An, L., Wang, Y., Zhou, H. (2026) Functional validation and predictive modeling of chicken enhancers and regulatory variants using massively parallel reporter assays, *Proceedings 13th World Congress on Genetics Applied to Livestock Production*: 24371
https://doi.org/10.31274/wcgalp.24371 ©2026 *An, et al.* CC BY-NC-ND

Functional validation and predictive modeling of chicken enhancers and regulatory variants using massively parallel reporter assays

L. An[1*], Y. Wang[1], H. Zhou[1]

[1]*University of California, Davis, CA, USA;* [*]*Corresponding author: lqan@ucdavis.edu*

Introduction

Chickens (Gallus gallus) are important domestic animals as they serve as a primary and accessible protein source for human consumption. They are also important model organisms in immunology, developmental biology, evolutionary biology, genomics of domestication, and the epigenetic regulation of gene expression (Brown et al., 2003; Hillier et al., 2004; Burt, 2007; Beacon and Davie, 2021). The functional annotation of regulatory elements in the chicken genome is crucial for understanding the molecular mechanisms underlying economically significant complex traits in chickens, including growth, reproduction, and disease resistance.

The Functional Annotation of Animal Genomes (FAANG) consortium has made great progress in annotating the chicken genome. As part of FAANG consortium, our lab has identified thousands of enhancers across 23 chicken tissues using histone modifications through ChIP–seq and chromatin accessibility through ATAC-seq (Kern et al., 2021; Pan et al., 2023). However, the functional activity of these elements remains largely unverified. ChickenGTEx project has presented a pilot atlas of identified single nucleotide polymorphisms (SNPs) associated with gene expression, most of which are found in non-coding regions of the genome without a clearly defined function(Guan et al., 2025). Due to the correlation of nearby genetic variants (linkage disequilibrium or LD), it remains challenging to distinguish the causal variants. Detecting allelic cis-regulatory effects of these variants will aid in pinpointing the causal variants. However, current computational methods for predicting variant effects rely heavily on conservation or biochemical signatures, and are often trained on human or mouse data, making them less accurate for chickens. To address this gap, we established a Massively Parallel Reporter Assay (MPRA), a high-throughput experiment designed to measure the regulatory activity of short DNA sequences using barcoded expression constructs, in chicken DF-1 cells to systematically assess enhancer activity and allele-specific regulatory effects.

Materials & Methods

Library Design and Construction. We designed a pilot library containing 999 oligonucleotides. This included 340 expression quantitative trait loci (eQTL) variants (represented by both reference and alternative alleles), along with 190 positive controls (conserved enhancers) and 100 negative controls (scrambled sequences). Oligonucleotides (299 bp) were synthesized (Twist Bioscience) containing 269 bp of genomic context and 15 bp of adapter sequences. Following the Tewhey Lab ENCODE protocol (Tewhey et al., 2016), unique 20-bp barcodes were added to the oligonucleotides via PCR using primers containing degenerate sequences. The barcoded library was cloned into the SfiI-digested pGL4.23ΔxbaΔluc vector backbone using Gibson assembly, forming the intermediate library pMPRA:Δorf. To generate the final transfection library (pMPRA:eGFP/minP), pMPRA:Δorf was linearized with AsiSI. The minimal promoter and eGFP reporter gene were amplified from pGL4.23minP-GFP and subsequently inserted via Gibson assembly. The final construct was treated with RecBCD and AsiSI to remove any remaining uncut vector before large-scale transformation.

Transfection and RNA Capture. Chicken DF-1 fibroblast cells were cultured in DMEM supplemented with 10% FBS. Transfection conditions were optimized using a GFP plasmid (PB513B), achieving ~40% efficiency with Lipofectamine 3000. The MPRA plasmid library was transfected into DF-1 cells in triplicate. Cells were harvested 36 hours post-transfection.
Total RNA was extracted from harvested cells using the RNeasy Maxi Kit (Qiagen). To eliminate plasmid contamination, RNA underwent a pre-capture TURBO DNase treatment followed by 5 minutes at 70°C. Specific eGFP transcripts were pulled down using a biotinylated GFP probe mix and Streptavidin C1 Dynabeads. Following capture, a second overnight DNase treatment was performed to ensure the complete removal of residual DNA.

Sequencing and Data Analysis. Enriched RNA was reverse-transcribed using a GFP-specific primer and SuperScript III. cDNA was quantified via qPCR to determine the optimal cycle number for first-round PCR. Libraries were generated via a two-step PCR to capture the barcode region and attach Illumina index adapters.
The pMPRA:Δorf plasmid library was sequenced on an Illumina MiSeq platform to establish barcode/oligo associations. For transcript quantification, libraries were sequenced on an Illumina NovaSeq platform. Barcodes were mapped to oligos, and MPRA activity was quantified as the ratio of normalized mRNA counts to plasmid DNA counts using the MPRAnalyze statistical framework.
We trained gapped k-mer support vector machine (gkm-SVM) models using chromatin features (H3K4me3, H3K27ac, ATAC-seq, H3K4me1, H3K27me3) generated from DF-1 cells via Cleavage Under Targets and Release Using Nuclease (CUT&RUN). Model performance was evaluated using Area Under the Curve (AUC) and Spearman's rank correlation against the experimental MPRA values.

Results

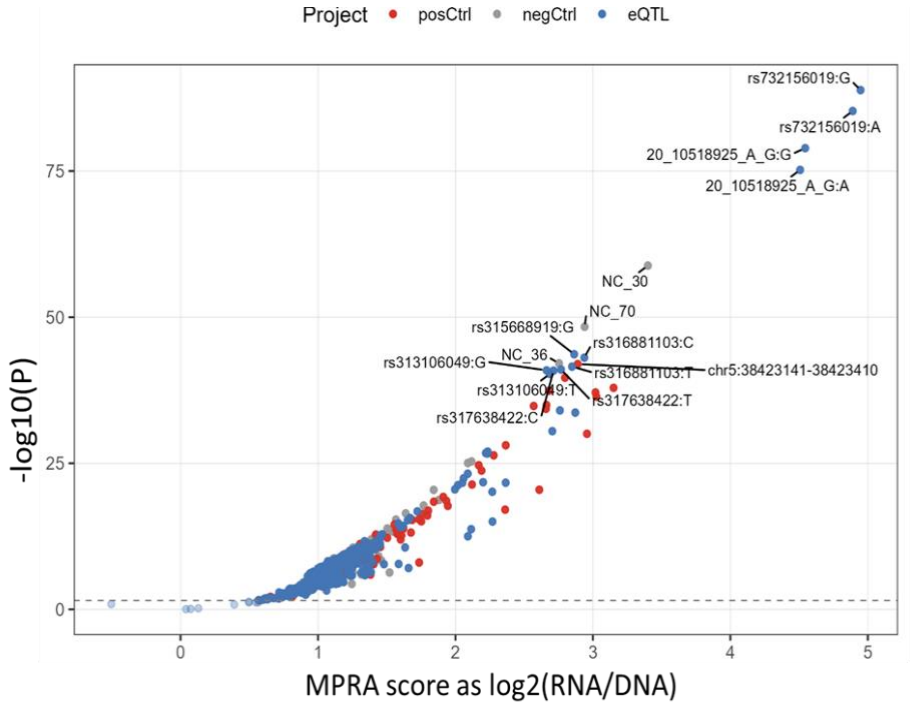

Figure 1. Regulatory activity of Tested sequences in chicken DF-1 cells. Volcano plot displaying the massively parallel reporter assay (MPRA) results for 999 oligonucleotides. X-axis represents the regulatory activity quantified as the log2 ratio of normalized mRNA counts (RNA) to plasmid library counts (DNA). Y-axis shows the statistical significance as -log10(P). Individual data points are color-coded by project category: eQTL variants (blue), positive controls (red), and negative controls (grey). High-confidence regulatory variants are labeled. Dashed horizontal line indicates threshold for statistical significance (P<0.05)

Technical Performance and Functional Characterization. The MPRA system achieved a 99.9% capture rate and high reproducibility between replicates ($r > 0.95$). Sequencing analysis confirmed that 95.9% of the barcodes were correctly matched to their input oligonucleotides. Of the 340 variants tested, 26% showed regulatory activity, with 82% enhancing expression (Fig. 1). Seven variants exhibited significant allelic skew, including rs14718121, whose C allele drove 1.32-fold higher expression (FDR = 3.9×10^{-5}; Fig. 2A). Predicted target genes of rs14718121 were enriched for cell-differentiation and immune-signaling pathways (Fig. 2B), suggesting allele-specific functional impacts on development and immunity.

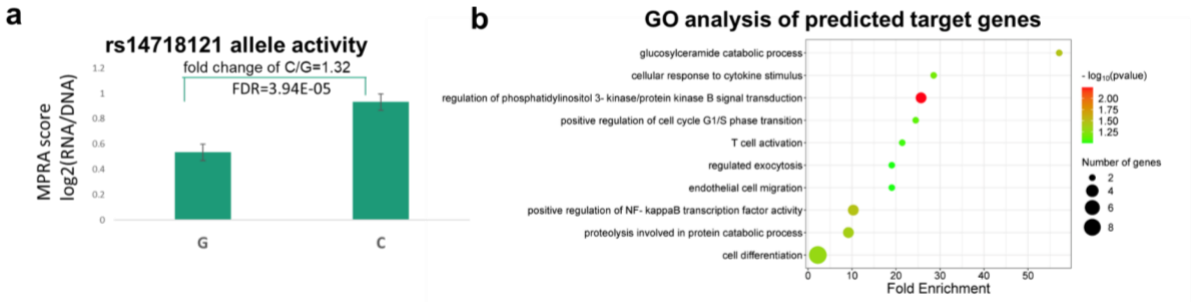

Figure 2. Allelic Effects and Functional Enrichment of Regulatory Variants. (A) Differential Allelic Activity: Representative bar plot demonstrating significant allelic skew for rs14718121. The C allele exhibits a 1.32-fold higher regulatory activity compared to the G allele. Error bars represent standard deviation across three biological replicates. (B) Functional Pathway Analysis: Gene Ontology (GO) enrichment of predicted target genes for active regulatory elements. The x-axis represents fold enrichment, while dot size and color indicate the number of genes and the statistical significance, respectively. Top enriched terms include the PI3K/Akt signaling pathway and T cell activation, highlighting the potential biological impact of identified variants on cellular signaling and immune response in chicken DF-1 cells.

Integrative Predictive Modeling. Integrative modeling via gkm-SVM achieved strong classification accuracy, particularly for H3K4me3 (AUC = 0.88; Fig. 3A). Experimental values showed high concordance with predictions (max $\rho = 0.55$), demonstrating that regulatory potential can be reliably inferred from intrinsic sequence and chromatin state (Fig. 3B-C).

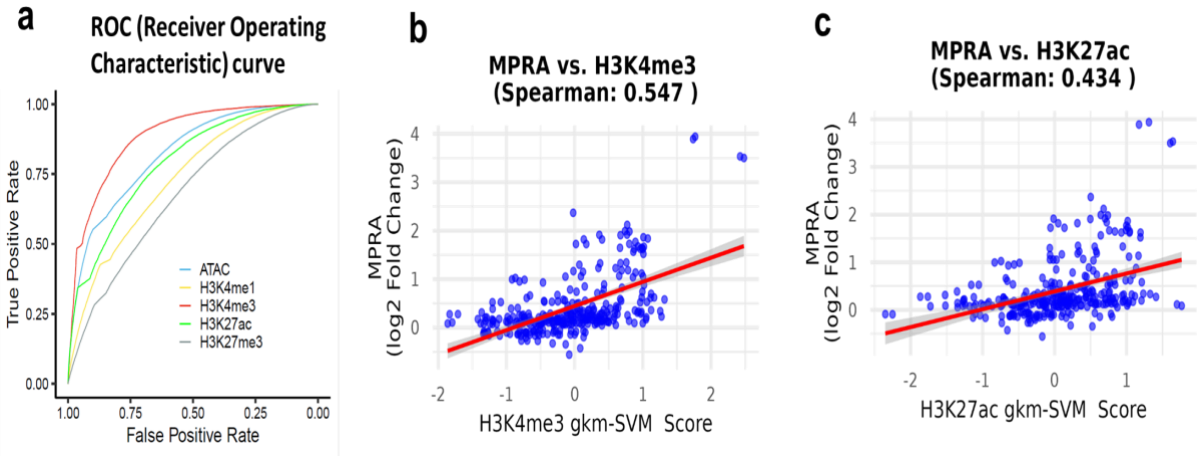

Figure 3. Predictive Performance of gkm-SVM Models. (A) Receiver Operating Characteristic (ROC) Curves: Classification performance of gkm-SVM models trained on various chromatin features (ATAC-seq, H3K4me1, H3K4me3, H3K27ac, H3K27me3). H3K4me3 (red) shows the highest predictive accuracy. (B-C) Correlation Analysis: Spearman correlation between experimental MPRA fold changes and predicted gkm-SVM scores for H3K4me3 ($\rho = 0.547$) and H3K27ac ($\rho = 0.434$), demonstrating that chromatin state accurately reflects measured regulatory potential.

Discussion

We successfully established the first high-throughput MPRA pipeline in chicken cells, overcoming traditional avian transfection challenges. Identifying functional variants like rs14718121 highlights MPRA's power to pinpoint causal mutations masked by linkage disequilibrium. Our results suggest that sequence-based modeling can complement experimental assays to prioritize variants not yet tested. We are currently scaling this platform to screen approximately 15,000 sequences, including FAANG-predicted enhancers across multiple tissues and regulatory variants associated with Marek's disease resistance. This expanded effort will generate a comprehensive functional atlas of the chicken regulome.

Conclusions

Our integrative framework successfully bridges computational prediction and experimental validation. This provides a scalable foundation for a functional atlas of the chicken regulome, and the prioritization of variants associated with disease resistance and production traits.

Acknowledgements

We thank Pardis Sabeti and Ryan Tewhey (The Jackson Laboratory) for the pGL4.23ΔxbaΔluc vector (Tewhey et al., 2016), and Xiaoming Lu and Leah Kottyan (Cincinnati Children's Hospital Medical Center) for the pGL4.23minP-GFP vector (Lu et al., 2021).

References

Beacon T. H., and Davie J. R. (2021). Genome 64, 476–489. doi: 10.1139/gen-2020-0129

Brown W. R. A., Hubbard S. J., Tickle C., and Wilson S. A. (2003). Nat Rev Genet 4, 87–98. doi: 10.1038/nrg998

Burt D. W. (2007). Poultry Science 86, 1460–1471. doi: 10.1093/ps/86.7.1460

Guan D., Bai Z., Zhu X., Zhong C., Hou Y., et al. (2025). Nat Genet 57, 1298–1308. doi: 10.1038/s41588-025-02155-9

Hillier L. W., Miller W., Birney E., Warren W., Hardison R. C., et al. (2004). Nature 432, 695–716. doi: 10.1038/nature03154

Kern C., Wang Y., Xu X., Pan Z., Halstead M., et al. (2021). Nat Commun 12, 1821. doi: 10.1038/s41467-021-22100-8

Lu X., Chen X., Forney C., Donmez O., Miller D., et al. (2021). Nat Commun 12, 1611. doi: 10.1038/s41467-021-21854-5

Pan Z., Wang Y., Wang M., Wang Y., Zhu X., et al. (2023). Science Advances 9, eade1204. doi: 10.1126/sciadv.ade1204

Tewhey R., Kotliar D., Park D. S., Liu B., Winnicki S., et al. (2016). Cell 165, 1519–1529. doi: 10.1016/j.cell.2016.04.027