SUPPLEMENTAL MATERIALS

Common and specific gene regulatory programs in zebrafish caudal fin regeneration at single-cell resolution

Yujie Chen,[1,2,3,†] Yiran Hou,[1,2,3,4,†] Qinglin Zeng,[1,2,3] Irene Wang,[1,2] Meiru Shang,[1,2] Kwangdeok Shin,[5] Christopher Hemauer,[1,2,6] Xiaoyun Xing,[1,2,3] Junsu Kang,[5] Guoyan Zhao[1,7,8,*] & Ting Wang[1,2,3,*]

[1] Department of Genetics, Washington University School of Medicine, St Louis, MO 63110, USA.

[2] The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63110, USA.

[3] McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO, USA.

[4] Present address: Department of Medical Microbiology and Immunology, School of Medicine and Public Health, University of Wisconsin, Madison, WI 53706, USA.

[5] Department of Cell and Regenerative Biology, School of Medicine and Public Health, University of Wisconsin Madison, WI 53705, USA.

[6] Present address: Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA.

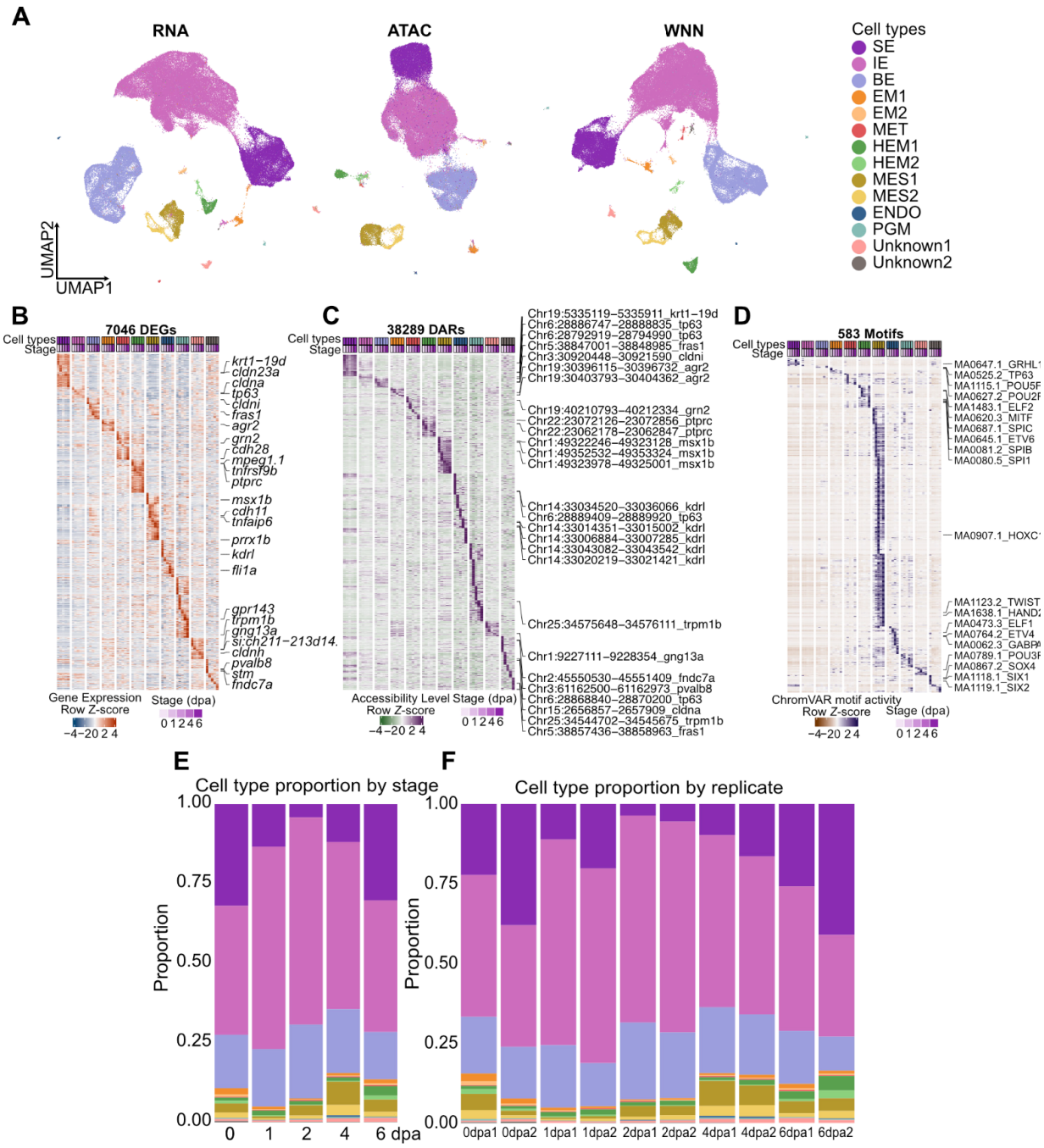
[7] Department of Neurology, Washington University School of Medicine, St Louis, MO 63110, USA.

[8] Department of Pathology and Immunology, Washington University School of Medicine, St Louis, MO 63110, USA.

[†] These authors contributed equally to this work.

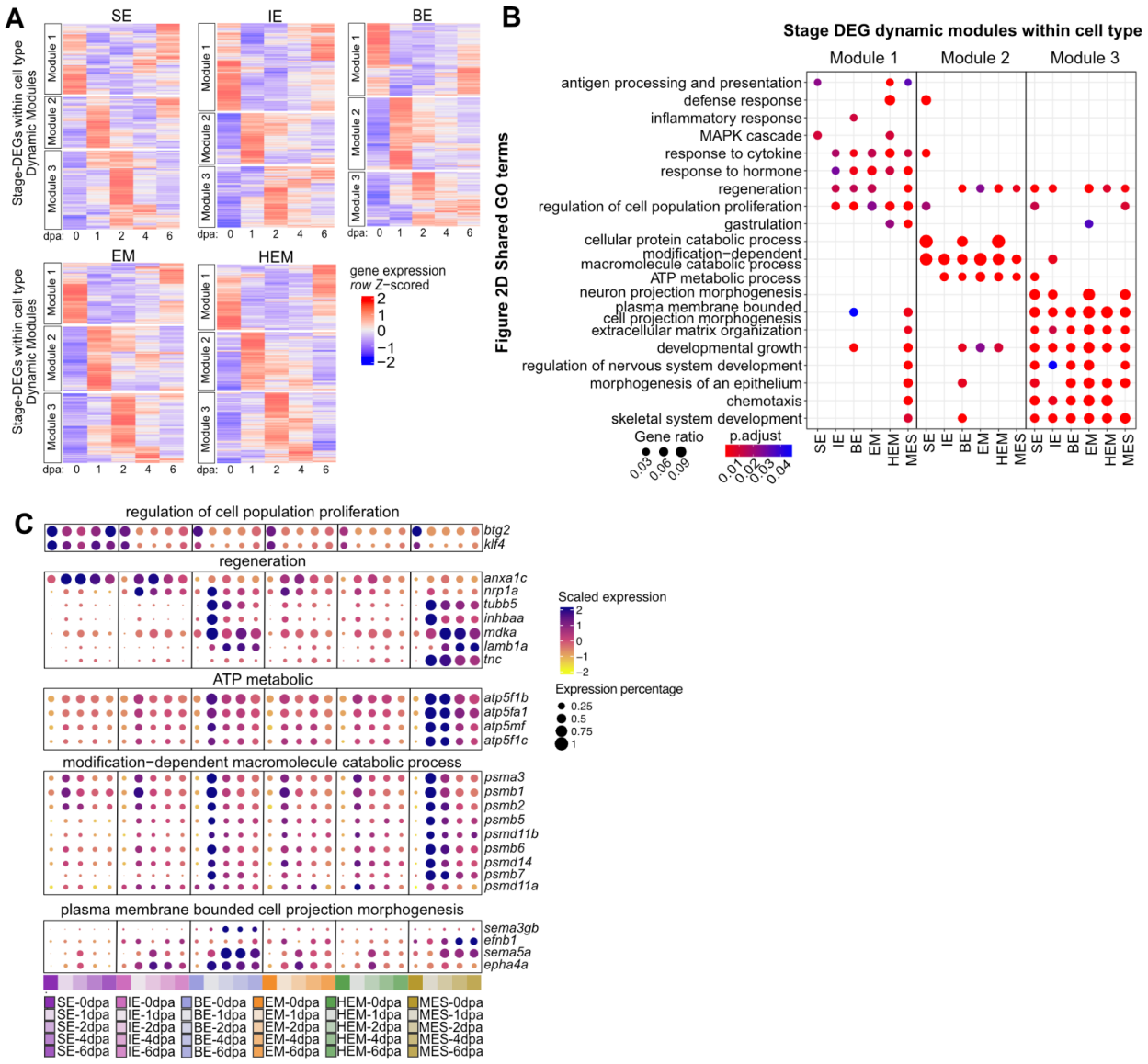
[*] Correspondence: twang@wustl.edu; gzhao@wustl.edu

Table of Contents

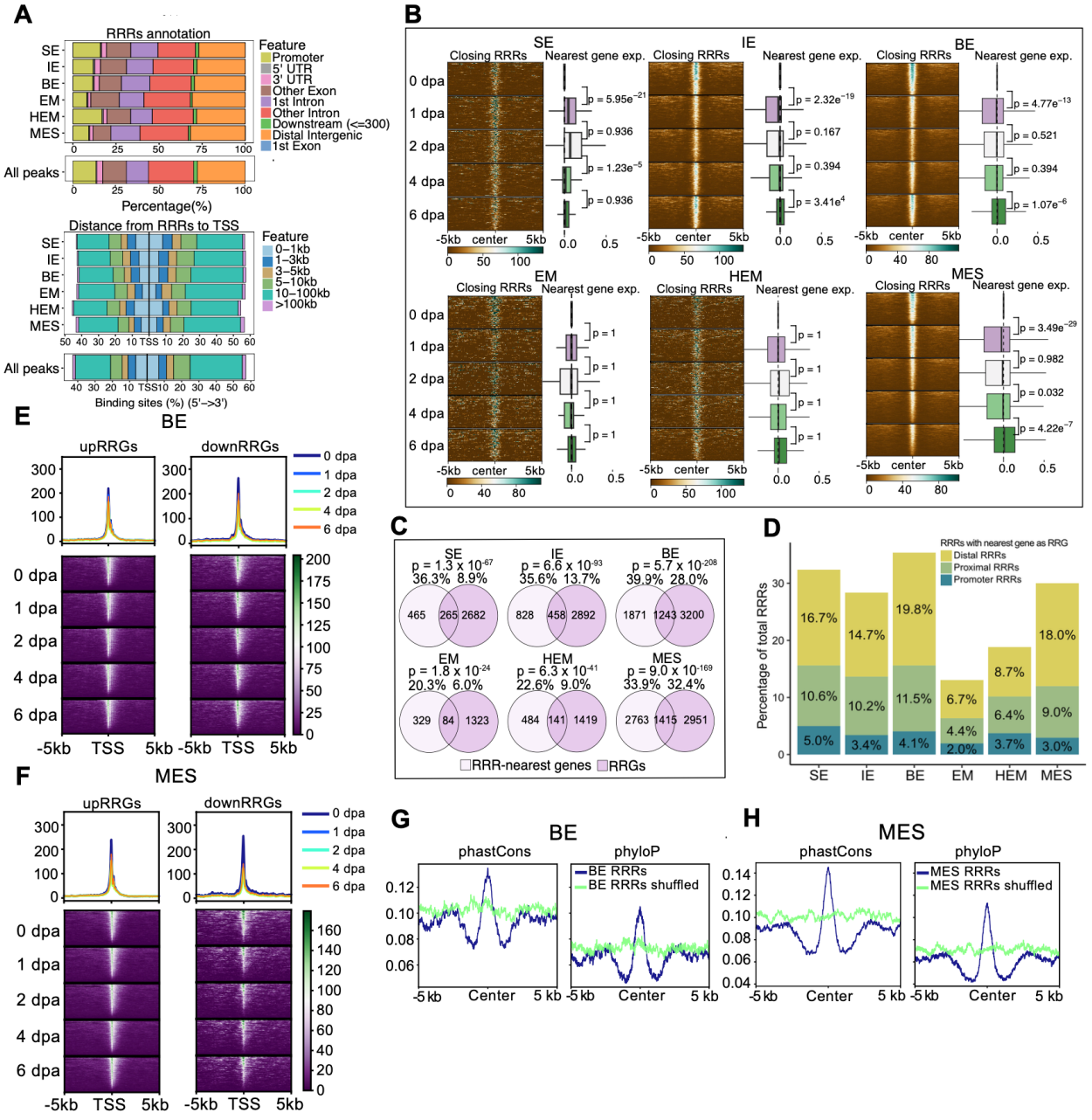


Supplemental Figure S1. Cell type annotations and differential expression analysis across modalities. (*A*) UMAP visualizations of all cells that passed quality control, represented through RNA embedding (left), ATAC embedding (middle), and weighted nearest neighbor (WNN) embedding (right). Each color encodes a distinct cell type. (*B*) A gene expression heatmap depicts DEGs associated with each cell type. The heatmap is organized with genes along the rows and cell type signals by stage across the columns. Selected cell type marker genes for each cell type are highlighted on the right. (*C*) A heatmap of chromatin accessibility signals displays the DARs for each cell type. Peaks are organized in rows, while the average signals for each cell type by stage are displayed across the columns. Key DARs located near marker genes are indicated on the right. (*D*) This motif activity heatmap shows the motifs with

differential activity across cell types, with motifs arranged in rows and cell type by stage signals in columns. Significant motifs, the TF of which are specific to certain cell types, are annotated on the right side of the heatmap. (*E*) Cell type proportions across each stage. Color code is the same as in (*A*). (*F*) Cell type proportions across each sample. Color code is the same as in (*A*).

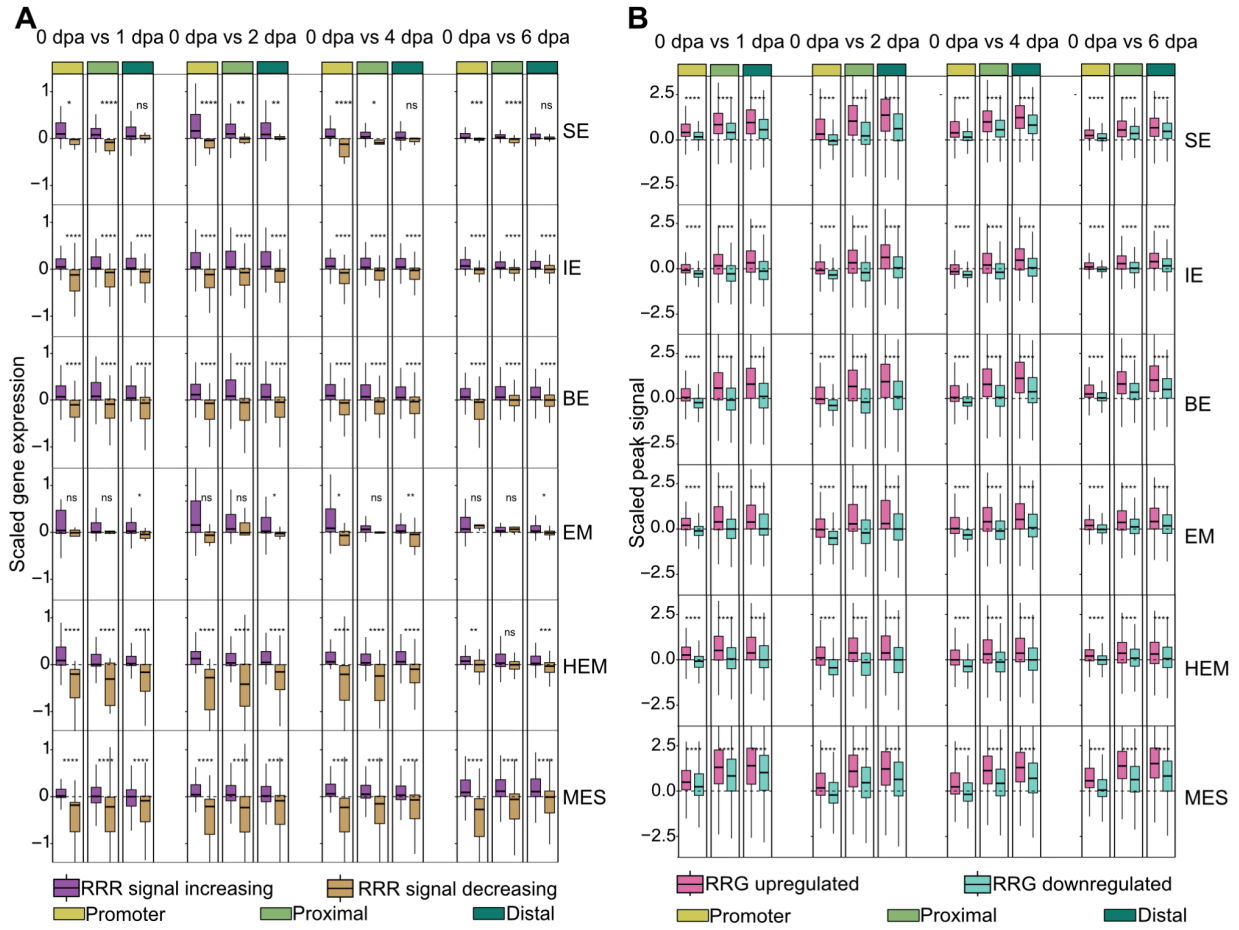


Supplemental Figure S2. Dynamic three modules across cell types and enriched GO terms. (*A*) Heatmaps showing scaled expression of all stage DEGs identified within each cell type (MES heatmap is shown in Figure 2E), with genes clustered into three modules. (*B*) GO terms enriched in each dynamic module. GO terms enrichment analyses were done across all dynamic modules and major cell types. GO terms shown in Figure 2D were extracted and displayed here, facilitating a better comparison to show dynamics and transition of shared processes across stages and cell types. (*C*) Dot plots represent scaled expression of genes selected from GO terms (*B*) labeled on the top of each group. Cell types and stages are encoded by a gradient color set at bottom.

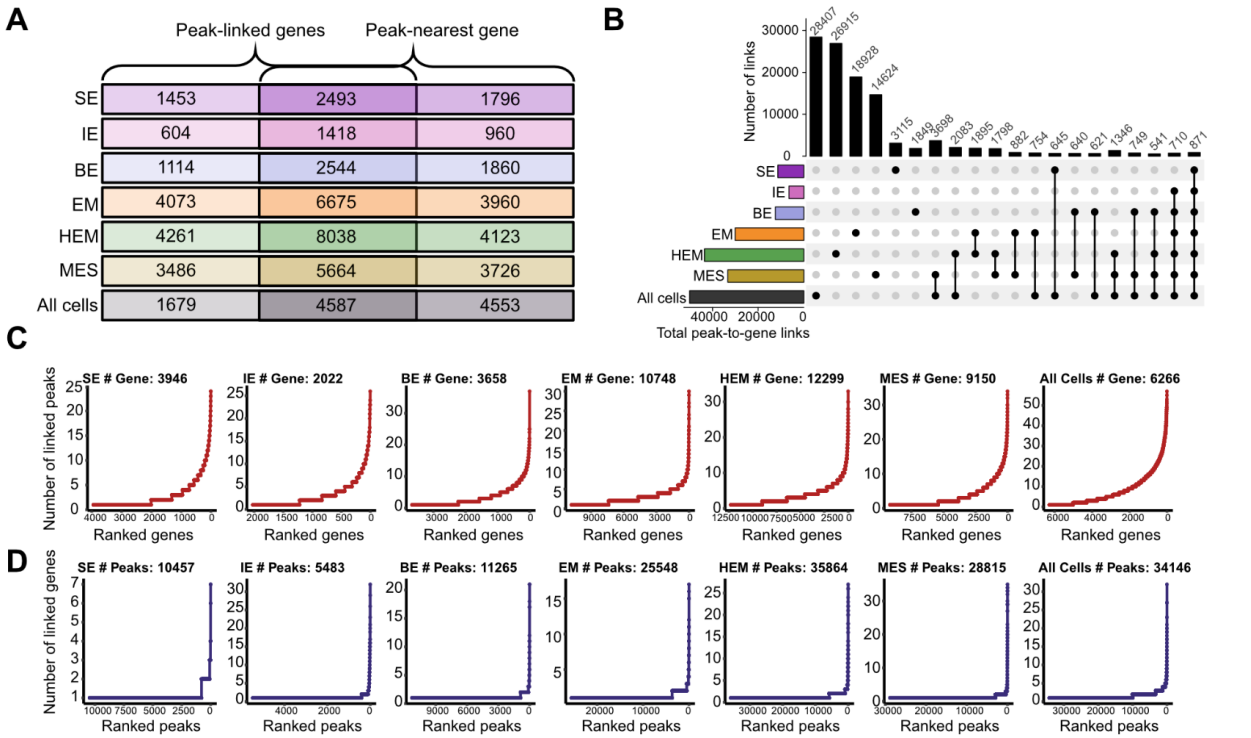


Supplemental Figure S3. Comparative analysis of RRRs and RRGs. (*A*) Annotation of RRRs identified within each cell type, as well as for a union set of peaks called within each cell type. (*B*) Aggregated snATAC-seq signal profiles by time points, plotted over 5-kb regions flanking the center of closing RRRs for each cell type. Box box plots on the right of each heatmap represent the log2 fold change of expression levels of genes nearest to RRRs across time point, measured across different time points compared to 0 dpa. The p-values were calculated using Wilcoxon signed-rank test for comparison between each two time points. (*C*) Venn diagrams illustrate the intersection between the nearest genes to RRRs and RRGs within each cell type. P-values calculated using hypergeometric test, with the total detected gene count as the reference, provide a measure of the significance of these overlaps. Percentages indicate the proportion of overlapping genes relative to all genes nearest to RRRs (left) and all RRGs (right). (*D*) Stacked bar plots display the percentage of total RRRs associated with the nearest genes classified as RRGs across major cell types. The

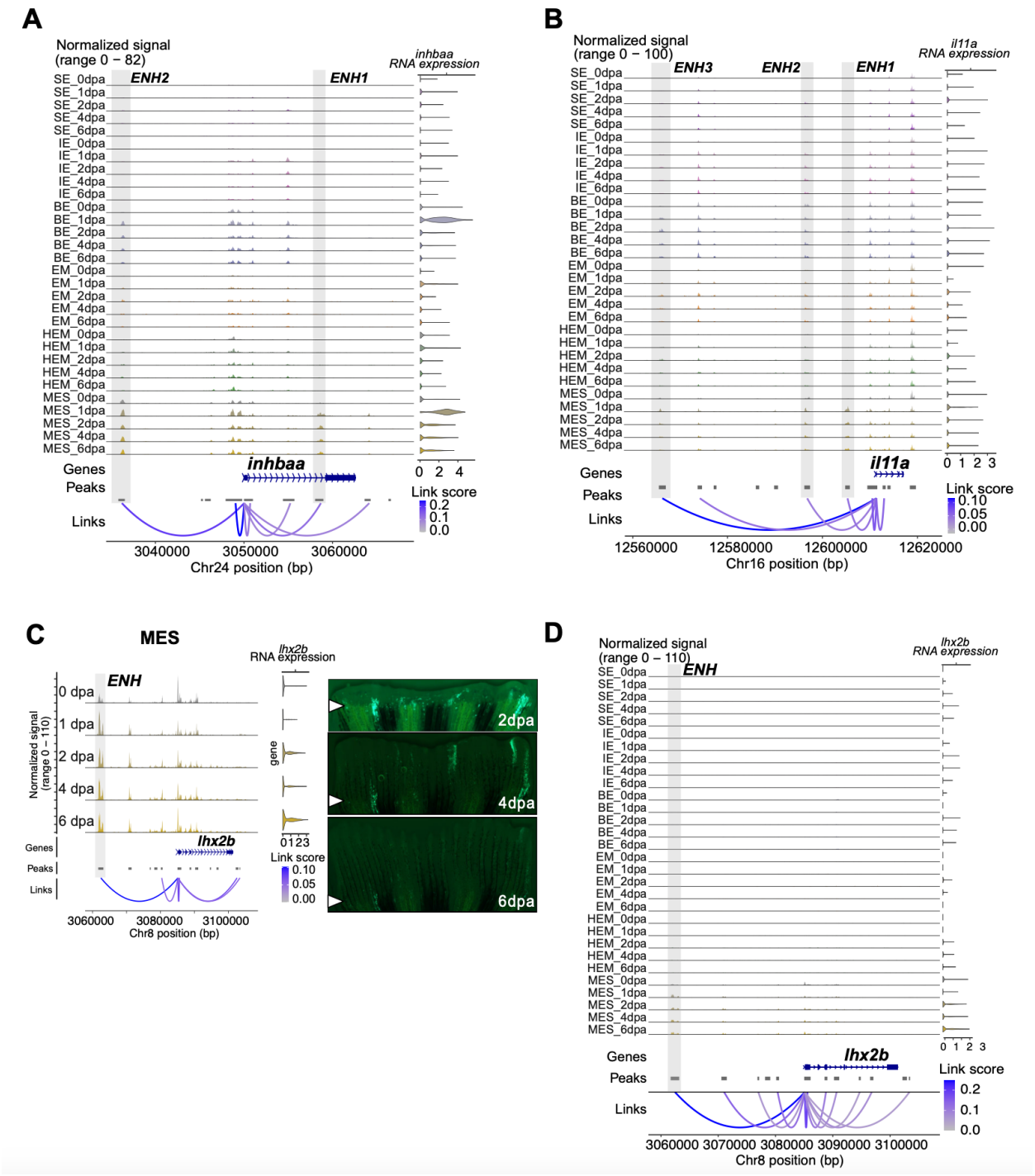
colors indicate the categories of RRRs based on their proximity to the TSS: Promoter (±1 kb around TSS), Proximal (±10 kb around TSS), and Distal (>±10 kb and <±100 kb around TSS). (*E*)(*F*) These panels depict aggregated snATAC-seq signal intensity over time across 5-kb regions flanking the promoters of upregulated/downregulated RRGs in basal epithelial cells and mesenchymal cells, respectively. (*G*)(*H*) Average vertebrate conservation scores (phastCons) and phylogenetic p-scores (phyloP) across 5-kb regions flanking the center of cell type-RRRs and their shuffled regions are plotted for basal epithelial cells and mesenchymal cells, respectively.



Supplemental Figure S4. Association between changes of signals from RRRs and expression of their nearest genes. (A) This panel presents the expression changes as scaled changes for genes located near RRRs whose signals increase or decrease between compared time points. The analysis is segmented into three zones based on their proximity to the TSS: the promoter region (±1kb around TSS), the proximal region (±10kb around TSS, excluding the promoter region), and the distal region (±100kb around TSS, excluding both promoter and proximal regions). This categorization helps delineate the impact of chromatin accessibility changes in different genomic contexts on gene expression. The p-values were calculated using Wilcoxon rank-sum test for comparison between each two groups. ****: p<0.0001, ***: p<0.001, **: p <0.01, *: p<0.1, ns: not significant. (B) Changes in ATAC-seq signal are also depicted as scaled changes, specifically focusing on ATAC peaks closest to genes that identified as RRGs. The p-values were calculated using Wilcoxon signed-rank test for comparison between each two groups. ****: p<0.0001, ***: p<0.001, **: p <0.01, *: p<0.1, ns: not significant.

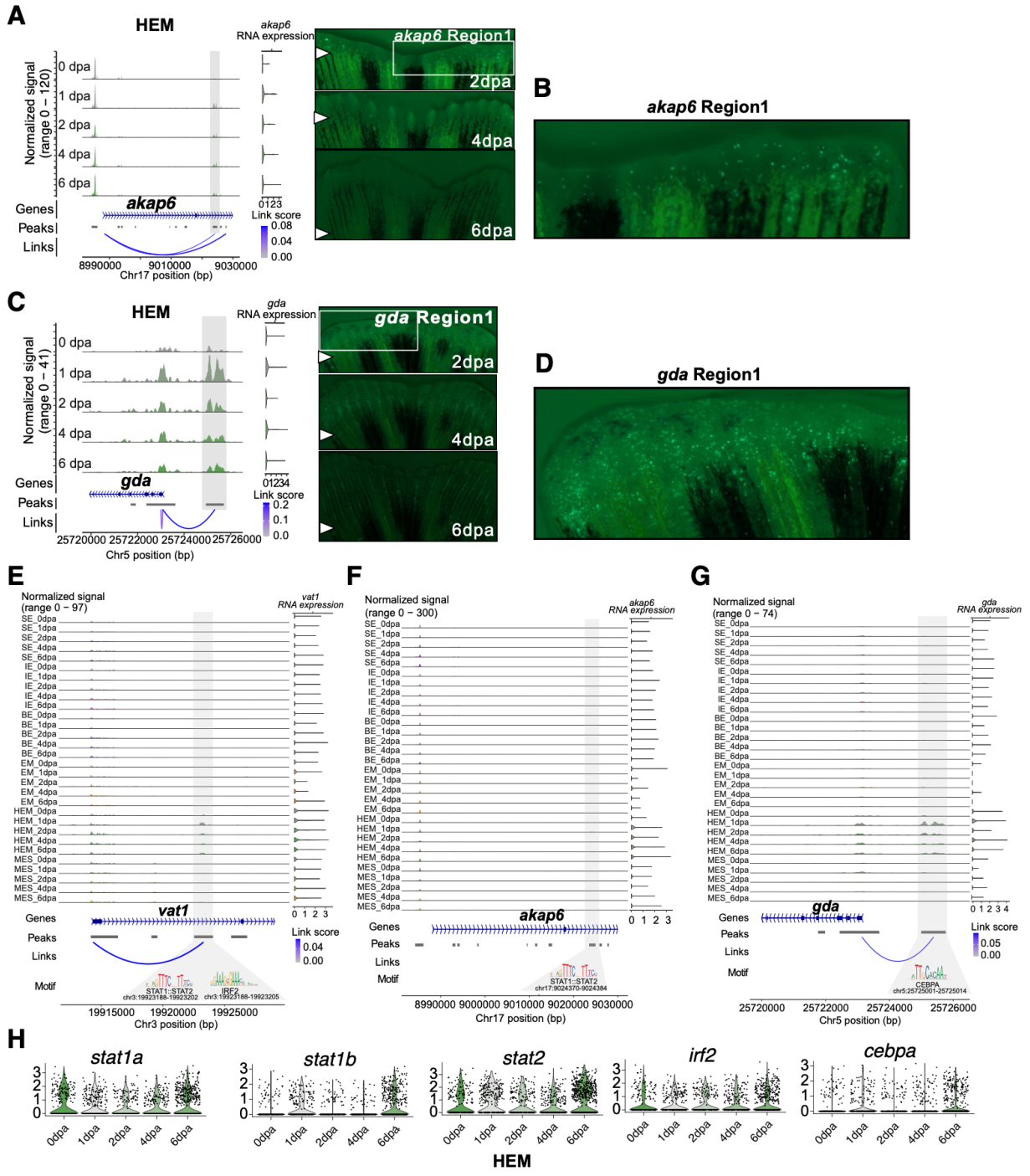


Supplemental Figure S5. Peak-to-gene link metrics. (A) This panel illustrates the overlaps between genes linked to peaks and those nearby peaks in each cell type subset, as well as in the whole dataset. (B) An upset plot shows the intersection of peak-to-gene links identified from cell type subsets and the whole dataset. (C) Elbow plots show genes ranked by the number of peak-to-gene links identified for each gene in each cell type subset and the whole dataset, respectively. Total linked genes are labeled on the top of each elbow plot. (D) Elbow plots show peaks ranked by the number of peak-to-gene links identified for each peak in each cell type subset and the whole dataset, respectively. Total linked peaks are labeled on the top of each elbow plot.



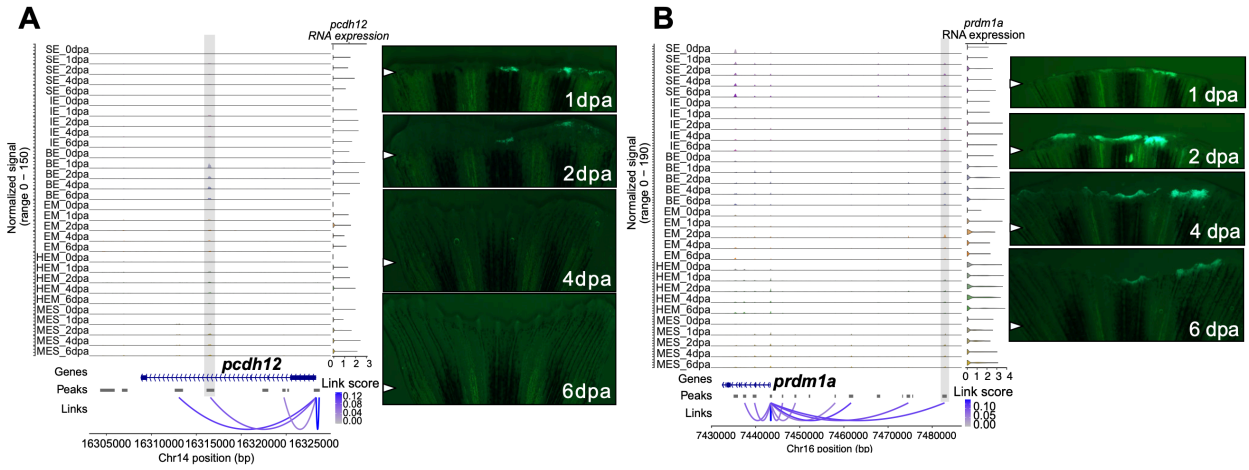
Supplemental Figure S6. Overview of chromatin accessibility signals of mesenchymal-specific enhancer candidates. (A) A genome browser view generated with Signac CoveragePlot displays pseudobulk snATAC-seq signals for each cell type and time point around *inhbaa.* Annotations at the bottom indicate the transcript region, called peaks, and peak-to-gene links identified using the whole dataset. Violin plot on the right side shows the normalized expression levels of *cdh11* across each cell group. (B) Similar to (A), a genome browser view of region around *il11a.* (C) On the left, a genome browser view generated with Signac CoveragePlot shows aggregated and normalized snATAC-seq signals across time points

surrounding gene *lhx2b* in mesenchymal cells. Regions of candidate RREs are shaded in grey. Annotations at the bottom indicate the transcript region, called peaks, and peak-to-gene links identified within mesenchymal cells. On the right, *in vivo* enhancer reporter assay results for this enhancer region are shown across various stages. The white arrows indicate amputation site. (D) Similar to (A), a genome browser view of region around *lhx2b.*

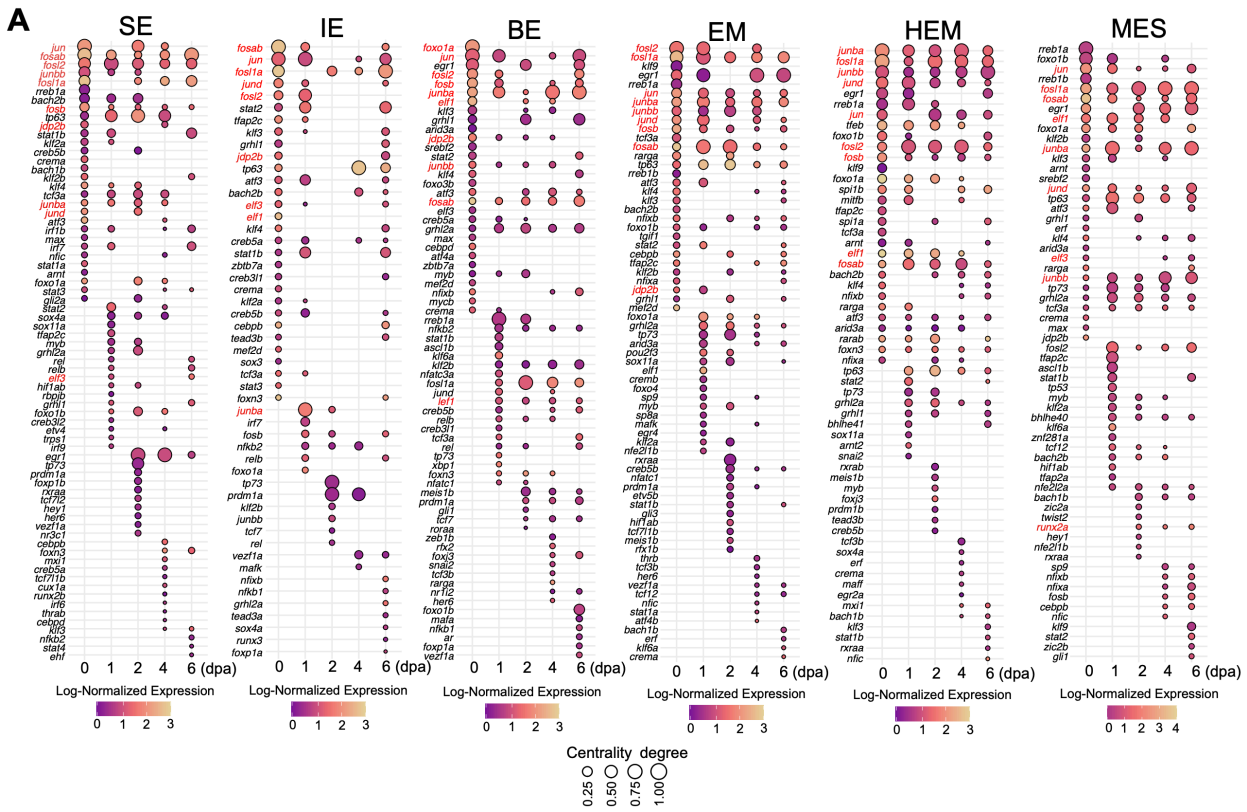


Supplemental Figure S7. Overview of chromatin accessibility signals of hematopoietic-specific enhancer candidates. (A) On the left, a genome browser view generated with Signac CoveragePlot shows aggregated and normalized snATAC-seq signals across time points around gene *akap6* in hematopoietic cells. The candidate RRE region is shaded in grey. Annotations at the bottom indicate the gene transcript region, called peaks, and peak-to-gene links identified within hematopoietic cells. In the middle, a violin plot shows normalized gene expression of *akap6* across time points. To the right, *in vivo* enhancer reporter assay results for this enhancer region are shown across various stages with amputation sites marked by white arrows. (B) This panel provides a zoomed-in views of the 2 dpa regenerating fin shown in (A). (C) Similar to (A), but

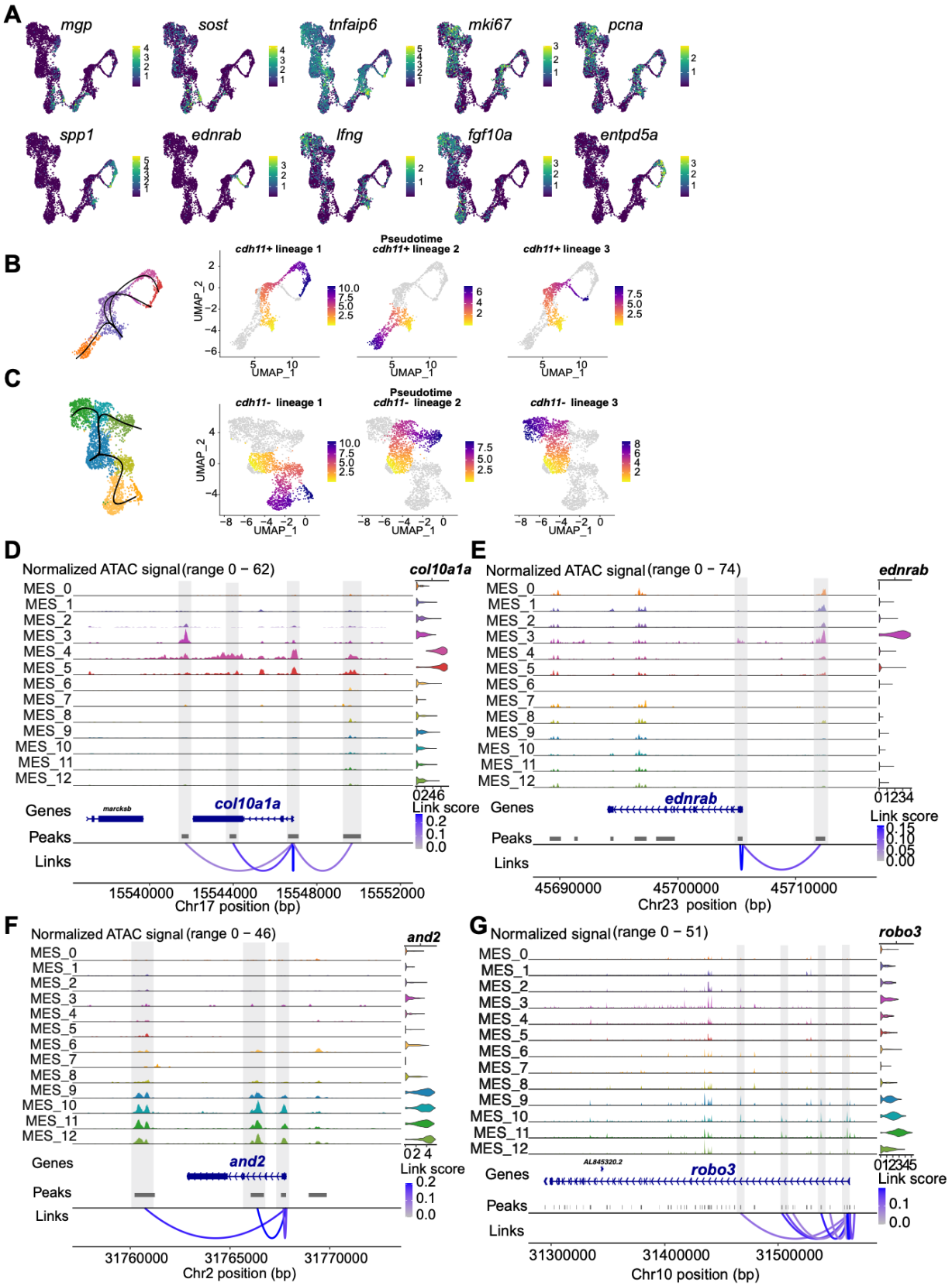
showing regions around *gda*. (D) This panel provides a zoomed-in views from the 2 dpa regenerating fin shown in (C). (E)(F)(G) These genome browser view present aggregated and normalized snATAC-seq signals across time points and cell types of regions around *vat1*, *akap6*, and *gda*, respectively. Regions of candidate RREs are shaded in grey. Annotations at the bottom indicate the transcript region, called peaks, peak-to-gene links, and motifs detected with binding footprint signals in the RREs. On the right, violin plot shows normalized gene expression of the corresponding gene across cell types and time points. (H) Violin plots display normalized gene expression levels of example TFs with footprints detected in the validated hematopoietic enhancers.



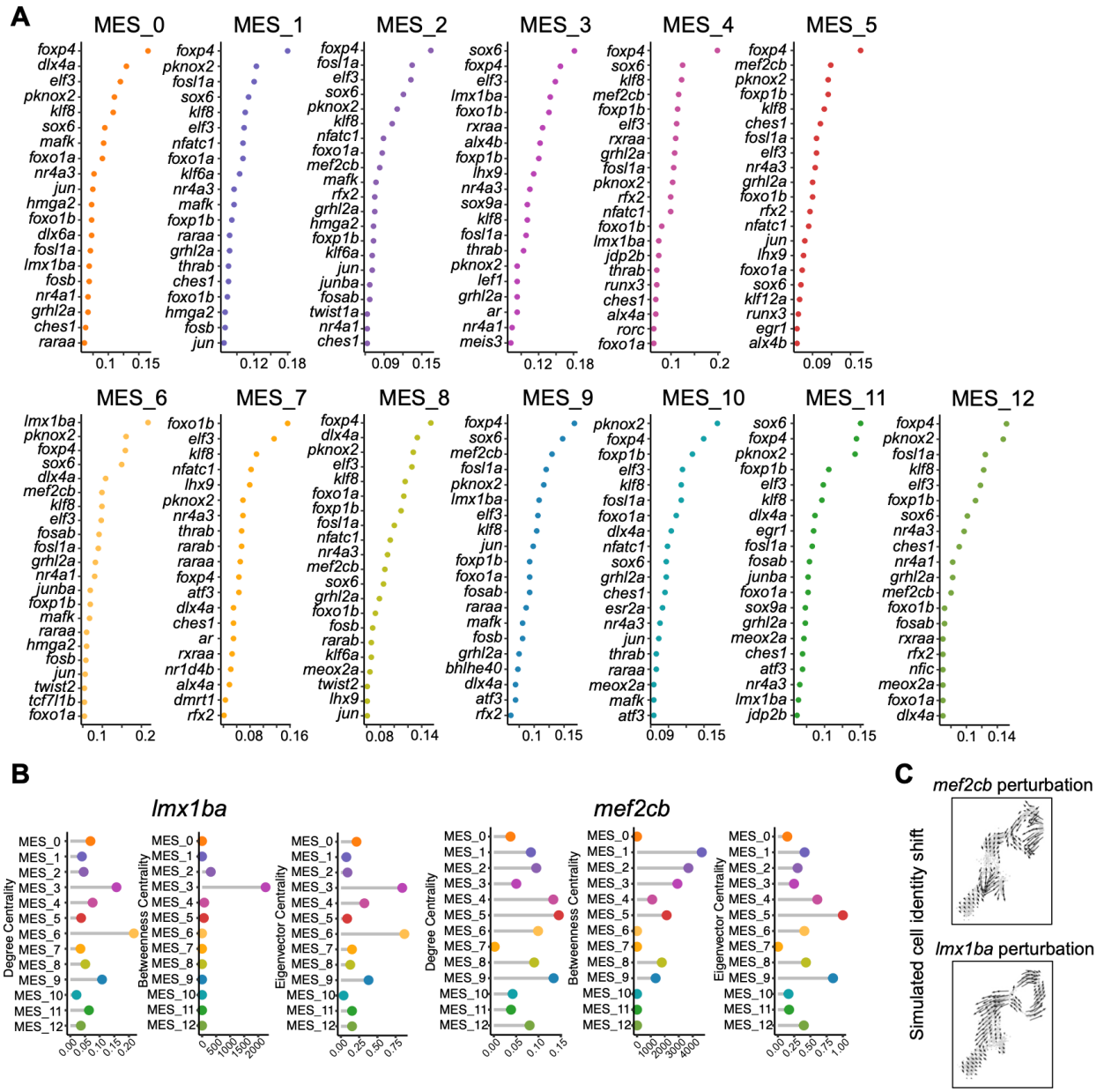
Supplemental Figure S8. Overview of chromatin accessibility signals of cell type-shared enhancer candidates across major cell type. (A) These genome browser view presents aggregated and normalized snATAC-seq signals across time points and cell types of regions around *pcdh12*. Regions of candidate RREs are shaded in grey. Annotations at the bottom indicate the transcript region, called peaks, and peak-to-gene links. In the middle, a violin plot shows normalized gene expression of *pcdh12* across time point. To the right, *in vivo* enhancer reporter assay results for this enhancer region are shown across various stages with amputation sites marked by white arrows. (B) Similar to (A), but showing the signals and reporter assay results of an enhancer near *prdm1a*.



Supplemental Figure S9. Top 30 TFs identified across cell types and stages. (A) A set of dot plots showing top 30 transcription factor regulators in each cell type and time point. TFs were ranked by centrality degree. The centrality degree values were normalized by the maximum degree in each cell type and time point. The dot size represents centrality degree, the dot color represents averaged, and log normalized expressions.


A
mgp
sost
tnfaip6
mki67
pcna
spp1
ednrab
lfng
fgf10a
entpd5a
B
Pseudotime
cdh11+ lineage 1
cdh11+ lineage 2
cdh11+ lineage 3
UMAP_1
UMAP_2
C
cdh11- lineage 1
cdh11- lineage 2
cdh11- lineage 3
D
Normalized ATAC signal (range 0 – 62)
col10a1a
MES_0
MES_1
MES_2
MES_3
MES_4
MES_5
MES_6
MES_7
MES_8
MES_9
MES_10
MES_11
MES_12
Genes
marcksb
col10a1a
Peaks
Links
Link score
15540000
15544000
15548000
15552000
Chr17 position (bp)
E
Normalized ATAC signal (range 0 – 74)
ednrab
ednrab
45690000
45700000
45710000
Chr23 position (bp)
F
Normalized ATAC signal (range 0 – 46)
and2
and2
31760000
31765000
31770000
Chr2 position (bp)
G
Normalized signal (range 0 – 51)
robo3
AL845320.2
robo3
31300000
31400000
31500000
Chr10 position (bp)

Supplemental Figure S10. Marker gene expression and lineage-associated peak signals. (A) UMAP plot showing normalized expression of marker genes in MES cells. (B) Pseudotime lineages inferred from the *cdh11*+ MES population. The first plot on the left displays three lineage trajectories overlaid on the UMAP, while the subsequent three plots show pseudotime progression embedded in the UMAP for each lineage. (C) Similar to (B), but depicting pseudotime lineage inference for the *cdh11*- MES population. (D)-(G) Genome browser views illustrating aggregated and normalized snATAC-seq signals across MES subclusters around the *col10a1a*, *ednrab, robo3*, and *and2* loci, respectively. Lineage-associated peak regions are highlighted in grey. Annotations below indicate transcript regions, called peaks, and peak-to-gene links. To the right, violin plots display the normalized expression levels of the corresponding genes.



Supplemental Figure S11. Overview of key TFs identified in mesenchymal subclusters. (A) Top 20 TFs ranked by centrality scores in each mesenchymal subcluster. (B) Three types of gene regulatory network centrality scores of *mef2cb* and *lmx1ba* across MES subclusters. (C) Simulated cell identify shifts after in-silico perturbation of *mef2cb* and *lmx1ba* in osteoblast MES population, which are embedded on osteoblast MES UMAP. Arrows indicate the direction of identity shifts induced by perturbation.