S23', 'S24', 'S25', 'S26', 'S27', respectively. We extracted the IDE2 model fits for each peak (available at the GEO accession) as a .tsv file, used a custom Python function to plot accessibility trajectories (will be available at the GitHub repository). To generate such plots, we recreated the accessibility plotting function from the ImpulseDE2 package in R as two Python functions, shown below:

```
def ImpulseFunction(t, vecImpulseParam):
    return (1/vecImpulseParam[3]) * (vecImpulseParam[2] + (vecImpulseParam[3] - vecImpulseParam[2]) * (1/(1 +
        np.exp(-vecImpulseParam[1] * (t - vecImpulseParam[5]))))) * (vecImpulseParam[4] + (vecImpulseParam[3] -
        vecImpulseParam[4]) * (1/(1 + np.exp(vecImpulseParam[1] * (t - vecImpulseParam[6])))))


def ImpulseEvaluate(vecTimepoints, vecImpulseParam):
    vecImpulseValue = [ImpulseFunction(i, vecImpulseParam) for i in vecTimepoints]
    vecImpulseOutput = [i if i > 10^(-10) else 10^(-10) for i in vecImpulseValue]
    return vecImpulseOutput
```

JASPAR2020 database

For RGT-HINT analyses, the software expects each motif as a simple matrix representing position weights at each base for A, T, G, and C. This format is not readily available for the JASPAR2020 database, which utilizes the JASPAR format. We downloaded the full JASPAR2020 CORE non-redundant database, split the database into individual JASPAR format files for each transcription factor, and then converted that JASPAR format tile into a plain matrix file using the "universalmotif" R package. We also utilized the metadata extracted from this package to generate an .mtf file as required by RGT-HINT. All files for this type of analysis will be available on GEO at time of publication.

Description of Observed Differences in Clusters

We first assessed the spatial category distribution of each cluster compared to the global average (Figure 5C). Clusters 1, 2, and 8 appeared to be enriched for "unknown" regions, while clusters 3-7 appeared to show slightly greater proportions of intronic elements than distal intergenic elements, counter to the global average. Cluster 8 is notable in that it is both the smallest cluster, and the cluster where a majority of peaks are found in "unknown" regions. In addition, we used the IDE2 model fits for each cluster to evaluate the general accessibility dynamics of each cluster (Supp. Fig. 5.1B, C). Clusters 1 and 2 appear to be enriched for peaks that achieve high accessibility early in development and decrease in accessibility over time. Clusters 3–6 appear to be enriched for peaks that achieve maximum accessibility at different timepoints along developmental time. Cluster 7 appears to be enriched for peaks that show low accessibility early in development, and increased accessibility late in development. Finally, clusters 8 and 9 showed a more striking pattern, with Cluster 8 composed of peaks that appear to have a pulsatile increase in accessibility around stages S21–S22, and Cluster 9 composed of peaks with an opposite pattern: those with a sudden pulsatile decrease around S21–S22. Thus, the 9 clusters identified in our analysis appeared to capture distinct patterns of accessibility throughout time.

We further investigated the clusters identified by Mfuzz using additional metrics in order to understand the biological significance of these clusters. First, we investigated the relationship between accessibility dynamics and nucleosome positioning between clusters (Fig. 5E, Supp. Fig. 5.1D). For each developmental time point, we visualized Omni-ATAC accessibility (solid line, black axis) compared to NucleoATAC signal (dashed line, blue axis) for peaks belonging to each cluster. We observed that peaks belonging to Clusters 3–7 and 9 all appeared to have an inverse relationship between accessibility and NucleoATAC signal; this inverse relationship was most obvious at the time points during which these peaks achieved their maximum accessibility (Fig. 5E, blue dashed boxes). Moreover, these peaks all appeared to have a depletion of nucleosome occupancy at peak centers relative to the region around peaks (Supp. Fig. 5.1D, solid line, pink axis). By contrast, peaks in clusters 1, 2, and 8 appeared to show a counterintuitive and opposite result: both high NucleoATAC signal and high nucleosome occupancy at timepoints during which those peaks achieved maximum accessibility (Fig. 5E, Supp. Fig. 5.1D, solid line, pink axis). As with the clusters that showed an inverse relationship between accessibility and occupancy, the clusters that showed a parallel relationship between accessibility and occupancy showed the strongest relationship at those timepoints during which they achieved maximum accessibility (Fig. 5E, red dashed boxes).

To assess the biological function of each of the 9 Mfuzz clusters, we performed GO term enrichment analysis. To assign functions to each peak, we extracted the *Drosophila melanogaster* gene assigned to the nearest gene for each peak, and performed GO enrichment analysis on the list of unique gene names associated with peaks from each cluster relative to the list of all OrthoFinder-assigned gene names in the genome. Fig. 6A illustrates the top 10 most-enriched GO terms for each of the clusters. While we were not able to identify significantly enriched GO-terms (FDR < 0.05) for Clusters 1, 2, and 8, Clusters 3–7 and 9 all showed enrichment of developmental GO-terms. These enriched GO-terms generally appeared to match the expected biological functions of peaks based on their accessibility dynamics. For example, Cluster 3 peaks, which generally achieve maximum accessibility during germ band elongation stages in development, showed enrichment for GO terms including head segmentation and neuroblast fate, two processes which take place during this developmental period. Meanwhile, Cluster 7 peaks appeared to be enriched for neuronal terms (neuropeptide signaling pathway, synaptic target recognition, neuron recognition), potentially reflecting the function of genes near these peaks which achieve their maximum accessibility towards the very end of the developmental time-course, during which much of the morphology of the organism has been specified, and the embryo begins to twitch inside the egg. Notably, Clusters 3–7 showed slightly greater enrichment of intronic peaks than distal intergenic peaks, suggesting that intronic peaks may be more likely to regulate developmental processes. These results resemble the observation from recent work showing that tissue-specific enhancers are enriched in intronic regions in various human cell types (Borsari et al., 2021).

To further investigate the clusters in our dataset, we examined the enrichment of TF footprints generated from HINT-ATAC in each of our clusters. For each cluster, we compiled all of the TF footprints predicted across developmental stages for all peaks within that cluster. We compared the enrichment of those TF footprints to our randomly generated background data, and examined all JASPAR CORE transcription factors, as well as the subset of JASPAR CORE transcription factors found in *Drosophila*.

We observed strong differential enrichment of different groups of transcription factors in each of our 9 clusters. Notably, particular families of transcription factors appeared enriched in several of the clusters. For example, Cluster 1 and Cluster 2 appeared to have enrichment for footprints matching C2H2 zinc-finger transcription factors and homeodomain transcription factors. Clusters 4 and 7 showed enrichment for footprints matching beta helix-loop-helix (b-HLH) transcription factors. Meanwhile, clusters 5, 8, and 9 showed strong enrichment for footprints of ethylene response factor (ERF), DM-type intertwined zinc-finger factor (DM IZF), and forkhead domain (FKH) transcription factor families, respectively.

Spontaneous Expression Patterns

In the course of our experiments, we observed numerous spontaneous expression patterns (Supp. Fig. 9). These expression patterns may represent rare integration events, or may reflect fortuitous insertions into regions of the genome with promiscuous regulatory sequences that facilitate expression.

One expression pattern observed from the Sp69-p3 enhancer appeared to be repeated in 5 embryos along the ~300 we injected. This expression pattern was characterized by strong expression in the head, antenna 1, and antenna 2 (Supp. Fig. 10). In the same set of experiments, we also observed three spontaneous expression patterns that appeared to differ amongst each other, labeling different regions of the limbs and gills. It is possible that the repeated expression observed from the Sp69-p3 enhancer represents a genuine expression pattern. We tested this hypothesis by raising three Sp69-p3 head-expressing animals to adulthood and mating the two adults that showed the strongest expression during development. However, none of the offspring of this cross showed obvious head expression, or any obvious DsRed expression elsewhere in the embryo.
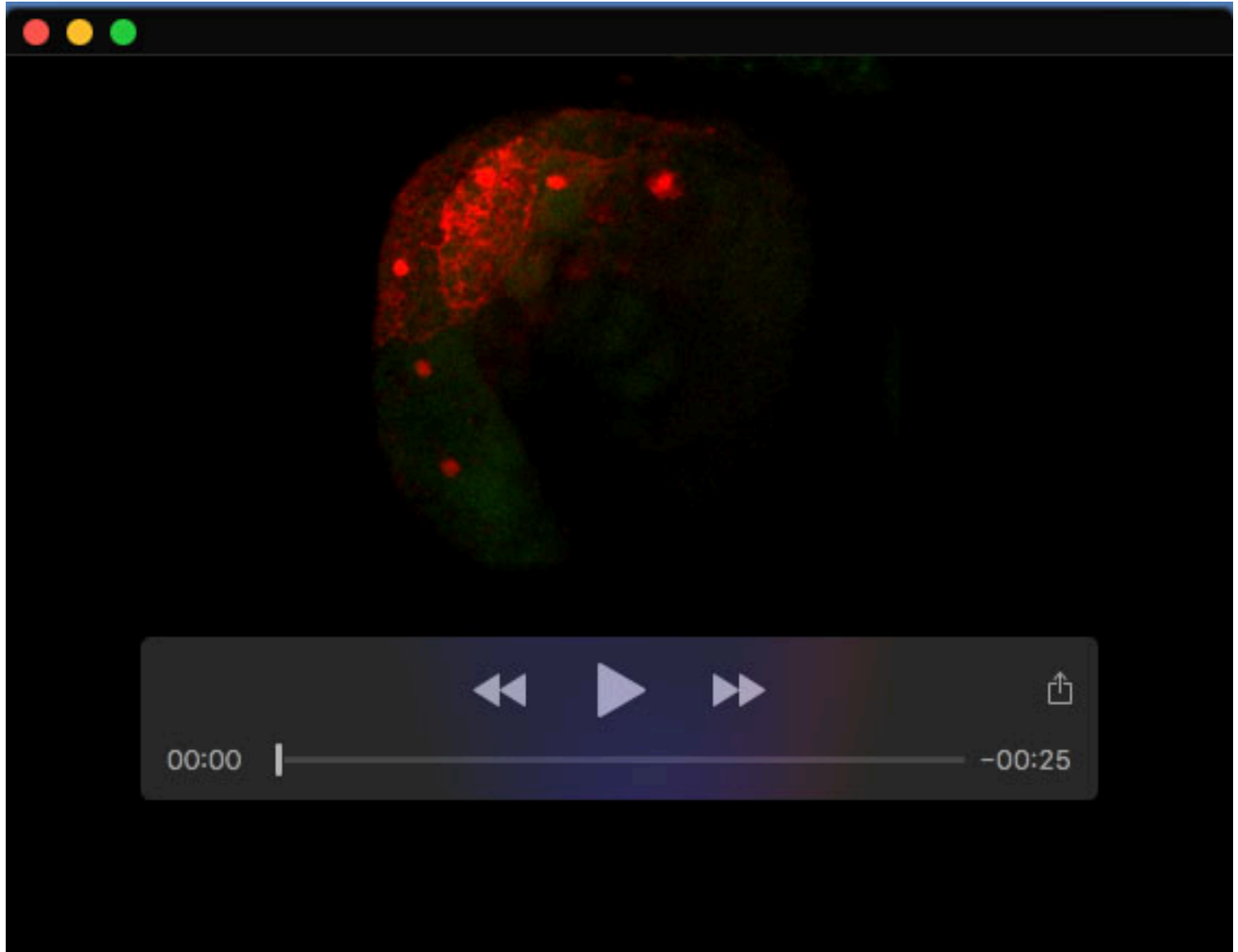
Movie 1. Yolk cells labeled by the Hsp70-p2 reporter Max intensity projection; frames taken 5 minutes apart.