

# UC Davis

## UC Davis Previously Published Works

### Title

Functional annotation of regulatory elements in rainbow trout uncovers roles of the epigenome in genetic selection and genome evolution.

### Permalink

<https://escholarship.org/uc/item/8r2414gg>

### Authors

Salem, Mohamed

Al-Tobasei, Rafet

Ali, Ali

et al.

### Publication Date

2024-01-02

### DOI

10.1093/gigascience/giae092

Peer reviewed

# Functional annotation of regulatory elements in rainbow trout uncovers roles of the epigenome in genetic selection and genome evolution

Mohamed Salem<sup>1,\*</sup>, Rafet Al-Tobasei<sup>2</sup>, Ali Ali<sup>1</sup>, Liqi An<sup>3</sup>, Ying Wang<sup>3</sup>, Xuechen Bai<sup>3</sup>, Ye Bi<sup>3</sup>, and Huaijun Zhou<sup>3</sup>

<sup>1</sup>Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742-231, USA

<sup>2</sup>Computational Science Program, Middle Tennessee State University, Murfreesboro, TN 37132, USA

<sup>3</sup>Department of Animal Science, University of California, Davis, Davis, CA 95616, USA

\*Correspondence address. Mohamed Salem, Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742-231, USA. E-mail: [mosalem@umd.edu](mailto:mosalem@umd.edu)

## Abstract

Rainbow trout (RBT) has gained widespread attention as a biological model across various fields and has been rapidly adopted for aquaculture and recreational purposes on 6 continents. Despite significant efforts to develop genome sequences for RBT, the functional genomic basis of RBT's environmental, phenotypic, and evolutionary variations still requires epigenome reference annotations. This study has produced a comprehensive catalog and epigenome annotation tracks of RBT, detecting gene regulatory elements, including chromatin histone modifications, chromatin accessibility, and DNA methylation. By integrating chromatin immunoprecipitation sequencing, ATAC sequencing, Methyl Mini-seq, and RNA sequencing data, this new regulatory element catalog has helped to characterize the epigenome dynamics and its correlation with gene expression. The study has also identified potential causal variants and transcription factors regulating complex domestication phenotypic traits. This research also provides valuable insights into the epigenome's role in gene evolution and the mechanism of duplicate gene retention 100 million years after RBT whole-genome duplication and during re-diploidization. The newly developed epigenome annotation maps are among the first in fish and are expected to enhance the accuracy and efficiency of genomic studies and applications, including genome-wide association studies, causative variation identification, and genomic selection in RBT and fish comparative genomics.

## Introduction

Rainbow trout (RBT) is among the most intensively studied fish in many research areas [1]. RBT, native to North America and Asia's Pacific Ocean, has been introduced to every state and province in North America and worldwide to every continent except Antarctica. In the United States, RBT is the most cultivated cool and cold freshwater fish [2]. Considerable biological knowledge has been developed for this species due to the RBT's widespread use as a model and cultivation as a food and sport fish. A plethora of knowledge is available for the biology of RBT, and it serves as a complementary research model for economically important fish other than RBT, such as Atlantic and Pacific salmon species [1].

The recent decade's considerable accumulation of genomic resources underscores the escalating requirement to employ genomic methodologies in RBT-focused research and applications in aquaculture and fisheries [3]. For example, RBT is an ideal model for delving into gene and genome evolution. Its status as a partially tetraploid organism, marked by a unique whole-genome duplication (WGD) event (salmonid-specific fourth WGD), with subsequent partial re-diploidization and significant genome rearrangements, is an appealing subject for genetic exploration. In addition, the potential of elevating aquaculture species, such as RBT, through genomic methodologies is critical to making superior germplasm with enhanced economic traits [3].

The availability of genome sequence references is essential for genomics-based selection. An accurately assembled and annotated genome sequence is the cornerstone, facilitating *in silico*

mapping and validation of single-nucleotide polymorphism (SNP) variants. This, in turn, streamlines the design of SNP chip assays, optimizing the precision of genetic analyses. Furthermore, the genome sequence facilitates functional genomics and proteomic approaches in RBT research [3], unraveling the intricacies of an overly complex and duplicated genomic landscape. This approach drives advancements in genetic understanding and lays the foundation for robust genomic analyses and improvement of the RBT.

Efforts to make a pangenome reference available for RBT have begun, and at least 3 chromosome-level genome assemblies are now available [4, 5]. However, epigenome reference annotations for RBT are lacking and needed to understand the functional genomic basis of the rapidly domesticating RBT's phenotypic, environmental, and evolutionary variations. Annotating the genome for chromatin histone modifications and accessibility is essential for identifying the genome regulatory elements. The chromatin organization of genomic regions involved in functional/regulatory interactions is more accessible to nucleases and other DNA-modifying enzymes due to altered structure and binding of transcription factors [6].

Epigenetics is vital in understanding the cellular and molecular processes, including cell type-specific regulation of gene expression, cellular differentiation, genomic imprinting, embryonic development, and chromosome inactivation. Regions of open chromatin identified by ATAC-seq, combined with expression analysis, allow for associating functional/regulatory ele-

Received: March 28, 2024. Revised: July 16, 2024. Accepted: October 24, 2024

© The Author(s) 2024. Published by Oxford University Press GigaScience. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

ments with transcribed genes [7–9]. Although the genomic DNA sequence is mainly identical in all cells, the chromatin context of the DNA changes from tissue to tissue. Some of the most significant differences are due to posttranslational histone modifications.

The ENCODE project has assayed more than a dozen different histone modifications. H3K4 methylation was first discovered in the RBT testis by Honda et al. [10]. A high abundance of H3K4me3 correlates with promoters of active genes and transcription start sites [11–14], while increased levels of H3K27me3, a repressive mark, are associated with promoters of inactive genes [15, 16]. H3K27ac is a chromatin mark of active regulatory elements and may differentiate active enhancers and promoters from their inactive counterparts [16]. H3K4me1 is a chromatin mark of regulatory elements correlated with enhancers and other distal elements but is also enriched downstream of TSS [16]. Elevated levels of H3K27ac and H3K4me1 are linked with enhancer regions and correlate with open chromatin sites [13, 17]. The combinatorial profile of these different epigenetic marks has been used to predict chromatin states in several species [18–23], including livestock. Using the profiles of histone marks in concert with open chromatin and transcription profiles allows an unprecedented view of the functional elements present in the RBT genome, which is the first in aquaculture species and among the first in fish.

DNA methylation is one of eukaryotes' major epigenetic/epigenomic mechanisms that modify the primary genetic code by converting cytosine into 5-methylcytosines (5mCs). However, in fish, large-scale gene expression studies that reveal the role of DNA methylation have been done in a few species [24–26]. Integrating the DNA methylation data with chromatin modification and accessibility can help understand the regulation of gene expression, tissue complexity, organismal development, and evolution at the systems biology level. Besides, it provides valuable molecular information for the genetic improvement of fish for food production and biomedical purposes.

The Functional Annotation of Animal Genomes (FAANG) consortium provided a functional annotations atlas of farm animal genomes, including pig, cattle, and chicken, for the first time [19, 27, 28]. Currently, there is a dearth of functional annotations for fish, especially aquaculture species. Moreover, epigenomic tracks have only been comprehensively established for zebrafish [29]. In the European Union, the AQUA-FAANG project aims to provide functional annotation tracks of 6 aquaculture species [30]. On the other hand, in the United States, over the past 10 years of the FAANG project, aquaculture has been represented by 1 species, the RBT. As part of this FAANG consortium, the main aim of this study was to annotate the RBT genome for chromatin histone modifications, chromatin accessibility, and DNA methylation by integrating data from chromatin immunoprecipitation sequencing (ChIP-seq), ATAC sequencing (ATAC-seq), and Methyl Mini-seq together with gene expression data from RNA sequencing (RNA-seq) across various tissues of the RBT. The study provides a unique RBT catalog/genome annotation tracks of several tissues in correlation with variation in gene expression. The study also reveals epigenetic functions of previously identified quantitative trait loci (QTL) for complex phenotypic traits important for domestication by mapping QTL onto genome tracks of the new gene regulatory elements, including promoters, enhancers, super-enhancers, and transcription factor binding sites. The study also offers insights into the epigenome's role in gene evolution after the genome duplication in RBT.

## Results

### Overview of the sequencing dataset

Approximately 1.59 billion ChIP-seq reads, 1.06 billion ATAC-seq, 0.53 billion RNA-seq, and 1.0 billion Methyl Mini-seq were used in these analyses, with average mapping rates of 97%, 94%, 81.3%, and 79%, respectively (Additional File 1). A total of 421,240, 1,057,603, 758,037, 1,392,453, and 1,628,755 peaks were obtained for H3K4me3, H3K4me1, H3K27ac, H3K27me3, and ATAC, with an average peak size of 749, 438, 604, 585, and 691 bp, respectively (Additional File 1).

Figure 1A shows the signal intensity of each epigenetic mark relative to the transcription start site (TSS) of the protein-coding genes. The ATAC-seq signal peaked around the TSS. The major peaks for H3K4me3 and H3K27ac were observed at about 500 nt in front of TSS, with minor peaks shortly after TSS. H3Kme1 showed moderate peaks about 1,000 nt upstream of TSS and right after.

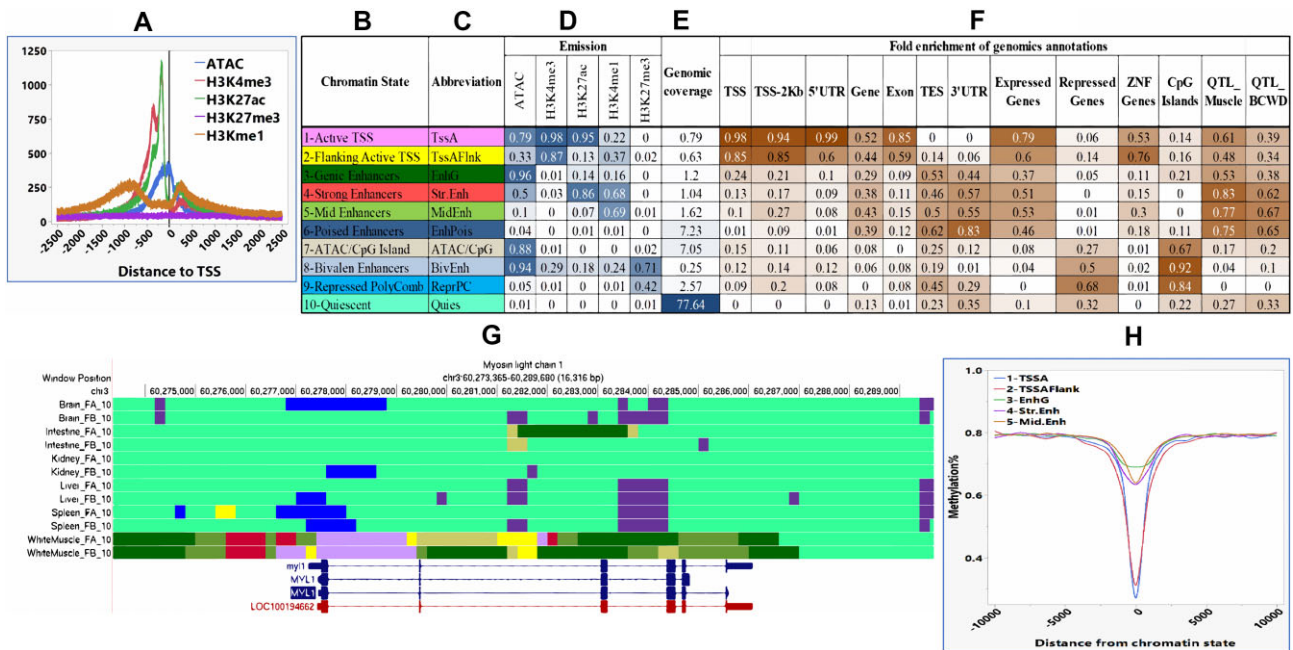
### Identification and characterization of 10 chromatin states in the rainbow trout genome

Genome-wide epigenomics mappings were generated by integrating 4 histone modifications ChIP-seq data sets (H3K4me3, H3K4me1, H3K27ac, and H3K27me3), chromatin accessibility (ATAC-seq), and DNA methylation (Methyl Mini-seq). Data from 6 major tissues (brain, liver, spleen, white muscle, intestine, and kidney) were included in all analyses except for ATAC-seq, where data from the first 3 tissues were available. The epigenomic marker integration predicted 10 categories of chromatin states in the RBT genome (Fig. 1B–F).

The first predicted 2 states were (i) active TSS (TssA), indicating active promoters, and (ii) flanking active TSS (TssAFlnk), together covering 1.42% of the genome. Strong epigenomic signals of H3K4me3, H3K27ac, and intermediate H3K4me1 signal, with no H3K27me3, characterized these 2 active chromatin states. TssA had higher ATAC-seq signals compared to the TssAFlnk state. As expected, these active promoter states were enriched around protein-coding gene TSS and TSS flanking regions (2 kb), Zink finger transaction factors, and highly transcribed (TPM > 2) genes but depleted in the repressed genes (TPM < 0.2) (Fig. 1B–F).

Chromatin states 3 to 6 are composed of 4 types of enhancers: (i) genic enhancers (EnhG) characterized by very strong open chromatin signal and moderate H3K27ac and H3K4me1 signals; (ii) strong active enhancers (Str.Enh) characterized by strong H3K27ac, and H3K4me1 signals but moderate open chromatin; (3) intermediate active enhancers (MidEnh) with moderate/strong H3K4me1 signal, and (4) poised enhancers (EnhPois). It is important to note that the EnhPois showed minimal chromatin modification and openness signals, yet ChromHMM identified them as a chromatin state. The first three active enhancer states (EnhG, Str.Enh, and MidEnh) cover 3.86%, while the EnhPois spans 7.2% of the genome. These enhancers were enriched in QTL (discussed below), highly expressed genes, the 3'UTR/TES (especially EnhPois), and gene bodies but depleted in the repressed genes (Fig. 1B–F).

The seventh chromatin emission state, covering 7.25% of the genome, was characterized by relatively strong ATAC-seq signals, enrichment in CpG island regions, and moderate enrichment in the suppressed genes. The eighth chromatin state, named bivalent enhancers (BivEnh), is characterized by open chromatin (ATAC-seq), strong repressor H3K27me3 signal, and weak promoter/enhancer signals from H3K4me3, H3K27ac, and H3K4me1. The ninth chromatin state represented the repressed/polycomb (ReprPC) regions spanning 2.57% of the genome and moderately enriched in 3'UTR/TES (Fig. 1B–F). Both BivEnh and ReprPC were



**Figure 1:** Discovery and characterization of chromatin marks and states in the rainbow trout genome. (A) Epigenetic mark's signal intensity from ATAC-seq, H3K4ME3, H3K27ac, H3Kme1, and H3K27me3 ChIP-seq relative to the protein-coding genes' TSS. (B, C) Names and abbreviations of 10 chromatin states identified in the rainbow trout genome. (D) Epigenetic mark probabilities associated with each chromatin state indicated in numbers (0–1) and color intensity. (E) Percentage of genomic coverage of each chromatin state. (F) Enrichment of each chromatin state associated with various genomic annotations, including genes, TSS and flanking regions ( $\pm 2$  kb around TSS and TES), expressed genes (TPM  $\geq 2$ ), and repressed genes (TPM  $< 0.2$ ), CpG islands, and QTL for fish/muscle growth, fillet quality, and bacterial cold water disease (BCWD). (G) UCSC genome browser tracks showing the landscape of the chromatin states at the MLC1 gene in 6 tissues. Only in muscle is MLC1 flanked by strong enhancers (red), weak enhancers (dark green), and active TSS states (purple). In the other tissues, MLC1 had quiescent (light green) or poised enhancers (blue). (H) Average methylation levels relative to the position of each active chromatin state (1–5).

**Table 1:** Count, percentage, and mean/median length (bp) of each chromatin state

State	Count	Percentage	Mean/median state length (bp)
1-TSSA	496,173	4.33	1,382/1,200
2-TSSAFlnk	604,309	5.27	1,081/1,000
3-EnhG	1,167,656	10.19	1,008/600
4-Str.Enh	651,095	5.68	1,550/1,000
5-MidEhh	1,439,241	12.56	993/600
6-EnhPois	1,520,923	13.27	6,295/3,800
7-ATAC-CpG	2,962,192	25.85	1,078/400
8-BivEnh	254,795	2.22	574/400
9-ReprPC	618,191	5.39	4,762/2,000
10-Quies	1,746,493	15.24	179,332/88,600
<b>Total</b>	<b>11,461,068</b>	<b>100.00</b>	

enriched in CpG islands and genes with no or minimal expression. The 10th chromatin status was quiescent (Quies), with poor chromatin modification signals covering most of the genome (77.67%) (Fig. 1B–F).

The chromatin states were used to generate genome annotation tracks available through the UCSC genome browser (see data availability). Table 1 summarizes the annotation tracks characterization with 515,159 chromatin states; the active chromatin states (1–5) represent 24.8% of the state counts, and the nonactive states (6–10) represent 75.2%. There were 47,433 active promoters, 80,404 active enhancers (EnhG, Str.Enh, and MidEnh), and 50,353 repressed enhancers (EnhPois and BivEnh). Table 1 also shows each chromatin state's mean and median length, with the

enhancers' medians ranging between 400 and 1,000 bp and a repressed polycomb median of 2,000 bp.

Figure 1G shows an example of the UCSC genome browser tracks displaying the chromatin regulatory states at the myosin light chain 1 (MLC1) gene in 6 tissues. Only in muscle is MLC1 flanked by strong enhancers and active TSS states; the other 5 tissues showed poised enhancers or quiescent states.

The density of each chromatin state relative to the position of TSS of the protein-coding genes is shown in [Additional File 2](#). The TssA and, to a lesser extent, TssAFlnk showed maximum enrichments at TSS. The other chromatin states showed enrichment around 5 kb on both sides of TSS.

The ATAC-seq data were missing in 3 tissues. The ChromHMM tool was found to be reliable for predicting states based on combinations of available chromatin tracks. In a study using the mouse ENCODE project data, ChromHMM accurately assigned chromatin states even when certain marks, such as ATAC-seq, were missing [31]. The research concluded that the absence of certain marks did not significantly impact the assignment. Although the states most affected were those most enriched in the missing mark in a 10-mark model, the opposite was not necessarily true, highlighting redundancy between the marks.

To address the effect of the missing ATAC-seq data in 3 tissues, we created a separate 10-state chromatin model without ATAC-seq data and uploaded the bed files to the journal portal *Folder name: TroutChromHMM\_State\_without\_ATAC\_seq*. A quick comparison of enhancers with and without ATAC-seq data revealed a 25% reduction in genetic enhancers and a 23% increase in strong enhancers using the muscle and kidney data. The mouse ENCODE paper reported similar differences in the number of enhancers [31].

## DNA methylation relative to the chromatin states

There were distinct patterns of DNA methylation near and within each chromatin state, as shown in Fig. 1H. All the active chromatin states (1–5) were hypomethylated compared to their flanking regions. As expected, the promoter TssA and its flanking regions TssAFlnk were strongly hypomethylated. Similarly, all the active enhancers (EnhG, Str.Enh, MidEnh) were moderately hypomethylated.

## Histone modification association with gene expression

### Histone marks correlation with gene expression

We characterized the enrichment of the tissue-specific histone marks at promoter regions of tissue-specific expressed genes among 6 tissues. To do that, genes showing more than 10-fold increases in expression compared to the rest of the tissues or more than 1 TPM value with zero TPM expression in other tissues were first identified. Then, histone marks uniquely identified within  $\pm 3$  kb from TSS (including  $-3$  kb of the promoter region) of the same gene showing tissue-specific expression were cross-listed. The number of the tissue-specific histone marks was divided by the total number of each histone mark in the genome to obtain a normalized relative abundance of each histone mark. Data showed that H3K4me1 was enriched in the tissue-specific genes compared to the same genes in other tissues where the genes are silent or scarce (chi-square  $P < 0.001$ ). Conversely, H3K27me3 was enriched in the silenced genes, compared to the tissue-specific expressed genes (chi-square  $P < 0.001$ , Fig. 2A, Additional File 1).

We also looked at the association of histone marks within  $\pm 3$  kb of TSS to gene expression. Densities of chromatin marks ATAC-seq, H3K4me1, H3K4me3, and H27Kac were higher in the genes with expression values more than 1 TPM (log<sub>10</sub> TPM equals zero). On the other hand, H3K27me3 chromatin mark density was higher in genes with less expression (Fig. 2B, Additional File 3). There was significant correlation between the histone marks and the log<sub>10</sub> TPM values ( $P < 0.001$ ,  $R^2 = 0.074$ ).

### Chromatin states correlation with gene expression

We identified 5,551 tissue-specific chromatin states within  $\pm 10$  kb of genes' TSS. There were 2,150 genes with tissue-specific gene expression and chromatin states, suggesting a correlation in gene expression (Additional File 1). All the active chromatin states (states 1–5, active promoter, and enhancers) were enriched in genes with tissue-specific expression, especially the strong enhancers. Notably, EnhPois and, to a lesser extent, ATAC-CpG states were also enriched, indicating the involvement of other epigenetic mechanisms in regulating gene expression (Fig. 2C). To get more insight into the correlation between chromatin state and gene expression, we looked at the distribution of each chromatin state density near genes with various relative gene expression levels. As seen in Fig. 2D, the chromatin state densities of the open chromatin states within  $\pm 3$  kb of TSS, including TssA and TssAFlnk, and enhancers, including EnhG, Str.Enh, MidEnh, and EnhPois, were higher in the genes with expression level more than 1 TPM (log<sub>10</sub> TPM equals zero). On the other hand, chromatin states RepPC, ATAC-CpG, and BivEnh did not show characteristic density patterns relative to gene expression. There was a negligible correlation between the chromatin states and the log<sub>10</sub> TPM values ( $R^2 < 0.01$ ), though.

## DNA methylation correlation with gene expression

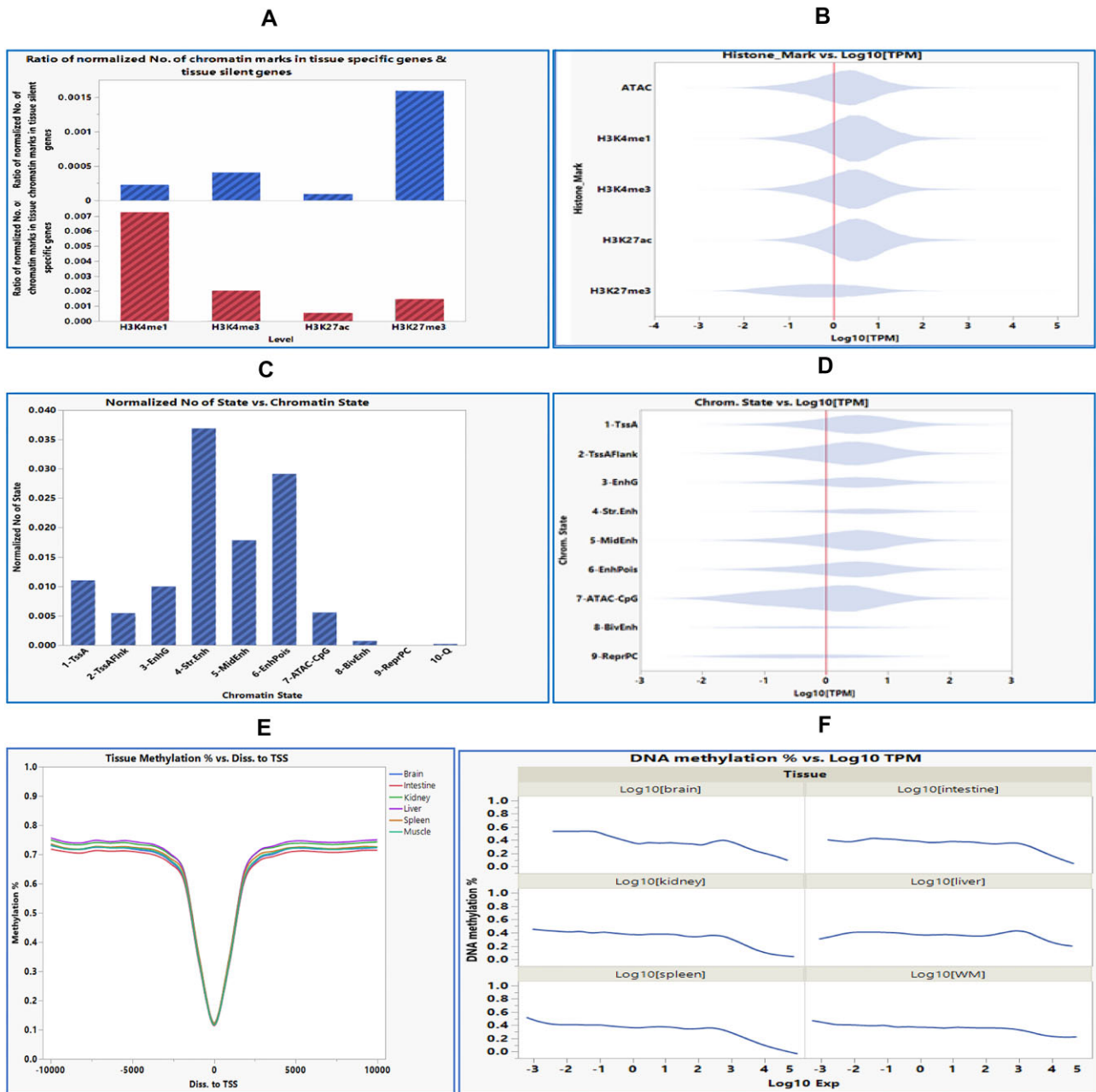
We characterized the methylation level near and within genes,  $\pm 10$  kb flanking TSS. The mean level of CpG methylation more than  $\pm 5$  kb flanking TSS was about 75%; however, a sharp decrease in DNA methylation to about 10% on average was observed at the TSS (Fig. 2E). Regarding the DNA methylation correlation with gene expression, our data showed a weak ( $R^2 = 0.002$ – $0.04$  depending on the distance to TSS) but statistically significant correlation between the average percentage of DNA methylation within  $\pm 3$  kb flanking TSS and gene transcription expression ( $P < 0.0001$ ) (Additional File 2). As seen in Fig. 2F, there was a trend of negative correlation between DNA methylation and gene expression, especially of the most highly expressed genes, with a log<sub>10</sub> TPM value of more than 3; the correlation varies between tissues, though.

## Detection and characterization of super-enhancers

We identified a total of 5,799 nonredundant super-enhancers (SEs) in all studied tissues (Additional File 4). Super-enhancers are clusters of enhancers enriched within 12.5 kb of the genome. Figure 3A shows the ranked SEs identified by HOMER based on an extremely high H3K27ac signal compared to conventional enhancers [32]. There was 5,104 SEs within or neighboring 4,120 genes within 10 kb. Of those SEs, there was an average of 850.5 SEs in all tissues, ranging from 630 in the spleen to 1,167 in the intestine (Fig. 3B, Additional File 4). The SEs had an average length of 25,234 bp, reaching a maximum length of 133 kb (Fig. 3C). Figure 3D shows the chromosome distribution of the SEs with an average of 159 SEs per chromosome. The SEs were generally shared between tissues, with 599 (13.8%) SEs ubiquitously existing in all tissues and only 805 (10.3%) SEs existing in a single tissue. For example, a muscle-specific SE at location NC\_048582.1:54022672-54039559 was associated with the muscle-specific gene guanosine monophosphate reductase (GMPR). The SEs were enriched around the gene TSS (Fig. 3E). SEs were also enriched in highly expressed genes, with 4,737 unique SEs overlapping with expressed genes (TPM values  $> 2$ ) and only 286 SEs overlapping in the repressed genes (TPM  $< 0.2$ ). Figure. 3F shows an example of a super-enhancer with H3K27ac signal flanking the C1QTNF4 gene only in muscle compared to a typical enhancer in all other tissues. Gene Ontology (GO) enrichment analysis of the SE neighboring genes showed involvement in important molecular functions, including catalytic activity, DNA, and metal/ion binding. In the biological process, SE genes were enriched in biosynthetic, cellular metabolic process, and transcription (Additional File 4).

## Enhancers in QTL

To demonstrate the utility of the new chromatin annotations in identifying potential causal variants for complex phenotypic traits important for domestication, we cross-matched previously identified QTL in the RBT genome with genome tracks of the new gene regulatory elements, including promoters and enhancers. We used previously identified QTL with known genomic locations for fish growth, muscle yield, fillet quality, and bacterial cold water disease (BCWD) [33–38]. We identified 2,074 Str.Enh, overlapped with QTL-harboring genes located on 15 chromosomes, with mean and median overlap lengths of 1,524 and 1,000 bp, respectively (Fig. 4A, Additional File 5). We also found 847 MidEnh overlapped with QTL-harboring genes located on 15 chromosomes, with mean and median overlapping lengths of 1,084 and 800 bp, respectively. Additionally, 3,975 EnhG enhancers over-



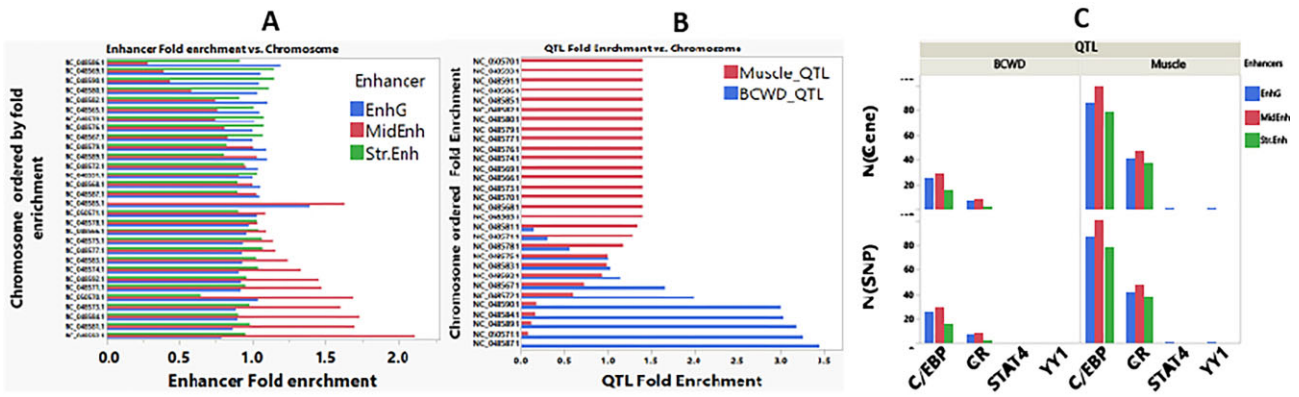
**Figure 2:** Histone modification and chromatin state correlation with gene expression. (A) Enrichment of H3K4me1 histone mark within  $\pm 3$  kb of TSS of the tissue-specific genes (top) and H3K27me3 in the genes silenced in other tissues (bottom). (B) Densities of the chromatin marks within  $\pm 3$  kb of TSS relative to gene expression. ATAC-seq, H3K4me1, H3K4me3, and H27Kac density were higher in the genes with expression levels of more than 1 TPM ( $\log_{10}$  TPM equals zero). Conversely, H3K27me3 density was higher in genes with less expression. (C) Enrichment of chromatin states (1–7), particularly Str.Enh within  $\pm 3$  kb of TSS of the tissue-specific genes. (D) Densities of the open chromatin states within  $\pm 3$  kb to TSS relative to gene expression. Chromatin state densities of TssA and TssAFlnk and enhancers EnhG, Str.Enh, MidEnh, and EnhPois were higher in the genes with expression levels of more than 1 TPM ( $\log_{10}$  TPM equals zero). Conversely, chromatin states RepPC and the ATAC-CpG and BivEnh did not show characteristic density patterns relative to gene expression. (E) Average methylation percentage relative to TSS. (F) DNA methylation percentage relative to gene expression ( $\log_{10}$  TPM).

lapped with QTL-containing genes on all chromosomes, with mean and median overlap lengths of 874 and 600 bp, respectively (Additional File 4). Figure 4A, B shows the QTL and enhancers' fold enrichment (observed/expected) per chromosome. There were 124 fish/muscle growth and 84 BCWD unique QTL overlapping with 239 unique SEs (Additional File 5).

To further investigate the epigenetic function of the SNPs in QTL, we looked at SNPs within QTL that overlap with the genic, strong, and mid-enhancers and have transcription factor bind-

ing motifs (TFBMs). A total of 112 SNPs that met these criteria were located within 4 TFBM spanning 85 genes involved in fish/muscle growth, fillet quality, and BCWD (Fig. 4C, Additional File 5). Interestingly, most TFBMs (99%) were classified into only 2 families. The first TFBM family was C/EBP (with 3 TF members, C/EBP alpha, beta, and delta), making up 69.4% of the TFBMs. The second TFBM family comprises glucocorticoid receptor (GR) and GR beta, constituting 30.1% of TFBMs (Fig. 4B). These data suggest a significant role of C/EBP and GR transcrip-





**Figure 4:** Enhancers in QTL. Fold enrichment per chromosome of active enhancers (A) that overlap with QTL (B) for fish/muscle growth, fillet quality, and BCWD. (C) Number of SNPs and genes within QTL that overlap with genic, strong, and mid-enhancers and have transcription factor binding motifs (TFBMs) mainly belonging to glucocorticoid receptors (GRs) and C/EBP transcription factors.

ancestral genes of the Northern pike. The analysis revealed the presence of 73.6% gene conservation cases, 14.2% neofunctionalization cases, 12% specialization cases, and 0.2% subfunctionalization cases (Additional File 6).

We compared the fold enrichment of the histone marks and the abundance of chromatin states within the promoter region located 2 kb upstream of the TSS of each gene copy. Compared to neofunctionalized genes, there was less divergence in the histone modification profiles of conserved gene paralogues (Wilcoxon test,  $P = 7.13E-270$ ) (Fig. 5A). H3K27ac of the conserved gene pairs exhibited the highest correlation compared to H3K4me1 (Wilcoxon test,  $P = 7.42E-99$ ) and H3K4me3 (Wilcoxon test,  $P = 4.84E-08$ ). The H3K4me3 profile of the neofunctionalized gene pairs showed the most significant dissimilarity compared to the conserved genes (Wilcoxon test,  $P = 1.07E-163$ ).

Similarly, the chromatin states in the promoter region of conserved gene pairs exhibited the highest correlation compared to neofunctionalized (Wilcoxon test,  $P = 4.53E-46$ ) and specialized genes (Wilcoxon test,  $P = 4.89E-11$ ) (Fig. 5B). In addition, we observed less abundance of Str.Enh within the first 7 chromatin states, upstream of the TSS of conserved genes. Except for BivEnh and RepPC, it was observed that the abundance of states upstream of the TSS was higher in gene pairs that are maintained through conservation (Wilcoxon test,  $P < 2.2e-16$ ) (Fig. 5C). Table 2 also shows the relative enrichment of all the chromatin states in each gene category. The single-copy genes had strong promoter and moderate signals compared to the conserved genes, which had strong promoter and enhancer signals. The neofunctionalized and specialized genes had moderate enhancer signals and very weak promoter signals.

## Discussion

The pioneering ENCODE projects built the foundations for discovering the regulatory element and their functions in humans and mammalian model species [20, 40, 41]. Following the ENCODE models, in the past decade, the FAANG consortium provided a functional annotations atlas of the farm animal genomes, including pig, cattle, and chicken, for the first time [19, 27, 28]. However, functional annotations of fish genomes are still in their infancy, with comprehensive epigenomics tracks available perhaps only for zebrafish [29]. In the United States, over the past 10 years of the FAANG project, aquaculture was represented by 1 species, the RBT. As a part of the FAANG project, this study thus aimed to iden-

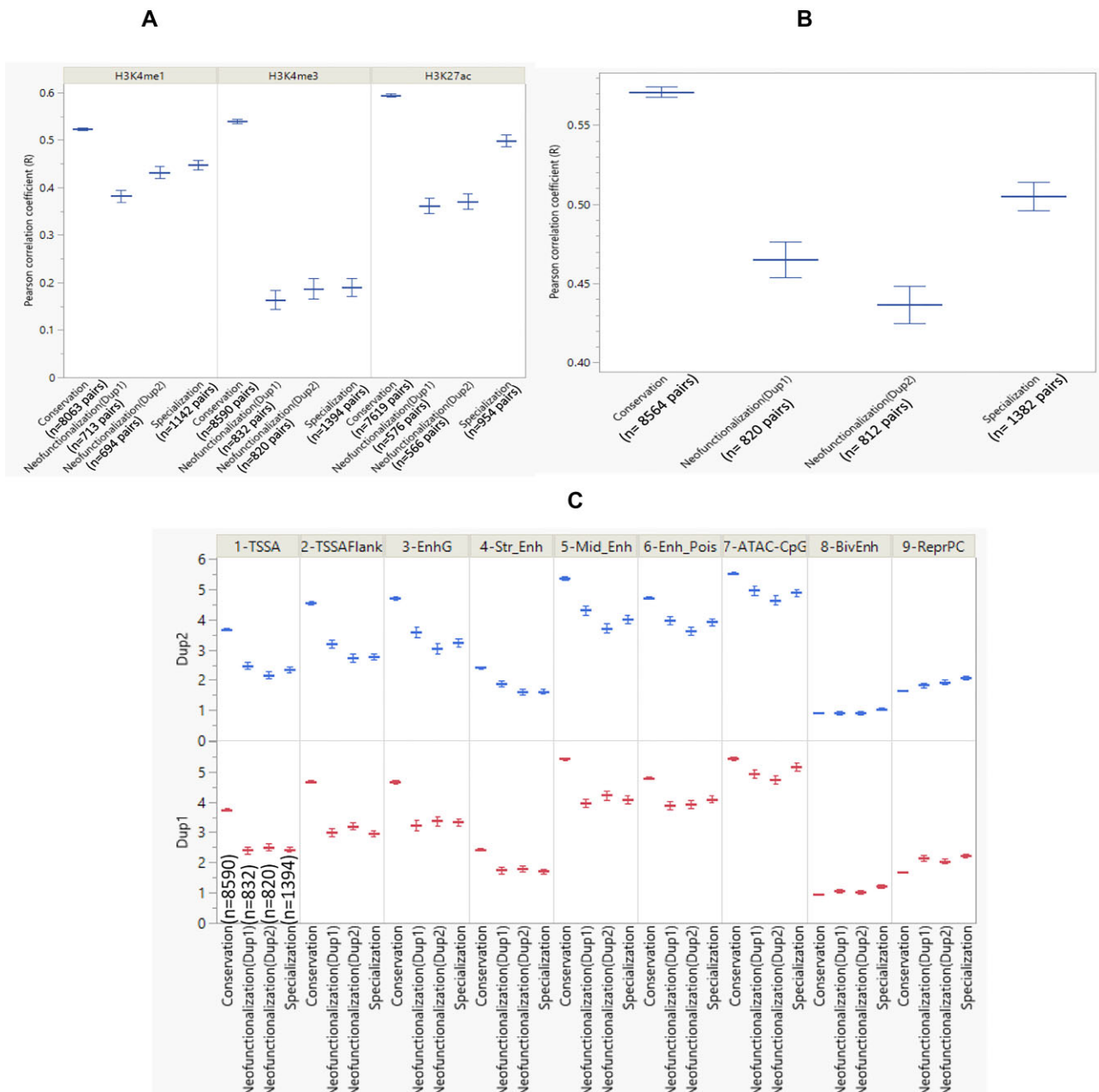
tify and characterize an atlas of regulatory elements and provide epigenome annotation tracks from RBT populations in the United States. We developed and characterized an atlas of regulatory elements and epigenome annotation tracks of the RBT. ChIP-seq, ATAC-seq, Methyl Mini-seq, and RNA-seq data were integrated across RBT tissues to identify gene regulatory elements, including chromatin histone modifications, chromatin accessibility, and DNA methylation.

This study identified regulatory elements, including 47,433 active promoters (19,784 TssA and 27,649 TssAFlnk). When this manuscript was ready for publication, the Ensemble genome browser released chromatin tracks for RBT, including promoters, enhancers, and open chromatin stats. For comparison, the Ensemble genome annotation browser has 23,394 promoters [42]. A total of 29,302 active promoters in our study were shared with promoters in the Ensemble genome browser (>100 nt). We also identified 80,404 active enhancers and 50,353 repressed enhancers, together (130,757) covering about 11.34% of the genome. The Ensemble genome annotation has 102,440 enhancers. Of all the enhancers identified in our study, 71,382 overlapped in genome positions with enhancers in the Ensemble genome browser (>100 nt). Variation in the numbers of the regulatory elements between our results and the Ensemble browser is expected due to differences in fish populations, tissues, physiological conditions, and the bioinformatics pipelines. In zebrafish, efforts to characterize the chromatin landscape identified 140,000 cis-regulatory elements [29]. And in mice, 33% of the genome had a chromatin signature of promoter, enhancer, transcriptional, and heterochromatin states [41].

In this study, the RBT active promoter and enhancer chromatin states were enriched around the genes TSS and TSS-flanking regions and zinc finger transcription factors and were highly transcribed but depleted in the repressed genes (Fig. 1). The RBT enhancers were also enriched in the expressed genes. Consistent with our results, the chicken genome promoters were more enriched in TSS, 5'UTR, and CpG islands than in enhancers. The chicken active promoters and enhancers were more enriched in the TSS and the gene body of the highly expressed than the repressed genes [18].

This study also identified distinct patterns of DNA methylation associated with each chromatin state (Fig. 1). All the active chromatin states (1–5) were hypomethylated compared to their flanking regions. On the other hand, the poised enhancers and quiescent genome regions were hypermethylated. The bivalent en-





**Figure 5:** Association of the chromatin mark/state with the mechanism of duplicate gene retention. (A) The histone mark divergence of each category of the duplicate gene pair was quantified using the Pearson correlation coefficient of each histone mark profile for the duplicated gene pairs. Gene pairs with conserved expression exhibit the highest correlation of histone mark profiles upstream of TSS. (B) Correlation of state counts between gene pairs within each category. Gene pairs with conserved expressions demonstrate the highest correlation of state counts. (C) The shared number of states 1–7 within 2 kb upstream of the TSS is higher in conserved gene pairs. The shared numbers of bivalent enhancer (BivEnh) and repressive polycomb complex (RepPC) showed a lower correlation.

hancers were strongly hypomethylated, the ATAC-CpG state was slightly hypermethylated, and the repressed polycomb showed no change in the methylation levels. Similar DNA methylation patterns were observed in the pig genome, where the promoter and the TSS transcribed states were hypomethylated, and the enhancer states showed intermediate methylation levels [19]. Previously, we also reported a sharp decline in DNA methylation within the  $\pm 2$  kb of the TSS of the muscle genes [43].

We characterized the enrichment of the tissue-specific chromatin marks at promoter regions of tissue-specific expressed genes. H3K4me1 was enriched in the tissue-specific genes compared to the same genes in other tissues (silenced genes). On

the other hand, H3K27me3 was enriched in the tissue-silenced genes compared to the tissue-specific expressed genes (Fig. 2). In addition, chromatin marks ATAC-seq, H3K4me1, H3K4me3, and H27Kac were enriched in the expressed genes ( $>1$  TPM), while H3K27me3 was enriched in promoters of the silenced genes (Fig. 2). Similarly, the open chromatin states involving promoters and enhancers were more enriched in the genes with more than 1 TPM expression value. On the other hand, the repressed chromatin states RepPC, ATAC-CpG, and BivEnh did not show characteristic density patterns relative to gene expression (Fig. 2). Consistent with our results, in the pig genome, the active chromatin states (promoters, transcribed regions, and enhancers) were en-

**Table 2:** Relative enrichment of the chromatin states in the single-copy genes compared to the conserved, neofunctionalized, and specialized genes

State	Single-copy	WGD	Conservation	Neofunctionalization	Specialization
TSSA	0.75	0.51	0.67	0.00	0.00
TSSAFlnk	0.82	0.41	0.52	0.09	0.08
EnhG	0.17	0.32	0.39	0.19	0.30
StrEnh	0.16	0.45	0.52	0.32	0.39
MidEnh	0.28	0.50	0.58	0.24	0.30
EnhPois	0.25	0.45	0.52	0.26	0.33
ATAC/CpG	0.14	0.05	0.12	0.00	0.06
BivEnh	0.08	0.03	0.07	0.01	0.13
ReprPC	0.00	0.00	0.00	0.09	0.16
Quies	0.16	0.12	0.19	0.09	0.10

riched in tissue-specific genes, while the repressed states were depleted [19]. In cattle, the relationships between chromatin states and gene expression showed that genes with TssA had the highest expression compared to genes with EnhPois, BivFlnk, and ReprPC [28].

Regarding the DNA methylation, we noticed a sharp decline in DNA methylation level within  $\pm 3$  kb, flanking the genes' TSS. There was a trend of a weak negative correlation between DNA methylation and gene expression, especially in the most highly expressed genes ( $\log_{10}$  TPM  $> 3$ ), and the correlation varied between tissues (Fig. 2). These data confirm our previous reports showing a weak to moderate negative correlation between DNA methylation levels and gene transcription expression in muscle. The correlation was dependent on CpG position relative to TSS. The correlation was negative within  $\pm 1$  kb of the TSS and positive in the gene body [43].

We have identified a total of 5,799 unique SEs in the RBT genome. Each tissue contained an average of 850.5 genes overlapping/neighbors SEs (Fig. 3B). The SEs were generally shared between tissues, with only 805 (10.3%) SEs existing in a single tissue, and the rest were shared between more than 1 tissue. In total, 599 (13.8%) SEs were ubiquitously existing in all tissues. SEs in zebrafish showed more tissue specificity in 4 of 5 tissues than regular enhancers [44].

GO analysis of the SEs' neighboring genes revealed functions relevant to essential molecular functions, including catalytic activity, DNA and metal/ion binding, and biological processes, including biosynthetic and cellular metabolic processes and transcription. SEs play a crucial role in determining cell identity and have been linked to the development of diseases [45]. Genes located within or near SEs had a higher gene expression than other genes, consistent with previous reports in mammals [46].

This study explored the potential epigenetic functions of previously identified QTL for complex phenotypic traits important for domestication by mapping QTL onto genome tracks of the regulatory elements. The active enhancer states (EnhG, StrEnh, and MidEnh) and the EnhPois were enriched in genome regions spanning QTL. We identified 2,074 StrEnh, 847 MidEnh, and 3,975 EnhG enhancers overlapped with QTL-containing genes on all chromosomes. Similar to our data, a recent study on cattle confirmed that active promoters/transcripts exhibited the highest enrichment for QTL. The cattle study also showed that weak enhancers had the

highest enrichment for expression quantitative trait loci (eQTLs) compared to 14 other chromatin states [28].

We took a closer look to investigate the potential epigenetic functions of the SNPs in QTL that overlap with the enhancers and have TFBMs. Out of 108 SNP markers within 84 genes involved in fish/muscle growth, fillet quality, and BCWD, we identified 8 TFBMs (Fig. 4). Interestingly, almost all the TFBMs (99%) were classified into only 2 families: the C/EBP and the GRs.

The glucocorticoid hormone is key in regulating muscle mass, and prolonged cell exposure to it causes muscle atrophy [47]. Muscle-specific deletion of GR in mice skeletal muscle increases muscle mass, reducing fat mass and muscle atrophy [48, 49]. Similarly, C/EBP $\beta$  is a central regulator of cancer muscle mass loss (cachexia) via promoting the expression of atrophy-inducing factors [50]. In RBT, stress increases cortisol levels and susceptibility to BCWD. A recent study found that rainbow trout BCWD-resistant fish are less sensitive to a cortisol-induced IgM response than susceptible/control fish [51]. Another recent study by de Laval et al. [52] revealed that short-term lipopolysaccharide-induced immune signaling can activate C/EBP $\beta$ -dependent chromatin accessibility, leading to trained immunity in hematopoietic stem cells during secondary infection. This establishes an epigenetic mechanism of memory function in innate immunity.

Our data regarding the C/EBP and GR warrant further studies to include CRISPR-Cas9 gene editing to confirm the causative nature of the SNPs involved in C/EBP and GR transcription factors and their role in regulating muscle growth, fillet quality, and BCWD. The muscle growth and quality and BCWD QTL analysis targeted in this study is an example of the potential utility of the genome annotation tracks generated as valuable tools in prioritizing genetic variants when searching for causal variants and alleles with major effects on domestication traits and genomic selection.

The ancestral genome of teleost fish underwent a teleost-specific third WGD (Ts3R), estimated to have occurred 225–333 million years ago [53], followed by the divergence of the Salmonidae family, which underwent a fourth salmonid-specific WGD (Ss4R), estimated to have occurred ~80–100 million years ago [39]. The recent salmonid-specific WGD and the existence of large genome segments as duplicate regions make RBT unique as a model organism to study the early stages of gene evolution. Therefore, we sought to identify the evolutionary processes that drive the preservation of gene duplicates and gain a better un-

derstanding of the role of epigenomes in gene evolution following WGD and during the rediploidization of RBT.

To distinguish the evolutionary processes that drive the preservation/neofunctionalization of gene duplicates after WGD, gene expression profile divergence was quantified among 11,654 ohnologue pairs in RBT and their ancestral singletons in the Northern pike. This phylogenetic approach was initially developed by Assis and Bachtrog [54]. The analysis revealed the presence of 73.6% gene conservation cases, 14.2% neofunctionalization cases, 12% specialization cases, and 0.2% subfunctionalization cases. These results indicate that conservation maintains the majority of the gene duplicates following WGD. In Atlantic salmon, Lien et al. [39] reported that 42% of the Ss4R duplicates displayed conserved co-expression with their orthologs in Northern pike.

Enhancers and promoters predominantly enrich epigenetic signatures [55, 56]. We thus postulated that genes displaying noticeable variations in gene expression would also exhibit contrasting epigenetic patterns. To validate this hypothesis, we compared the fold enrichment of the histone marks and the abundance of chromatin states within the promoter region located 2 kb upstream of the TSS of each gene copy. Compared to neofunctionalized genes, there was less divergence in the histone modification profiles of conserved gene paralogues. H3K27ac of the conserved gene pairs exhibited the highest correlation compared to H3K4me1 and H3K4me3 (Fig. 4A). In their recent study on Atlantic salmon, Verta et al. [57] reported that the transcriptional divergence observed in duplicated genes resulting from WGD is correlated with variations in the number of nearby regulatory elements, suggesting that the functional divergence between ohnologues following WGD is primarily driven by enhancers. In this study, the H3K4me3 profile of the neofunctionalized gene pairs showed the most significant dissimilarity compared to the H3K4me3 profile of the conserved genes, which aligns with the divergence observed in gene expression. Our results suggest a role for the promoters in the functional divergence between ohnologues following WGD.

Similarly, the chromatin states in the promoter region of conserved gene pairs exhibited the highest correlation compared to neofunctionalized and specialized genes, which may help explain their increased stability and conservation (Fig. 5B). Furthermore, compared to other enhancers, we observed less abundance of Str.Enh upstream of the TSS in the conserved genes. Also, except for BivEnh and RepPC, the abundance of the chromatin state upstream of the TSS was higher in gene pairs maintained through conservation (Fig. 5C). Together, our study reveals significant enrichment of distinct epigenetic signatures in ohnologue pairs exhibiting divergent gene expression modes.

Overall, this study provides a new atlas of regulatory elements in the RBT genome, which will help accelerate the genetic selection efforts, mainly through genome-wide association studies and genomic selections, to improve essential production traits in RBT for domestication. In addition, the new chromatin atlas will help in understanding the functional genomic basis of RBT's phenotypic, environmental, and evolutionary variations.

## Methods

### Animals and tissues

Six tissues (brain, intestine, liver, kidney, spleen, and white muscle) were collected at Washington State University, Dr. Gary Thorgaard's laboratory, from 2 individual doubled haploid Swanson clonal line fish. Tissues were flash-frozen in liquid nitrogen before being stored at  $-80^{\circ}\text{C}$  until further processing. The Institu-

tional Animal Care and Use Committee at Washington State University reviewed and approved the animal study under protocol #02456.

### ChIP-seq and ATAC-seq

ChIP-seq (H3K4me3, H3K27ac, H3K4me1, and H3K27me3) library preparations were performed using the iDeal ChIP-seq kit (Diagenode, cat. C01010059), as previously described [18, 19]. In brief, approximately 20–30 mg powdered tissue was cross-linked using 1% formaldehyde for 8 minutes before quenching with 100  $\mu\text{L}$  glycine for 10 minutes. Cell nuclei were isolated by centrifugation at  $2,000 \times g$  for 5 minutes, resuspended in 600  $\mu\text{L}$  iS1 buffer, and incubated on ice for 30 minutes. Chromatin was sheared using a Bioruptor Pico for 10 to 15 cycles, depending on the tissues. For immunoprecipitation, about 1–1.5  $\mu\text{g}$  of sheared chromatin was used as input with 1  $\mu\text{g}$  of the specific histone mark antibody according to the manufacturer's protocol: H3K4me3 (part of the Diagenode iDeal Histone kit #C01010059), H3K27me3 (#C15410069), H3K27ac (#C15410174), and H3K4me1 (#C15410037). An input with no antibody was used as a negative control for each sample. NEBNext Ultra DNA library prep kit (#E7645L) from New England Biolabs was used for library construction. Libraries were sequenced using an Illumina HiSeq 4000 platform (RRID:SCR\_016386) with a single-end read length of 50 bp. Additionally, ATAC-seq libraries were prepared using a modified Omni-ATAC57 protocol on cryopreserved nuclei [58]. The DNA sequencing was performed on Illumina's NextSeq platform, with a 40-bp paired-end read length. Sequencing reads were trimmed with Trim Galore (RRID:SCR\_011847) (v.0.6.5) [59] and aligned with bowtie2 (RRID:SCR\_016368) [60] (v.2.5.4a) to the RBT genome (NCBI Accession GCA\_013265735.3), and then duplicates were marked using Picard (RRID:SCR\_006525) (v.2.18.7). MACS2 was used to call regions of signal enrichment ("peaks") [61]. The correlations between assays, tissues, and biological replicates were performed by deepTools (RRID:SCR\_016366) [62].

### Chromatin state annotation

ChromHMM69 (v.1.20) was used to predict the chromatin state by integrating ChIP-seq (H3K4me3, H3K4me1, H3K27ac, H3K27me3, and input control) from 2 biological replicates of all 6 tissues and ATAC-seq data from 3 tissues (brain, liver, and spleen). A 10-state model was chosen to represent the most appropriate number of distinct states based on the histone marks and accessibility combinations and their enrichment [18, 19]. In addition, the fold enrichment of each chromatin state for each gene annotation element (e.g., TSS, 5'UTR, and QTL) was calculated by  $(C/A)/(B/D)$ , where A, B, C, D are the number of bases in a chromatin state, a gene element, overlapped between a chromatin state and a gene element, in the genome, respectively.

### RNA sequencing data

RNA sequencing data for the 6 tissues used in this study were downloaded from our previously described NCBI BioProject PRJNA389609. Sequence read mapping to genome reference and assessment of TPM expression values per gene was performed using the CLC genomics workbench (Qiagen).

### Methyl-MiniSeq

Genome-wide bisulfite library preparation and sequencing were done using the Methyl-MiniSeq Service at Zymo Research as previously described [43]. Briefly, DNA was extracted using Quick-DNA Plus Miniprep Kit. Then, 500 ng genomic DNA was digested with 60

units of Taq $\alpha$ I followed by 30 units of MspI (NEB) and then purified with Zymo Research DNA Clean & Concentrator-5. According to Illumina's guidelines, DNA fragments were ligated to adapters containing 5'-methylcytosine instead of cytosine. The adaptor-ligated fragments of 150–250 bp and 250–350 bp were retrieved from a 2.5% NuSieve 1:1 agarose gel using the Zymoclean Gel DNA Recovery Kit. The EZ DNA Methylation-Lightning Kit was used for the bisulfite treatment. PCR was performed, and then the products were purified using DNA Clean & Concentrator-5 for sequencing on an Illumina HiSeq.

Raw FASTQ files were adapter- and quality-trimmed using TrimGalore 0.6.5 [59]. Filled-in nucleotides were also trimmed using TrimGalore 0.6.5. Reads with a quality <20 were removed. Bismark 0.22.3 was used to align the sequence reads to the RBT genome (NCBI Accession GCA\_013265735.3) [63]. The methylated and unmethylated read totals for each CpG site were retrieved using the Bismark Methylation Extractor. CpG sites with fewer than 10 read depths or more than the 99.9th percentile of coverage in each sample were filtered out to account for PCR bias. The methylation level of the cytosines was calculated as the number of reads calling C divided by the total number of reads calling C and T, as previously described [43]. JMP Pro Version 15 (SAS Institute) was used to generate figures and statistical measures of the association between DNA methylation percent and gene transcription expression levels.

### Histone marks correlation with gene expression

To assess the enrichment of the chromatin marks and states around the TSS of the tissue-specific expressed genes among tissues, we first determined the TPM value of each gene in each tissue. The expression level of each gene in a specific tissue was compared to its expression level in all remaining tissues. For a gene to be tissue specific, the fold-change in the expression level of the gene had to be  $\geq 10$ -fold than the sum of the TPM values in all other tissues, or the TPM value of the gene had to be  $\geq 1$  and the rest of the other tissues zero. The same genes were considered silenced genes in the other tissues (showing no or almost no expression) for comparison. Second, we identified the chromatin mark or state that uniquely exists in the tissue-specific genes within  $\pm 3$  kb of TSS in each gene. JMP Pro Version 15 (SAS Institute) was used to generate figures and statistical measures of the association between gene transcription expression levels and densities of the chromatin marks and states.

### Identification of super-enhancers

The HOMER algorithm findPeaks tool was utilized to identify peaks and calculate ChIP-seq tags from the H3K27ac ChIP-seq bam files. The parameter of finding histone-enriched regions (-style histone) was used. H3K27ac-enriched signals were used to identify enhancers [32, 64]. Enhancers that were located within 12.5 kb of each other were clustered together. The enhancer clusters were then ranked based on H3K27ac signals using the HOMER super-enhancer tool. Enhancers with a tangent slope greater than 1 were considered super-enhancers, while enhancers with a tangent slope less than or equal to 1 were considered conventional enhancers. Nonredundant super-enhancers were determined by merging (at least an overlap of 50% of SE length) across all tissues. Genes overlapped with SE were annotated for GO molecular functions and biological processes using DAVID [65].

### Enhancers and transcription factor binding sites in QTL

Previously identified QTL associated with fish growth, muscle growth, fillet quality, and BCWD were used as gene elements in the chromatin state analyses explained above [33–38]. Genes overlapped with enhancer states in QTL were identified. Then, we searched for SNPs within QTL that overlapped with the genic, strong, and mid-enhancers and were located within transcription factor binding motifs. The transcription factor binding motifs were identified by PROMO [22] using Version 8.3 of TRANSFAC software. SNPs within these motifs that may affect transcription factor binding were identified. The most common motifs associated with fish/muscle growth and fillet quality traits were presented.

### Histone mark/state role in gene evolution following WGD

#### Identification of genes in collinear blocks

The RBT protein sequences and genomic positions were obtained from the NCBI database (Accession number "GCA\_013265735.3"). For genes with multiple transcripts, the transcript with the longest coding sequence (CDS) was selected. To determine homology, protein-coding genes were compared against themselves using BLASTp, specifically the all-vs.-all local BLASTp approach. The top 5 hits, excluding self-hits, with an E-value threshold of less than  $10^{-5}$  for each protein sequence were recorded. This process allowed for identifying potential homologous proteins across the rainbow trout genome.

The MCScanX software package [66] was utilized to categorize genes into 5 distinct types based on their copy number and genomic distribution. These types include singletons, dispersed duplicates, tandem duplicates, proximal duplicates, and WGD/segmental duplicates. To execute the duplicate gene classifier, a core program of MCScanX, the BLASTp output, and the annotation file were used as input files.

The classification of gene duplication was determined as follows: initially, all genes were labeled as singletons and assigned ranks based on their order on chromosomes. Genes that exhibited BLASTp hits to other genes were then relabeled as dispersed duplicates. Gene pairs were classified as proximal duplicates if their difference in gene rank was less than 20 (configurable) or as tandem duplicates if the difference in gene rank was equal to 1. Finally, the MCScanX program was executed, and anchor genes within collinear blocks were relabeled as segmental/WGD duplicates.

In cases where a gene appeared in multiple hits, it was assigned to a unique class based on the following order of priority: WGD/segmental duplicates, tandem duplicates, proximal duplicates, and dispersed duplicates.

#### Divergence of histone modifications

We first calculated the log<sub>2</sub>-transformed fold enrichment ratio. Then, we converted these ratios into z scores using the formula  $Z_x = (\chi - \mu)/\delta$  as in [67]. In this equation,  $\chi$  represents the ratio value for a specific gene,  $\mu$  denotes the mean ratio of all genes, and  $\delta$  signifies the standard deviation of this ratio across all genes.

To assess the correlation and divergence of histone modification patterns between duplicate gene pairs, we utilized the Pearson correlation coefficient  $r$  of the histone modification profiles for the duplicated gene pair and dissimilarity index  $(1 - r)$ , respectively. By comparing the mean values of  $r$  or  $1 - r$  in each gene category, we determined the significance using the Wilcoxon rank-sum test.

## Quantification of gene expression

To quantify gene expression, we obtained the raw RNA-seq reads of RBT (Acc# SRP108798) and Northern pike (Acc# SRP040114) from the NCBI SRA database. To ensure data quality, these raw reads were then subjected to trimming using the CLC Genomics Workbench (version 22.0).

Next, we mapped the high-quality reads to the reference genome sequence (GCF\_013265735.2) using the HISAT2 aligner [68]. To retrieve the abundance levels of each gene, we utilized the BAM files and employed the TPMCalculator to calculate the gene expression levels based on the number of uniquely mapped reads to each gene.

## Identification of the mechanisms of duplicate gene preservation

The WGD duplicates, obtained from the output file that contains collinear blocks identified by MCScanX (RRID:SCR\_022067) [66], were subjected to a blast analysis against noncollinear genes from the Northern pike. If both members of the duplicate gene pair matched the same singleton (with an E-value  $<10^{-5}$ ), the gene triplet was selected for further downstream analysis.

We limited our analyses to triplets, where every gene copy is expressed in at least 1 tissue. To determine the expression prior to duplication, we used the singletons' expression profile in male Northern pike as a proxy. All absolute expression levels were then converted into relative expression levels, representing the proportions of contributions to total expression. These relative expression values were employed as gene expression profiles for comparison.

We employed the phylogenetic method developed by Assis and Bachtrog [54] and Perry and Assis [54, 69] to categorize the evolutionary processes and mechanisms that retain pairs of duplicate genes. To determine the preservation of these duplicates, we calculated the Euclidean distances between the expression profiles of D1 and ancestral copies ( $E_{D1,A}$ ), D2 and ancestral copies ( $E_{D2,A}$ ), and the combined D1–D2 expression profile and that of the ancestral copy ( $E_{D1+D2,A}$ ). To establish a baseline level of gene divergence, we also calculated the Euclidean distances between the expression profiles of singletons in sister species ( $E_{S1,S2}$ ). We explored various cutoff values to define expression divergence and ultimately selected the semi-interquartile range from the median due to its robustness to outliers. Based on previously established rules, we classified each pair of duplicates as conserved, neofunctionalized, subfunctionalized, or specialized. In cases where duplicates are conserved, we expect  $E_{D1,A} \leq E_{S1,S2}$  and  $E_{D2,A} \leq E_{S1,S2}$ . For neofunctionalization of D1, we anticipate  $E_{D1,A} > E_{S1,S2}$  and  $E_{D2,A} \leq E_{S1,S2}$ . Similarly, for neofunctionalization of D2, we expect  $E_{D1,A} \leq E_{S1,S2}$  and  $E_{D2,A} > E_{S1,S2}$ . In cases where duplicates are subfunctionalized, we anticipate  $E_{D1,A} > E_{S1,S2}$ ,  $E_{D2,A} > E_{S1,S2}$ , and  $E_{D1+D2,A} \leq E_{S1,S2}$ . Finally, for the specialized duplicates, we anticipate that  $E_{D1,A}$ ,  $E_{D2,A}$ , and  $E_{D1+D2,A}$  are all greater than  $E_{S1,S2}$ .

## Proofreading

Grammarly (2024) was used for text improving and proofreading [70].

## Additional Files

**Additional File 1.** Overview of the sequencing dataset, QC, and enrichment of histone marks/states in tissue-specific genes versus silenced genes.

**Additional File 2.** Density of each chromatin state relative to the position of TSS of the protein-coding genes and correlation between DNA methylation and gene expression.

**Additional File 3.** Association of histone marks within  $\pm 3$  kb of TSS to gene expression.

**Additional File 4.** Super enhancers.

**Additional File 5.** Enhancers, super-enhancers, TFBM in QTL.

**Additional File 6.** Retention mechanisms for rainbow trout gene duplicates—73.6% gene conservation cases, 14.2% neofunctionalization cases, 12% specialization cases, and 0.2% subfunctionalization cases.

**Additional File 7.** Description file showing the location of each chromatin state by chromosome.

## Abbreviations

5mC: 5-methylcytosine; BCWD: bacterial cold water disease; CDS: coding sequence; ChIP-seq: chromatin immunoprecipitation sequencing; FAANG: Functional Annotation of Animal Genomes; GMPR: guanosine monophosphate reductase; GO: Gene Ontology; GR: glucocorticoid receptor; MLC1: myosin light chain 1; RBT: rainbow trout; RNA-seq: RNA sequencing; SE: super-enhancer; SNP: single-nucleotide polymorphism; TFBM: transcription factor binding motif; TSS: transcription start site; WGD: whole-genome duplication.

## Author Contributions

M.S. and H.Z. designed the research. A.A., L.A., Y.W., X.B., and Y.B. performed the experiments. R.A., A.A., and M.S. analyzed the data, and M.S. wrote the manuscript. M.S. and R.A. contributed equally to the research.

## Funding

This study was supported by competitive grants 2020-67015-30770, 2021-67015-33388, and 2023-67015-39742 from the United States Department of Agriculture, National Institute of Food and Agriculture (MS).

## Data Availability

RNA sequence data for the 6 tissues used in this study are available via the NCBI BioProjects PRJNA389609. The ChIP-seq and ATAC-seq data have been submitted to the NCBI Geo database under accession number GSE245212. The genome reference is available through the NCBI accession number GCF\_013265735.2. Additional files are openly available in the OSF Platform [71]. The main supporting data, including the UCSC Genome Browser epigenome state and marks annotation tracks, are available in the GigaScience repository, GigaDB [72].

## Competing Interests

The authors declare that they have no competing interests.

## Ethics Approval and Consent to Participate

Fish tissues were collected at Washington State University, Dr. Gary Thorgaard's laboratory, from 2 individual doubled haploid Swanson clonal line fish. The Institutional Animal Care and Use Committee at Washington State University reviewed and approved the animal study under protocol #02456.

## References

1. Thorgaard GH, Bailey GS, Williams D, et al. Status and opportunities for genomics research with rainbow trout. *Comp Biochem Physiol B Biochem Mol Biol* 2002;133(4):609–46. [https://doi.org/10.1016/s1096-4959\(02\)00167-7](https://doi.org/10.1016/s1096-4959(02)00167-7).
2. Harvey DJ. *Aquaculture Outlook – U.S. Aquaculture Production Higher in 2006, Outlook Report*, Economic Research Service, United States Department of Agriculture (USDA), LDP-AQS-23, March 9, 2006.
3. Aquaculture Genomics G, Breeding W, Abdelrahman H, et al. Genetics and breeding in the United States: current status, challenges, and priorities for future research. *BMC Genomics [Electronic Resource]* 2017;18(1):191. <https://doi.org/10.1186/s12864-017-3557-1>.
4. Gao G, Magadan S, Waldbieser GC, et al. A long reads-based de-novo assembly of the genome of the Arlee homozygous line reveals chromosomal rearrangements in rainbow trout. *G3* 2021;11(4). <https://doi.org/10.1093/g3journal/jkab052>.
5. Pearse DE, Barson NJ, Nome T, et al. Sex-dependent dominance maintains migration supergene in rainbow trout. *Nat Ecol Evol* 2019;3(12):1731–42. <https://doi.org/10.1038/s41559-019-1044-6>.
6. Elgin SC. The formation and function of DNase I hypersensitive sites in the process of gene activation. *J Biol Chem* 1988;263(36):19259–62. [https://doi.org/10.1016/S0021-9258\(19\)77625-X](https://doi.org/10.1016/S0021-9258(19)77625-X).
7. Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. *Nature* 2012;489(7414):75–82. <https://doi.org/10.1038/nature11232>.
8. Stergachis AB, Neph S, Sandstrom R, et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* 2014;515(7527):365–70. <https://doi.org/10.1038/nature13972>.
9. Rendeiro AF, Schmidl C, Strefford JC, et al. Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat Commun* 2016;7:11938. <https://doi.org/10.1038/ncomms11938>.
10. Honda BM, Candido PM, Dixon GH. Histone methylation. Its occurrence in different cell types and relation to histone H4 metabolism in developing trout testis. *J Biol Chem* 1975;250(22):8686–89. [https://doi.org/10.1016/S0021-9258\(19\)40726-6](https://doi.org/10.1016/S0021-9258(19)40726-6).
11. Bernstein BE, Humphrey EL, Erlich RL, et al. Methylation of histone H3 lys 4 in coding regions of active genes. *Proc Natl Acad Sci USA* 2002;99(13):8695–8700. <https://doi.org/10.1073/pnas.082249499>.
12. Santos-Rosa HSR, Bannister AJ, Sherriff J, et al. Active genes are tri-methylated at K4 of histone H3. *Nature* 2002;419(6905):407–11. <https://doi.org/10.1038/nature01080>.
13. Shen Y, Yue F, McCleary DF, et al. A map of the cis-regulatory sequences in the mouse genome. *Nature* 2012;488(7409):116–20. <https://doi.org/10.1038/nature11243>.
14. Xiao S, Xie D, Cao X, et al. Comparative epigenomic annotation of regulatory DNA. *Cell* 2012;149(6):1381–92. <https://doi.org/10.1016/j.cell.2012.04.029>.
15. Heintzman ND, Stuart RK, Hon G, et al. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 2007;39(3):311–18. <https://doi.org/10.1038/ng1966>.
16. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489(7414):57–74. <https://doi.org/10.1038/nature11247>.
17. Greer EL, Shi Y. Histone methylation: a dynamic mark in health, disease and inheritance. *Nat Rev Genet* 2012;13(5):343–57. <https://doi.org/10.1038/nrg3173>.
18. Pan Z, Wang Y, Wang M, et al. An atlas of regulatory elements in chicken: a resource for chicken genetics and genomics. *Sci Adv* 2023;9(18):eade1204. <https://doi.org/10.1126/sciadv.ade1204>.
19. Pan Z, Yao Y, Yin H, et al. Pig genome functional annotation enhances the biological interpretation of complex traits and human disease. *Nat Commun* 2021;12(1):5848. <https://doi.org/10.1038/s41467-021-26153-7>.
20. Consortium EP, Moore JE, Purcaro MJ, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 2020;583(7818):699–710. <https://doi.org/10.1038/s41586-020-2493-4>.
21. Gerstein MB, Lu ZJ, Van Nostrand EL, et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* 2010;330(6012):1775–87. <https://doi.org/10.1126/science.1196914>.
22. Messeguer X, Escudero R, Farre D, et al. PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics* 2002;18(2):333–34. <https://doi.org/10.1093/bioinformatics/18.2.333>.
23. mod EC, Roy S, Ernst J, et al. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 2010;330(6012):1787–97. <https://doi.org/10.1126/science.1198374>.
24. Wan ZY, Xia JH, Lin G, et al. Genome-wide methylation analysis identified sexually dimorphic methylated regions in hybrid tilapia. *Sci Rep* 2016;6:35903. <https://doi.org/10.1038/srep35903>.
25. Beemelmanns A, Ribas L, Anastasiadi D, et al. DNA methylation dynamics in Atlantic salmon (*Salmo salar*) challenged with high temperature and moderate hypoxia. *Front Marine Sci* 2021;7. <https://doi.org/10.3389/fmars.2020.604878>.
26. Anastasiadi D, Diaz N, Piferrer F. Small ocean temperature increases elicit stage-dependent changes in DNA methylation and gene expression in a fish, the European sea bass. *Sci Rep* 2017;7(1):12401. <https://doi.org/10.1038/s41598-017-10861-6>.
27. Kern C, Wang Y, Xu X, et al. Functional annotations of three domestic animal genomes provide vital resources for comparative and agricultural research. *Nat Commun* 2021;12(1):1821. <https://doi.org/10.1038/s41467-021-22100-8>.
28. Fang L, Liu S, Liu M, et al. Functional annotation of the cattle genome through systematic discovery and characterization of chromatin states and butyrate-induced variations. *BMC Biol* 2019;17(1):68. <https://doi.org/10.1186/s12915-019-0687-8>.
29. Baranasic D, Hortenhuber M, Balwiercz PJ, et al. Multiomic atlas with functional stratification and developmental dynamics of zebrafish cis-regulatory elements. *Nat Genet* 2022;54(7):1037–50. <https://doi.org/10.1038/s41588-022-01089-w>.
30. Johnston IA, Kent MP, Boudinot P, et al. Advancing fish breeding in aquaculture through genome functional annotation. *Aquaculture* 2024;583:740589. <https://doi.org/10.1016/j.aquaculture.2024.740589>.
31. van der Velde A, Fan K, Tsuji J, et al. Annotation of chromatin states in 66 complete mouse epigenomes during development. *Commun Biol* 2021;4(1):239. <https://doi.org/10.1038/s42003-021-01756-4>.
32. Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;38(4):576–89. <https://doi.org/10.1016/j.molcel.2010.05.004>.
33. Ali A, Al-Tobasei R, Lourenco D, et al. Genome-wide association study identifies genomic loci affecting filet firmness and protein

- content in rainbow trout. *Front Genet* 2019;10:386. <https://doi.org/10.3389/fgene.2019.00386>.
34. Ali A, Al-Tobasei R, Lourenco D, et al. Genome-wide identification of loci associated with growth in rainbow trout. *BMC Genomics [Electronic Resource]* 2020;21(1):209. <https://doi.org/10.1186/s12864-020-6617-x>.
  35. Salem M, Al-Tobasei R, Ali A, et al. Genome-wide association analysis with a 50 K transcribed gene SNP-chip identifies QTL affecting muscle yield in rainbow trout. *Front Genet* 2018;9:387. <https://doi.org/10.3389/fgene.2018.00387>.
  36. Ali A, Al-Tobasei R, Lourenco D, et al. Genome-wide scan for common variants associated with intramuscular fat and moisture content in rainbow trout. *BMC Genomics [Electronic Resource]* 2020;21(1):529. <https://doi.org/10.1186/s12864-020-06932-0>.
  37. Liu S, Martin KE, Gao G, et al. Identification of haplotypes associated with resistance to bacterial cold water disease in rainbow trout using whole-genome resequencing. *Front Genet* 2022;13:936806. <https://doi.org/10.3389/fgene.2022.936806>.
  38. Vallejo RL, Evenhuis JP, Cheng H, et al. Genome-wide mapping of quantitative trait loci that can be used in marker-assisted selection for resistance to bacterial cold water disease in two commercial rainbow trout breeding populations. *Aquaculture* 2022;560:738574. <https://doi.org/https://doi.org/10.1016/j.aquaculture.2022.738574>.
  39. Lien S, Koop BF, Sandve SR, et al. The Atlantic salmon genome provides insights into rediploidization. *Nature* 2016;533(7602):200–5. <https://doi.org/10.1038/nature17164>.
  40. Breschi A, Munoz-Aguirre M, Wucher V, et al. A limited set of transcriptional programs define major cell types. *Genome Res* 2020;30(7):1047–59. <https://doi.org/10.1101/gr.263186.120>.
  41. Gorkin DU, Barozzi I, Zhao Y, et al. An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* 2020;583(7818):744–51. <https://doi.org/10.1038/s41586-020-2093-3>.
  42. Harrison PW, Amode MR, Austine-Orimoloye O, et al. Ensembl 2024. *Nucleic Acids Res* 2024;52(D1):D891–D9. <https://doi.org/10.1093/nar/gkad1049>.
  43. Salem M, Al-Tobasei R, Ali A, et al. Integrated analyses of DNA methylation and gene expression of rainbow trout muscle under variable ploidy and muscle atrophy conditions. *Genes (Basel)* 2022;13(7). <https://doi.org/10.3390/genes13071151>.
  44. Perez-Rico YA, Boeva V, Mallory AC, et al. Comparative analyses of super-enhancers reveal conserved elements in vertebrate genomes. *Genome Res* 2017;27(2):259–68. <https://doi.org/10.1101/gr.203679.115>.
  45. Hnisz D, Abraham BJ, Lee TI, et al. Super-enhancers in the control of cell identity and disease. *Cell* 2013;155(4):934–47. <https://doi.org/10.1016/j.cell.2013.09.053>.
  46. van Groningen T, Koster J, Valentijn LJ, et al. Neuroblastoma is composed of two super-enhancer-associated differentiation states. *Nat Genet* 2017;49(8):1261–66. <https://doi.org/10.1038/ng.3899>.
  47. Watson ML, Baehr LM, Reichardt HM, et al. A cell-autonomous role for the glucocorticoid receptor in skeletal muscle atrophy induced by systemic glucocorticoid exposure. *Am J Physiol Endocrinol Metab* 2012;302(10):E1210–20. <https://doi.org/10.1152/ajpendo.00512.2011>.
  48. Braun TP, Grossberg AJ, Krasnow SM, et al. Cancer- and endotoxin-induced cachexia require intact glucocorticoid signaling in skeletal muscle. *FASEB J* 2013;27(9):3572–82. <https://doi.org/10.1096/fj.13-230375>.
  49. Yamazaki H, Uehara M, Yoshikawa N, et al. The crucial role of muscle glucocorticoid signaling in accelerating obesity and glucose intolerance via hyperinsulinemia. *JCI Insight* 2023;8(8). <https://doi.org/10.1172/jci.insight.162382>.
  50. AlSudais H, Rajgara R, Saleh A, et al. C/EBPbeta promotes the expression of atrophy-inducing factors by tumours and is a central regulator of cancer cachexia. *J Cachexia Sarcopenia Muscle* 2022;13(1):743–57. <https://doi.org/10.1002/jcsm.12909>.
  51. Quddos F, Zwollo P. A BCWD-resistant line of rainbow trout is less sensitive to cortisol implant-induced changes in IgM response as compared to a susceptible (control) line. *Dev Comp Immunol* 2021;116:103921. <https://doi.org/10.1016/j.dci.2020.103921>.
  52. de Laval B, Maurizio J, Kandalla PK, et al. C/EBPbeta-dependent epigenetic memory induces trained immunity in hematopoietic stem cells. *Cell Stem Cell* 2023;30(1):112. <https://doi.org/10.1016/j.stem.2022.12.005>.
  53. Berthelot C, Brunet F, Chalopin D, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nat Commun* 2014;5(1):3657. <https://doi.org/10.1038/ncomms4657>.
  54. Assis R, Bachtrog D. Neofunctionalization of young duplicate genes in *Drosophila*. *Proc Natl Acad Sci USA* 2013;110(43):17409–14. <https://doi.org/10.1073/pnas.1313759110>.
  55. Zentner GE, Tesar PJ, Scacheri PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res* 2011;21(8):1273–83. <https://doi.org/10.1101/gr.122382.111>.
  56. Papait R, Cattaneo P, Kunderfranco P, et al. Genome-wide analysis of histone marks identifying an epigenetic signature of promoters and enhancers underlying cardiac hypertrophy. *Proc Natl Acad Sci* 2013;110(50):20164–69. <https://doi.org/10.1073/pnas.1315155110>.
  57. Verta J-P, Barton HJ, Pritchard V, et al. Genetic drift dominates genome-wide regulatory evolution following an ancient whole-genome duplication in Atlantic salmon. *Genome Biol Evol* 2021;13(5). <https://doi.org/10.1093/gbe/evab059>.
  58. Halstead MM, Kern C, Saelao P, et al. Systematic alteration of ATAC-seq for profiling open chromatin in cryopreserved nuclei preparations from livestock tissues. *Sci Rep* 2020;10(1):5230. <https://doi.org/10.1038/s41598-020-61678-9>.
  59. Krueger F. Trim Galore: a wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisulfite-Seq) libraries. Accessible date Nov. 13, 2024. [https://www.bioinformatics.braham.ac.uk/projects/trim\\_galore/](https://www.bioinformatics.braham.ac.uk/projects/trim_galore/).
  60. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9(4):357–59. <https://doi.org/10.1038/nmeth.1923>.
  61. Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008;9(9):R137. <https://doi.org/10.1186/gb-2008-9-9-r137>.
  62. Ramirez F, Ryan DP, Gruning B, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016;44(W1):W160–65. <https://doi.org/10.1093/nar/gkw257>.
  63. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 2011;27(11):1571–72. <https://doi.org/10.1093/bioinformatics/btr167>.
  64. Li X, Duan Y, Hao Y. Identification of super enhancer-associated key genes for prognosis of germinal center B-cell type diffuse

- large B-cell lymphoma by integrated analysis. *BMC Med Genomics* 2021;14(1):69. <https://doi.org/10.1186/s12920-021-00916-z>.
65. Sherman BT, Hao M, Qiu J, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res* 2022;50(W1):W216–W21. <https://doi.org/10.1093/nar/gkac194>.
66. Wang Y, Tang H, Debarry JD, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 2012;40(7):e49. <https://doi.org/10.1093/nar/gkr1293>.
67. Wang J, Orlov YL, Li X, et al. In situ dissecting the evolution of gene duplication with different histone modification patterns based on high-throughput data analysis in *Arabidopsis thaliana*. *PeerJ* 2021;9:e10426. <https://doi.org/10.7717/peerj.10426>.
68. Kim D, Paggi JM, Park C, et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 2019;37(8):907–15. <https://doi.org/10.1038/s41587-019-0201-4>.
69. Perry BR, Assis R. CDROM: classification of duplicate gene retention mechanisms. *BMC Evol Biol* 2016;16:82. <https://doi.org/10.1186/s12862-016-0644-x>.
70. Grammarly. 2024. <https://app.grammarly.com>, accession date Nov. 13, 2024.
71. Salem M, Al-Tobasei R, Ali A, et al. Additional files: functional annotation of regulatory elements in rainbow trout uncovers roles of the epigenome in genetic selection and genome evolution. OSF. 2024. <https://doi.org/10.17605/OSF.IO/87GYK>, access ion date Nov. 13, 2024.
72. Salem M, Al-Tobasei R, Ali A, et al. Supporting data for “Functional Annotation of Regulatory Elements in Rainbow Trout Uncovers Roles of the Epigenome in Genetic Selection and Genome Evolution.” *GigaScience Database*. 2024. <https://doi.org/10.5524/102592>.