HUNGARIAN UNIVERSITY OF AGRICULTURE AND LIFE SCIENCES

Doctoral School of Biological Sciences

Integrative Analysis of Allele-Specific Expression and De Novo Mutations: Leveraging a Family-Based Approach to Identify Regulatory Elements Using RNA and Whole-Genome Sequencing

DOI: 10.54598/007160

Doctoral (PhD) dissertation

Maher Alnajjar

Gödöllő

2025

The PhD Program

Name: Doctoral School of Biological Sciences

Discipline: Genomics

Head: Prof. Dr. Nagy Zoltán, DSc.

University Professor
MATE, Hungarian University of Agriculture and Life Sciences
Department of Plant Physiology and Plant Ecology

Supervisor(s): Dr. Barta Endre, Ph.D.

Scientific Advisor
MATE, Hungarian University of Agriculture and Life Sciences
Institute of Genetics and Biotechnology

....

Approval of the Head of Doctoral School Approval of the Supervisor(s)

1. INTRODUCTION AND OBJECTIVES

What drives the notable diversity among individuals, that even those sharing identical DNA, look and function differently? One answer lies in the way genes are expressed and regulated. Understanding genotype-phenotype interactions remains an essential quest for biologists.

Gene expression is a multidimensional product controlled by genetics, epigenetics, and environmental factors. Transcriptional regulation involves interactions between RNA Polymerase (RNAP) and regulatory elements; heterozygous variants in these elements can act as cis-acting factors, causing unequal expression of alleles in a phenomenon known as Allele-Specific Expression (ASE). ASE has drawn much research for its role in discovering rare variants and understanding gene regulation across tissue development. Detecting ASE and developing tools for its discovery have also become objectives in their own right.

The influence of cis-regulatory variants on ASE is evident in altered affinities between transcription factors (TFs) and transcription factor binding sites (TFBS). In RNA quantification, we usually obtain combined allele counts from both parents at each gene. Only when a heterozygous variant is present in the transcript can we infer the ratio of parental alleles. The problem is that not all genes have exonic variants to reliably detect ASE. ASE can also be studied through expression quantitative trait loci (eQTL) experiments and population-scale profiling using genome-wide association studies (GWAS). However, these approaches are limited by high cost, large sample sizes, and the complexity of RNA-seq results.

One way to overcome these issues is an intraspecific F1 hybrid study. By comparing allele expression from parents, the presence of Allelic Imbalance (AI) indicates a cis-regulatory variant. Many studies have used F1 hybrids in farm animals to study ASE, but none have done so in rabbits. Moreover, most focused on the parent-of-origin phenomenon and neglected shared information among progenies, treating each as a separate trio.

Therefore, we conducted our study to accurately characterize ASE genes and their putative variants in TFBS. We used a rabbit family with genetically divergent parents and their eight offspring, combining high-throughput RNA-seq and WGS data for each individual. Our approach relies on gene expression levels supported by Mendelian inheritance-based haplotype phasing of variants in any region. We also show that inheritance-based WGS analysis in a larger family is effective for accurately characterizing de novo mutations (DNMs), which arise in offspring but are absent in parents.

Objectives to achieve:

1. Utilizing rabbits in ASE and DNM studies for the first time.
2. Characterization of ASE as a means of distinguishing between the cis and trans-acting elements in terms of gene regulation by providing a novel pipeline that categorizes gene expression level patterns in a family model.
3. Providing sets of genes as well as Identifying putative regulatory variants within Transcription Factor Binding Sites (TFBSs) that might be related to meat quality and quantity in rabbits and other farm animals.
4. Analyzing DNMs in a larger family model.

2. MATERIALS AND METHODS:

2.1 Samples and Experimental Design

Samples were prepared from muscle tissue (thigh and back) derived from a divergent breeding pair of rabbits (*Oryctolagus cuniculus*). The mother belonged to the Hycole XXL line, and the father is a Thuringer rabbit, and they had eight offspring. The trial took place at a small-scale commercial rabbit farm, and animals were kept under standard livestock production conditions. All procedures were conducted in compliance with ethical guidelines. Prior to collection, animals were euthanized by mechanical stunning and decapitation, complying with Hungarian animal welfare regulations. The experiment and sample collection were carried out at the University of Veterinary in Budapest. The study involves both RNA-seq and Whole Genome Sequencing (WGS) for characterizing Allele-Specific Expression (ASE) and potential variants in cis-regulatory elements. A total of four biological replicates were obtained from each individual. RNA and DNA extraction was performed at the Institute of Genetics and Biotechnology.

2.2 Library Preparation and RNA Sequencing

High-throughput mRNA sequencing analysis was performed on the Illumina platform. RNA sample quality was checked by Agilent BioAnalyzer, and samples with RIN >7 were accepted. RNA-Seq libraries were prepared using an Ultra II RNA Sample Prep kit. Poly-A RNAs were captured by oligo-dT beads, fragmented, and reverse transcribed into cDNA. After adapter ligation and enrichment PCR, sequencing libraries were generated. Sequencing runs were executed on Illumina NextSeq 500 using paired-end 150 cycles.

2.3 Whole Genome Sequencing (WGS)

Family sequencing and library preparation were done by Novogene, following the standard WGS protocol (paired-end, 150 bp read length) on Illumina NovaSeq 6000 to an average depth of 35.4 (± 6.37).

2.4 Quality Control (QC)

QC for raw FASTQ files was achieved using fastQC (Andrews, 2010). For duplication rate check in RNA-seq, the fastp tool was used.

2.5 Workflow Design

The workflow combined RNA-seq and WGS data (Figure 1). Variants (SNVs and INDELs) were called from both RNA-seq and WGS. Variants from WGS are particularly important for noncoding regions. The read counts matrix was obtained from RNA-seq after mapping to the reference genome (OryCun3.0). Pipelines can be found at

https://github.com/Maher199/ASE-in-a-family.

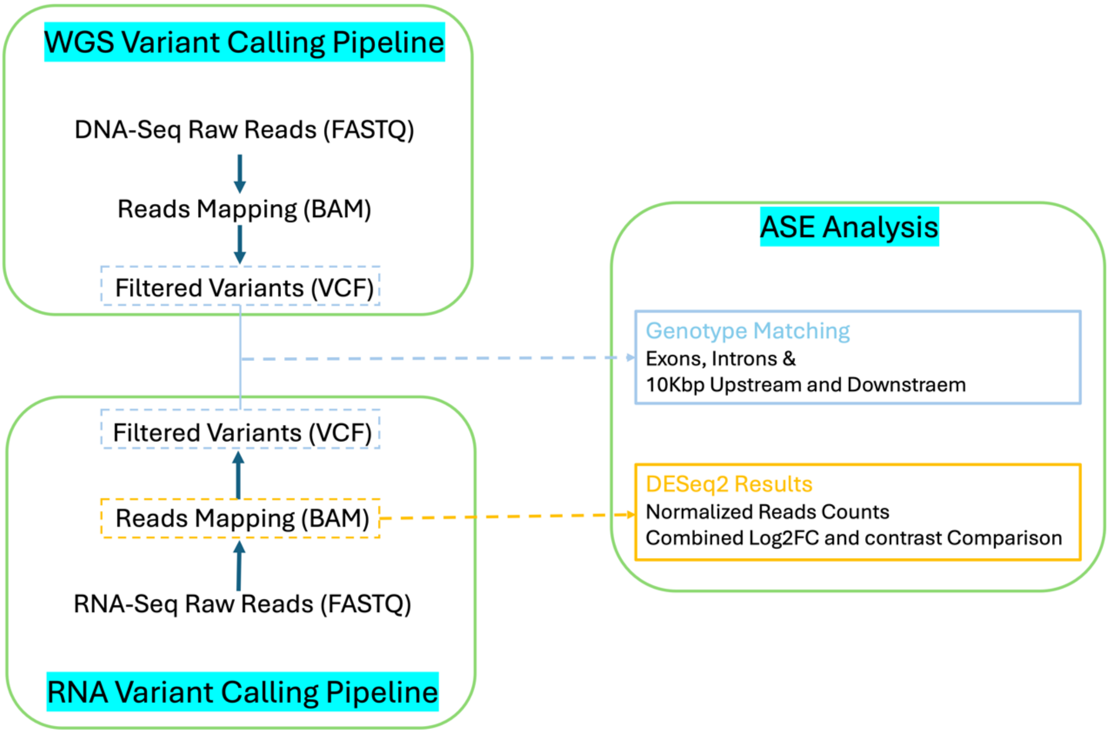
WGS Variant Calling Pipeline
DNA-Seq Raw Reads (FASTQ)
Reads Mapping (BAM)
Filtered Variants (VCF)
ASE Analysis
Genotype Matching
Exons, Introns &
10Kbp Upstream and Downstraem
Filtered Variants (VCF)
Reads Mapping (BAM)
RNA-Seq Raw Reads (FASTQ)
RNA Variant Calling Pipeline
DESeq2 Results
Normalized Reads Counts
Combined Log2FC and contrast Comparison

Figure 1: An overview of the pipeline design, harnessing the benefits of using RNA-seq and WGS in ASE analysis. Variants were called from both types of sequences. Read counts were obtained from RNA-seq BAM files to perform ASE downstream analysis and match the results with the variants' genotypes in coding and non-coding regions.

2.6 ASE Analysis in the Entire Family

ASE analysis was performed in R. The read count matrix from FeatureCounts (after removing mother_2) was used as input to DESeq2 v1.42.0. DESeq2 tests for differential expression using negative binomial generalized linear models, estimates dispersion, calculates log2 fold changes, and adjusts p-values (padj) using the Benjamini-Hochberg method. Genes with fewer than 10 reads in 3 samples were filtered out. DESeq2 was used to estimate contrasts for each individual relative to the father: {mother vs. father}, {offspring_1 vs. father}, … {offspring_8 vs. father}. The resulting log2 fold changes and padj-values were merged into one dataset to examine and compare the entire family. Cases where at least one individual had [|log2FC| > 1] and [padj <0.05] were selected for ASE analysis.

2.7 Pinpointing ASE Cases (Genes)

The selected dataset was used to identify genes that demonstrate ASE. Criteria were set in a Python script (https://github.com/Maher199/ASE-in-a-family). In this study, we refer to High Expression (H), Low Expression (L), and Moderate Expression (M), order (Mother_Father). All comparisons are relative to the father:

- H_L: mother log2FC > 1; offspring span [-0.2, mother + 0.2].
- L_H: mother log2FC < -1; offspring [mother - 0.2, 0.2].
- H_M / M_L: mother log2FC ≥ 0.8; group1 ≥ 0.8, group2 [-0.4, 0.4]; one offspring > 1 with padj < 0.05.
- L_M / M_H: mother log2FC ≤ -0.8; group1 ≤ -0.8, group2 [-0.4, 0.4]; one offspring ≤ -1 with padj < 0.05.

- M_M: mother log2FC [-0.4, 0.4]; offspring can be H, M, or L; two groups not empty; one offspring |log2FC| ≥ 1 with padj < 0.05.

Allele Ratio was also calculated from RNA-seq variants in the VCF to validate the haplotype.

2.8 Haplotype Phasing

Depending on the blocks of Homozygosity uninterrupted by recombination, phase_M.py was built (https://github.com/Maher199/ASE-in-a-family). The program phases the parents' haplotypes in the offspring and reports the separated parents' haplotypes in each child. The condition is to have heterozygous (HET) variants at one parent and homozygous (HOM) at the other; if not, the region should be extended. Additionally, the program outputs a bar plot illustrating the haplotype of each child.

The inputs are the VCF file, chromosome name, and start and end of the region. The program works in these steps:

1. Iterates through the region and separates variants (one parent HET, other HOM) into two tables.
2. Iterates each row, checks grouping: children that are HET are in the first group, HOM in another, and creates a dictionary.
3. The reported grouping is the one with the highest number of occurrences.
4. Optionally, it plots the bar plot.

2.9 De Novo Mutations (DNMs) Discovery

In order to accurately discover novel variants in the offspring, stringent criteria were applied in a Python script (Discover_DNMs.py) (https://github.com/Maher199/Discover_DNMs) on the WGS VCF file:

- A DNM should be a heterozygous locus at the given child.
- Both parents should be homozygous for the reference.
- The variant should not be present in the parents, i.e., no read supporting the alternative allele.
- The variant should be present in the given child only, while the rest of the children should be homozygous for the reference.
- Minimum coverage threshold 15 and maximum 45 to limit false variant calls.
- PL > 450 (normalized likelihoods of genotypes).
- At least 50% of the minimum depth (7–8 reads) should support the alternative allele at the DNM.

3. RESULTS AND DISCUSSION

3.1 Gene expression differences between the two parents

Our primary aim was to detect and explore ASE in rabbit muscle tissue. The father was a Thuringer rabbit, and the mother a meat-producing Hycole breed, which had undergone rigorous selection for meat quantity and quality. We expected significant differences in muscle gene expression between the parents (Fekete et al., 2025).

To examine this, we generated a gene-level count matrix for both parents and conducted differential gene expression (DGE) analysis on iDEP2.0 using DESeq2 (Ge et al., 2018b). We identified 773 differentially expressed genes (DEGs), with 410 upregulated and 363 downregulated in the mother (Figure 2A). Figure 2 B highlights the most significant DEGs in an Enhanced Volcano Plot; ENSOCUG00000021647 had the highest significance. This novel gene is located within an intron of ENSOCUG00000002686 (FAM162A) and corresponds to the human ortholog CORO2B, involved in cytoskeletal organization and signal transduction. Figure 2 C shows the distribution of parent samples by gene expression profiles, suggesting distinct biological expression or phenotypic differences. The full list of DEGs with GO enrichment is available on GitHub.

(https://github.com/Maher199/ASE-in-a-family).

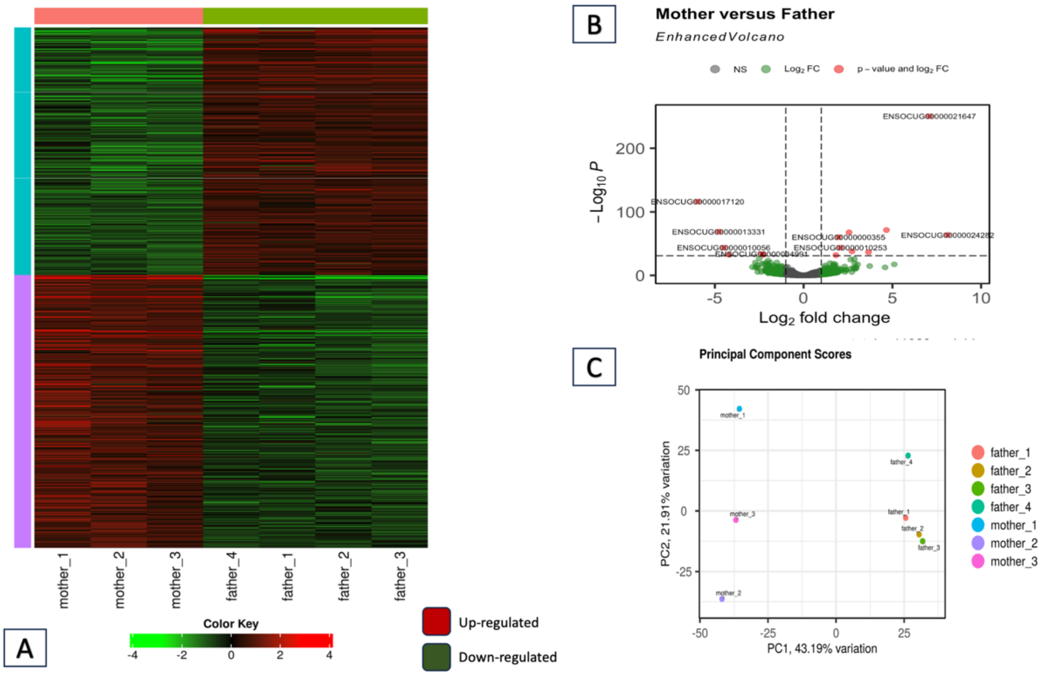

Figure 2: DEGs analysis between the parents. A- Heatmap represents all DEG set between the parents, red: up-regulated, green: down-regulated in the mother. B- Enhanced Volcano plot created in R (Blighe, 2018) and highlights the top DEG, cutoffs used: log2FC=1, p-value=0.001. C- Principal Component Scores for the divergent parents' samples, transformed data with EdgeR (Y. Chen et al., 2025), showing the clustered samples by parent in PCA1 and PCA2.

Using the iDEP 2.0 platform, we conducted GO Enrichment analysis on up- and down-regulated genes. The Gene Ontology Molecular Functions (GOMF) results indicate an up-regulation of several pathways correlated to phenotype differences between the parents, including growth factor binding, Platelet-derived growth factor binding, extracellular matrix components, skeletal system development, glycosaminoglycan binding, collagen binding, and fibronectin binding. On the down-regulation side, the most significant hits are related to carnitine metabolism. Table 1 lists the top GO Enrichment pathways divided as up/down-regulated groups.

Table 1: Top 12 up and top 12 down- regulated Enrichment pathways groups in the mother relative to the father. With a cutoff of 0.1 for the (False Discovery Rate) FDR and Fold enrichment > 1. nGenes: represents the number of genes assigned to each group of (GOMF) pathways for each up/down-regulated set of genes.

group	FDR	nGenes	Pathway size	Fold enriched	Pathway
Upregulated	2.54e-05	15	103	5.69	Glycosaminoglycan binding
Upregulated	4.04e-05	9	35	9.97	Extracellular matrix structural constituent
Upregulated	2.29e-04	11	67	6.21	Heparin binding
Upregulated	1.31e-03	4	5	22.14	2-5-prime-oligoadenylate synthetase activity
Upregulated	1.31e-03	23	1536	2.76	Transmembrane signaling receptor activity
Upregulated	1.32e-03	13	114	4.18	Sulfur compound binding
Upregulated	1.63e-03	27	1652	2.41	Signaling receptor activity
Upregulated	1.63e-03	27	1652	2.41	Molecular transducer activity
Upregulated	4.75e-03	5	15	10.66	Fibronectin binding
Upregulated	6.88e-03	8	47	5.27	Collagen binding
Upregulated	6.88e-03	4	9	13.84	Platelet-derived growth factor binding
Upregulated	7.67e-03	10	87	4.13	Growth factor binding
Downregulated	1.02e-02	3	3	31.96	Carnitine O-palmitoyltransferase activity

Downregulated	1.02e-02	3	3	31.96	O-palmitoyltransferase activity
Downregulated	2.67e-02	3	4	23.97	Carnitine O-acyltransferase activity
Downregulated	7.11e-02	9	141	3.94	Tetrapyrrole binding
Downregulated	7.33e-02	5	40	6.95	O-acyltransferase activity
Downregulated	7.33e-02	2	2	31.96	Sphingolipid floppase activity
Downregulated	7.33e-02	2	2	31.96	Phosphatidylcholine floppase activity
Downregulated	7.33e-02	2	2	31.96	Estrogen 2-hydroxylase activity
Downregulated	7.33e-02	2	2	31.96	Floppase activity
Downregulated	8.10e-02	8	134	3.76	Heme binding
Downregulated	8.34e-02	26	642	1.88	Oxidoreductase activity

3.2 Allele-Specific Expression Characterization

ASE, where parental alleles are expressed at different levels in the offspring, can arise from variation in cis- or trans-acting regulatory elements. It can be detected by eQTL analysis across populations or by hybrid-based studies, such as crosses between genetically distinct breeds, where ASE is assessed in F_1 offspring. We designed our pipeline based on the hypothesis that a heterozygous locus in a cis-regulatory element, where a transcription factor binds, is necessary to induce ASE. Unlike conventional approaches, our setup can detect ASE in a family without requiring heterozygous sites in the children's transcripts, though any heterozygous variant can serve as validation.

3.3 Experiment Design Overview

The family-model approach consists of the parents and eight offspring. Considering gene expression as the phenotype, we matched it with variant genotypes in the hypothetical driving TFBS. Based on intermediate inheritance, we classified expression levels into High (H), Low (L), and Moderate (M). Figure 3 illustrates these possible matching scenarios.

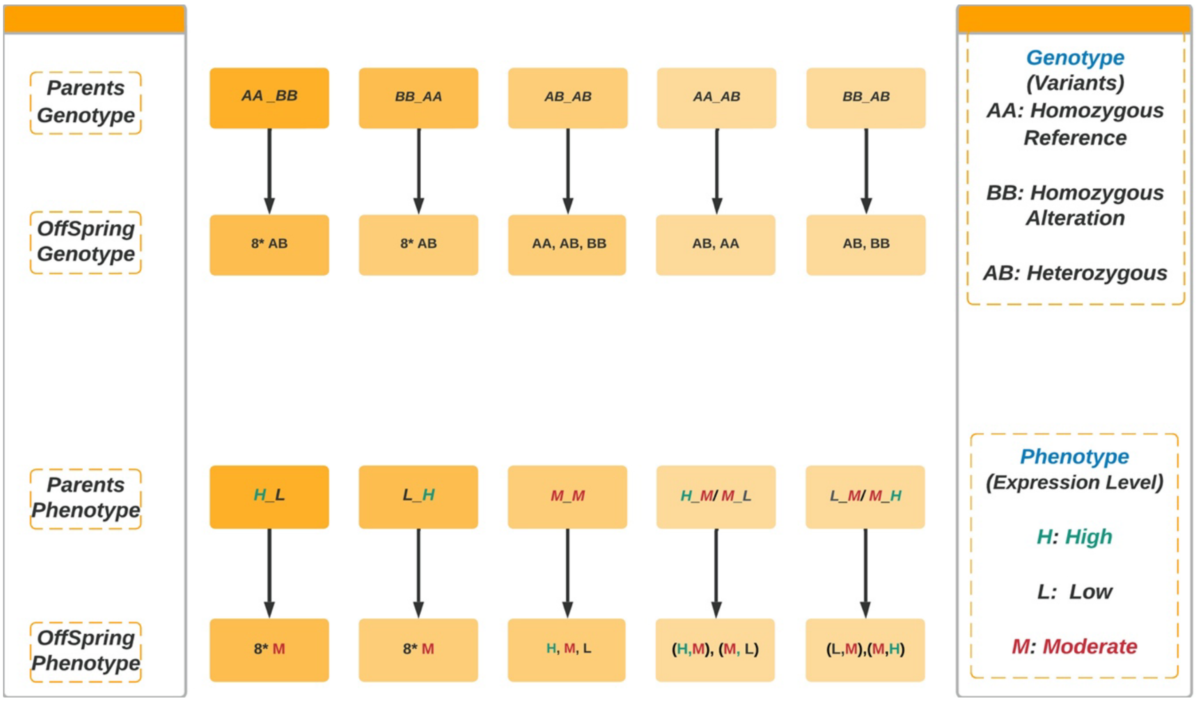

Figure 3: Possible combination of Genotype and Phenotype inherited patterns in accordance with of the Mendelian Laws. Color-coded as the expression level (H: green, L: black, and M: red)

3.4 Phenotype Patterns Prediction

In the simplest case, where one allele is responsible for the higher expression level and the other allele is responsible for a lower expression level, we can assume that the individuals with a heterozygous locus will show an intermediate expression. As in figures 4 and 5, we refer to the high expression with the letter ‘H’, the low expression with the letter ‘L’, and the moderate with the letter ‘M’ (an intermediate between the high and low levels). Accordingly, we hypothesized seven possible parents’ expression level combinations at the different genes, which are the following: H_L, L_H, H_M, M_H, L_M, M_L, and M_M. In every case, the first letter represents the mother.

We developed a script that uses the expression values of the two parents and eight children (Log2FC relative to the father) at each gene to predict the presence and type of ASE (Figure 4). After applying a threshold based on experimental and visual checks, the analysis identified 97 H_L, 110 L_H, 469 M_M, 119 (H_M or M_L), and 133 (L_M or M_H) genes (Table 2). The plots, scripts, normalized read counts, and predicted ASE types are available in the repository (https://github.com/Maher199/ASE-in-a-family). Notably, most cases fall into the M_M group, while only eleven cases include children across all three expression categories (H, M, and L). M_M cases represent the heterozygosity in both parents and that can only be detected in the segregation of the children. In these cases, we hypothesize that both parents are heterozygous in a regulatory site; therefore, the children will segregate due to this variation. Ideally, among the eight children, two L, four M, and two H cases should be observed, but, of course, the random inheritance of the parental alleles can result in other ratios.

Table 2: ASE Cases Summary: A summary table of ASE genes, including the numbers and percentage of evident variants matched at each level of the analysis. For each expression case (H_L, L_H, M_M, H_M or M_L, and L_M or M_H), the table summarizes the number of genes that were found to have variants and the number of matched variants identified at the corresponding level of evidence analysis: RNA variants, DNA in the Exonic region, DNA in the Intronic region and DNA in the surrounding 10Kb regions upstream and downstream. The last row shows the number and percentage of the sum of the unique genes after accounting for the overlaps.

Cases	**H_L**			**L_H**			**M_M**			**H_M or M_L**			**L_M or M_H**		
No.Genes	**97**			**104**			**469**			**110**			**133**		
	All	matched	%	All	matched	%	All	matched	%	All	matched	%	All	matched	%
RNA	65	27	41.5	72	35	48.6	341	1	0.3	72	3	4.2	94	8	8.5
DNA_EXON	81	34	42	88	45	51.1	374	1	0.3	89	6	6.7	105	14	13.3
DNA_INTRON	84	42	50	97	60	61.9	423	16	3.8	103	14	13.6	116	24	20.7
DNA_Outside	97	49	50.5	103	61	59.2	462	13	2.8	110	17	15.5	132	21	15.9
No. of Unique Genes	97	55	**56.7**	104	65	**62.5**	467	25	**7.3**	110	21	**29.2**	132	31	**33**

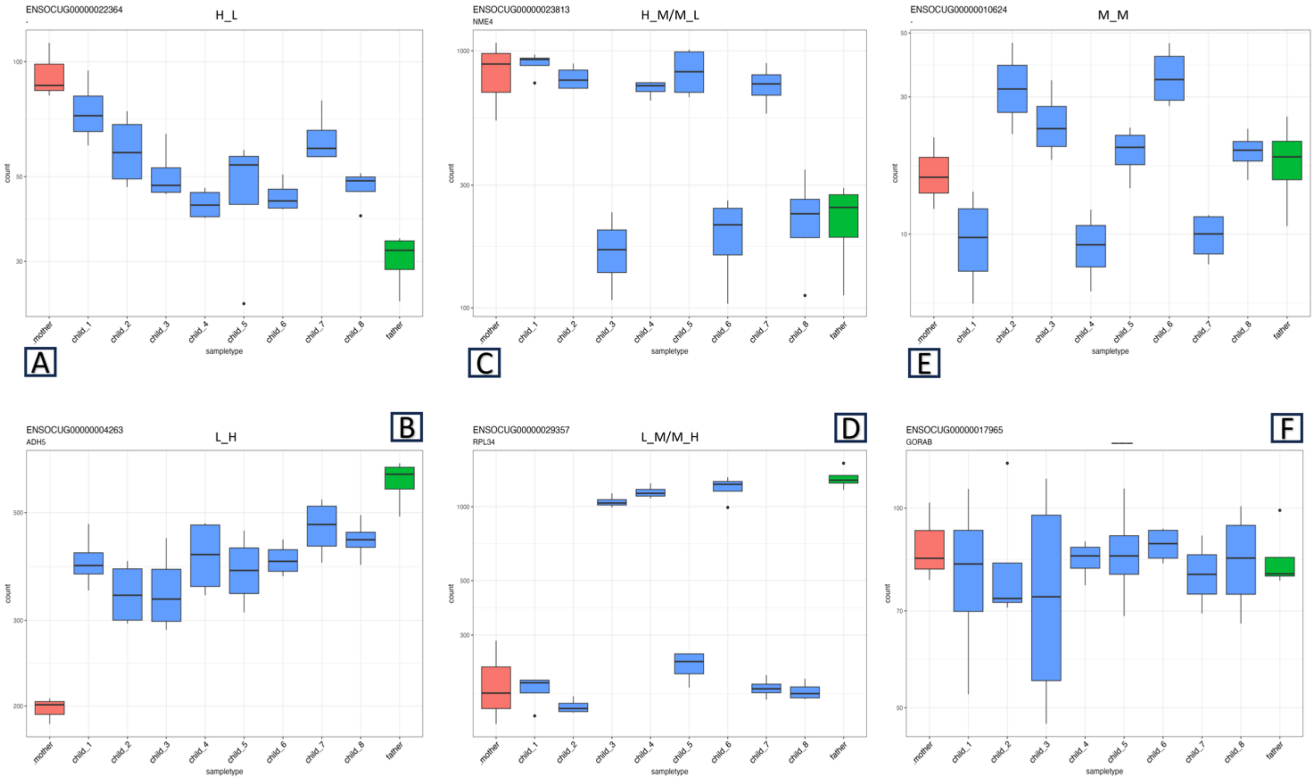

Figure 41: Inheritance Patterns of Gene Expression Across the Family. Cases examples of possible gene expression in our family model. The X-axis lists the family members, and the Y-axis shows the RNA-seq normalized read counts by DESeq2. A & B: (H_L) and (L_H) cases respectively, where the parents are either High (H) or Low (L) expressing level, and all children represent the Moderate (M) expression level demonstrating the intermediate inheritance. C & D: (H_M/M_L) and (L_M/M_H) cases where the children split into two groups each matching the parent's gene expression level. E- (M_M) case demonstrates the hidden intermediate inheritance where both parents are Moderately expressed, while the children have H, L, and M-expressing levels. F- No Allele-Specific Expression. Genes with no significant differences among any individuals in the family.

3.5 Validating predicted phenotypes by the conventional approach

Our described approach analysis is based only on considering expression levels at each gene by comparing the log2FC relative to one parent (the father). However, in the traditional approach involving crossing two breeds where RNA-seq is available for the entire family, ASE can be detected only if at least one heterozygous site is found in the transcript. To demonstrate the feasibility of our approach and its advantage, we carried out the traditional method to ensure that our pipeline includes all such cases plus others lacking variants in the transcripts. We counted allele-specific reads at heterozygous transcript variants in the eight children and compared them to our expression-based predictions. After variant calling from RNA-seq, we checked the allele ratio at each heterozygous variant in ASE genes discovered by our method and considered only variants conforming to the family's expression patterns. Not all genes have variants in RNA-seq—about 19% of expressed genes (2440/12659). We found 42% of genes in H_L, 51% in L_H, 13% in L_M or M_H, 6.7% in H_M or M_L, and 0.3% in M_M to be reliable and conforming. The full dataset available at (https://github.com/Maher199/ASE-in-a-family).

3.6 Identification of regulatory variants:

The availability of whole genome sequences allowed us to test the hypothesis that ASE always presumes a variation in a regulatory region. To find such regulatory SNPs (rSNPs), it is essential to predict TFBSs in the rabbit genome. Since it lacks a comprehensive ChIP-seq-based TFBS mapping, we utilized our human ChIPSummitDB database (Czipa et al., 2020). First, the human hg19 genome was aligned to the rabbit OryCun3.0 reference genome. Next, using the chain files and the human ChIP-seq-based consensus transcription factor binding site collection, TFBSs conserved between human and rabbit were determined. Finally, the ASE-predicted genes and their potential regulatory elements (intronic and 10kb surrounding regions) were searched for variants in the predicted TFBSs. The entire dataset for TFBS is available at (https://github.com/Maher199/ASE-in-a-family). We found 222 conserved TFBSs altogether at 90 genes. They all contain a variation with an identical inheritance pattern as predicted at the given ASE gene.

Figure 5 shows an example of a variant found in the intronic region, that matches the expression pattern and overlaps with a conserved TFBS. Based on the gene expression data, the ZACN gene was predicted as an (L_M or M_H) ASE pattern. In the first intron of this gene, a conserved Esrra binding site was identified. This means that in a human ChIP-seq experiment using an Esrra antibody, a peak was observed at a homologous position that contains an Esrra binding site. The rabbit OryCun3.0 reference genome contains an AGGTCgcGGTCA (capital letters match the consensus) site in a conserved position. AGGTCA is the consensus binding site for a nuclear hormone receptor, and it is not a complete DR0 nuclear hormone receptor dimer binding site. From panel C in figure 5, the mother is homozygous reference (HOM_REF) at this variant, while the father is heterozygous (HET). Among the children, only children 3, 6, and 8 remain HOM, which perfectly matches the ASE pattern.

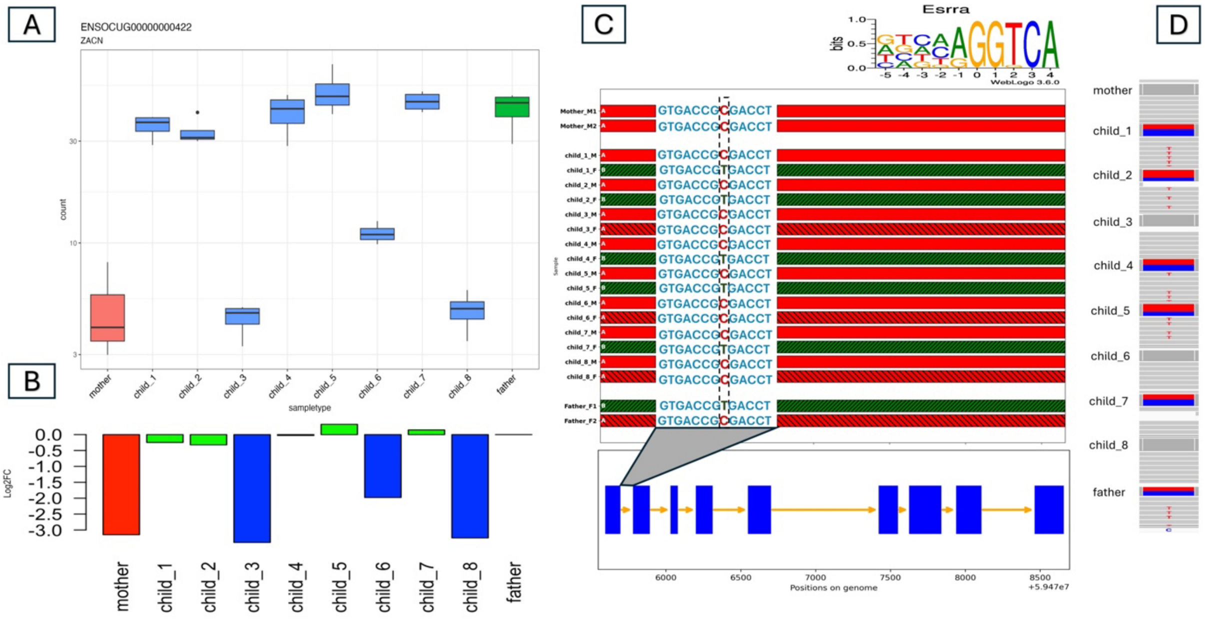

Figure 5: ZACN Gene Analysis Demonstrating an Example of L_M Expression Pattern Across Family Members. A: Normalized RNA-seq read counts with color-coded members - mother (red), father (green) and children (blue). The mother and three children (3,6, and 8) exhibit Low (L) expressions, while the father and the remaining children have Moderate (M) expressions. B: Log2 Fold Change (Log2FC) relative to father. The L-expressing individuals display Log2FC of < -0.8, whereas the M-expressing individuals have Log2FC close to zero as shown in the Y-axis. C: Haplotype phasing of the ZACN gene across the family with colored haplotype panels based on the parental imputation. The Y-axis lists the family members, and the X-axis expands the gene region position in the genome. A conserved TFBS for Esrra is highlighted in the first intron with the variant C/T in the M-expressing individuals. The consensus sequence from SummitDB for Esrra is shown above the plot. D: IGV view of the variant in the conserved TFBS. The mother and the L-expressing offspring are homozygous (C/C), while the father and the M-expressing children are heterozygous (C/T).

Most key ASE findings rely on overlapping reads on a variation in a transcribed region, usually using RNA-seq from F_1 offspring after crossing two genetically distant parents. However, this requires at least one heterozygous site in the exonic region, and not all genes have exonic variants; we found about 19% of expressed genes lack any reliable variant in RNA-seq.

Expression quantitative trait loci (eQTL) mapping is another approach, but rabbits have not previously been used in ASE studies. We selected this species because the two breeds under investigation are genetically divergent and harbor a relatively high number of heterozygous variants.

In this work, we demonstrated that a family-based approach can detect ASE without pre-existing variants in transcribed regions and predict putative cis-regulatory variants without resequencing many individuals. We developed a novel pipeline combining WGS and RNA-seq, relying on gene expression supported by Mendelian inheritance-based haplotype phasing of variants. Gene expression can be confirmed by variants either transcribed or in non-coding regions.

Using ChIPSummitDB, we determined rabbit-conserved TFBSs. Cis-element effects were hypothesized when a variant exhibits the same inheritance pattern as a nearby ASE gene. Seven theoretical expression combinations exist (L, H, M), but only six are observable. Simple cases are H_L or L_H, where offspring show intermediate expression, suggesting cis regulation. M_M, where parents have similar expression, can produce diverse offspring levels, revealing subtle functional impacts.

Our family-based setup improves accuracy in identifying cis-elements without sequencing many individuals. On average, 21% of ASE genes may be regulated by cis-elements, while the remaining 79% likely involve trans-regulatory mechanisms or complex multi-factor regulation. H_L and L_H categories showed the highest cis-element match rates (56% and 62%), while M_M genes had the lowest (7.3%). Among 773 differentially expressed genes between parents, 307 exhibit ASE.

3.7 De Novo Mutation Discovery and Filtration

The experiment setup consists of two genetically divergent parents (mother Hycole, father Thuringer) with their eight offspring. WGS for each individual helped us investigate de novo mutations (DNMs).

To discover DNMs, the variant should be present only in the child and not in the parents. To avoid false positives from sequencing errors, the variant should also be absent in other siblings. We developed an algorithm with strict filtering in a Python script (https://github.com/Maher199/Discover_DNMs) for detecting DNMs in trios or larger families. The script classifies DNMs into SNVs and INDELs.

To understand the influence of the number of children, we created combinations of all possible trios, gradually adding siblings, and ran the script. Figure 6 illustrates the decline in DNMs after adding more siblings, especially after the 3rd sibling. Initially, treating each child as an only child (trio) reported ~4650 DNMs per child. Adding more siblings reduced false positives to ~170 DNMs per child.

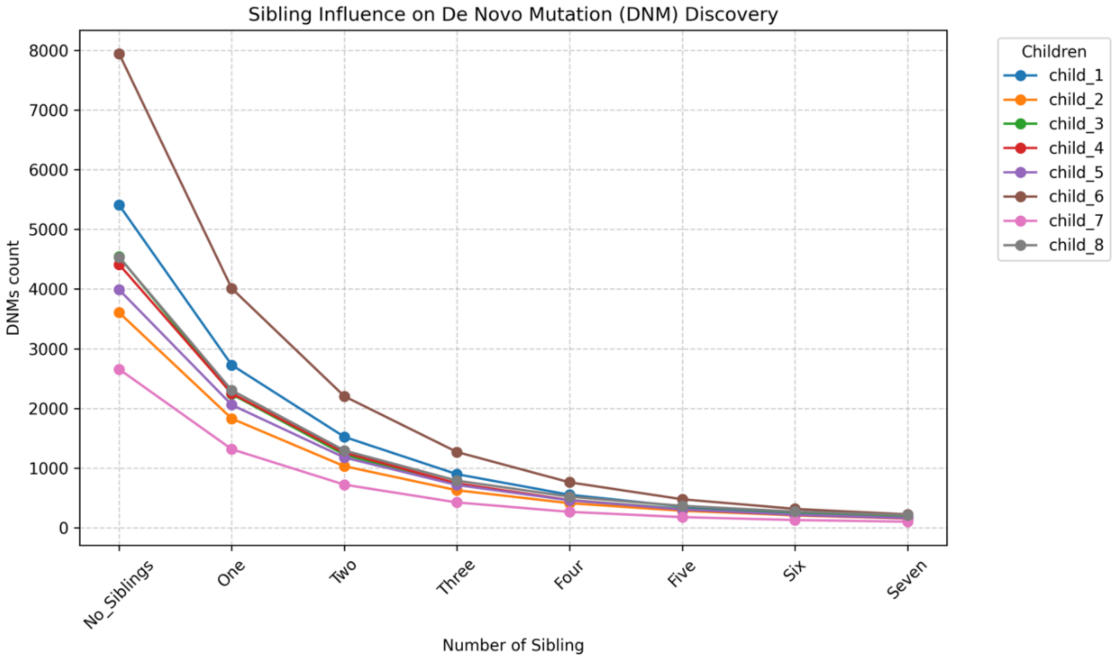

Figure 6 The impact of a larger family size on the number of reported DNMs. X-axis indicates the number of siblings starting with no sibling (i.e. trio family), while Y-axis refers to the putative DNMs reported. Lines colored by children.

We demonstrate the power of utilizing a larger family in the purpose of DNMs identification from the Whole Genome Sequences (WGS). In special cases, like DNM detection, a high level of fidelity is crucial, and a special case should be considered when looking for new mutations. For this aim, we developed a script (Discover_DNMs.py https://github.com/Maher199/Discover_DNMs) that consists of a set of stringent criteria in order to preserve the accurate variants for downstream analysis. The script has flexible options and can be executed on different family sizes beginning with trio families, while maintaining a computationally feasible runtime. To the best of our knowledge, rabbits were not used in a DNMs study. The script was performed on a large rabbit family consisting of the parents and 8 offspring. The results show a massive drop of the number of DNMs discovered in case we deal with each child alone without considering its sibling.

In summary, we have developed a Python script with flexible options to pinpoint the putative DNMs after a stringent filtering, demonstrating the impact of a larger family on the analysis. We are convinced that utilizing longer reads technology will significantly increase the accuracy especially in the repetitive regions, also more research is needed in this area to analyze DNMs hotspots and their consequences.

4. CONCLUSIONS AND RECOMMENDATIONS

Allele-Specific Expression (ASE) occurs when maternal and paternal gene copies are unequally expressed, implying differential regulatory control, often due to cis-acting variants, although trans effects can also contribute. ASE helps differentiate cis from trans effects and detect rare variants influencing disease and trait phenotypes. ASE has been reported in farm animals affecting muscle growth, meat quality, and production.

Most ASE findings rely on counting overlapping reads at a transcribed variant (Lin et al., 2023; Quan et al., 2024). This requires at least one heterozygous site, but ~19% of expressed genes (2440/12,659) lack reliable exonic variants. Here, we used a family-based approach to detect ASE without exonic variants by comparing mRNA expression across family members.

eQTL studies link regulatory DNA variation to gene expression and complex traits (Renganaath & Albert, 2023) but require many individuals. Our method identifies putative causal regulatory variants in TFBS without resequencing large populations.

We developed a pipeline that matches gene expression patterns across a family with genotype patterns in TFBS, filtering out false variants. On average, 21% of predicted ASE genes were potentially cis-regulated, suggesting 79% are influenced by trans factors or complex interactions, consistent with mouse studies (12–24%).

Applied to a hybrid rabbit family, we detected 913 ASE genes, categorized by expression patterns. Using whole-genome alignment between human H19 and rabbit OryCun 3.0, conserved TFBSs from ChIP-SummitDB were mapped, intersected with variants near ASE genes, and 222 TFBSs potentially regulated 90 ASE genes.

This is the first study to explore ASE and DEGs in rabbits, identifying 773 DEGs between divergent parents, 307 of which exhibit ASE. These include genes relevant to meat production, some novel and some reported in pigs and other farm animals.

The pipeline can be improved using long-read sequencing, reciprocal crosses, and machine learning models integrating multiple variant effects. TFBS variants reported here provide a valuable resource for studying cis-regulation of muscle development in rabbits.

De novo mutations (DNMs) were also studied in rabbits for the first time. We developed a Python script showing that extended family data improves DNM detection. Long-read sequencing will further enhance accuracy, especially in repetitive regions. On average, 135 DNMs per child were identified in the eight-offspring family, though further research is needed to understand hotspots and functional impacts.

5. NEW SCIENTIFIC RESULTS

1. Utilizing rabbits to characterize ASE and DNM analysis for the first time.
2. Extending the definition of ASE by providing a novel pipeline for discovering ASE in a hybrid family without the necessity of a heterozygous variant to be present in the transcript and without involving a large number of samples/individuals.
3. Providing a dataset of potential TFBS that can be utilized in meat quality and quantity in farm animals.
4. Developed and reported haplotype phasing software. The software resolves the parental haplotypes in the children in a given region.
5. Conducting DEGs in rabbits between two divergent breeds and reporting genes that might be related to meat quality and production.
6. Providing a Python program that pinpoints DNM from a nuclear family with any size.
7. Reporting DNMs hotspots.

6. LIST OF PUBLICATIONS

Fekete, Z., Német, Z., Ninausz, N., Fehér, P., Schiller, **M., Alnajjar**, M., Szenes, Á., Nagy, T., Stéger, V., Kontra, L., & Barta, E. (2025). **Whole-Genome Sequencing-Based Population Genetic Analysis of Wild and Domestic Rabbit Breeds**. *Animals*, *15*(6), 775. **https://doi.org/10.3390/ani15060775**

Maher Alnajjar, Zsófia Fekete, Tibor Nagy et al. **A New Family-based Approach for Detecting Allele-Specific Expression and for Mapping Possible eQTLs**, October 2025, Animals, available at Research Square. **https://doi.org/10.21203/rs.3.rs-6515696/v1**

T. Pintér, M. Urbán, M. Alnajjar, E. Barta, O.I. Hoffmann, Z. Szőke, E. Gócza, L. Hiripi, L. Bodrogi, **P25-03** Exploration of the novel role of NADPH oxidases in the biotransformation of T2 mycotoxin,Toxicology Letters, Volume 399, Supplement 2, 2024, Pages S353-S354, ISSN 0378-4274, https://doi.org/10.1016/j.toxlet.2024.07.841.
(https://www.sciencedirect.com/science/article/pii/S0378427424019209)
Presentations:

Maher Alnajjar, Levente Kontra, Zsófia Fekete, Tibor Nagy, Mátyás Schiller, Nóra Ninausz, Viktor Stéger, Zoltán Német, Endre Barta: Meat RNA-seq analysis of a rabbit family with 10 members reveals gene expression differences between the hobby and meat-producing parents. **MBK napok**: National Center of Agriculture and Innovation (NAIK), Gödöllő, Hungary. 20.11.2020

Maher Alnajjar, Levente Kontra, Zsófia Fekete, Tibor Nagy, Mátyás Schiller, Nóra Ninausz, Viktor Stéger, Zoltán Német, Endre Barta: Meat RNA-seq analysis of a rabbit family with 10 members reveals gene expression differences between the hobby and meat-producing parents. **Molecular, cell and immune biology, winter symposium**, Debrecen University, Hungary, 08.01.2021

Maher Alnajjar, Levente Kontra, Zsófia Fekete, Tibor Nagy, Mátyás Schiller, Nóra Ninausz, Viktor Stéger, Zoltán Német, Endre Barta. Combined RNA-seq and WGS analysis of a rabbit family reveals clear genotype supported intermediate inheritance of gene expression levels at many genes: **Hungarian Molecular Biology Conference**, Eger, Hungary, 05-07.11.2021.

Maher Alnajjar, Levente Kontra, Zsófia Fekete, Tibor Nagy, Mátyás Schiller, Nóra Ninausz, Viktor Stéger, Zoltán Német, Endre Barta. Whole genome sequencing of a big family allows determining the haplotypes of the parents' chromosomes, the crossing over positions, the de novo mutations and the exact CNVs. **FIBOK2022**, Gödöllő, Hungary, 11-12.04.2022.

Maher Alnajjar, Tibor Nagy, Levente Kontra, Zsófia Fekete, Nóra Ninausz, Viktor Stéger, Zoltán Német, Endre Barta. A family based whole genome genotyping approach for identifying recombinations, haplotypes and denovo mutations. GBI Nap, Gödöllő, Hungary. 18.11.2022.

Maher Alnajjar, Levente Kontra, Zsófia Fekete, Tibor Nagy, Nóra Ninausz, Viktor Stéger, Zoltán Német, Endre Barta. A Large FAMILY-BASED APPROACH FOR Determining THE Parents' HAPLOTYPES, Recombination events, De Novo mutations, and exact CNVs. **Population Genomics EMBO practical course**. Procida, Naples, Italy. 13-19.03.2023.

Maher Alnajjar, Levente Kontra, Zsófia Fekete, Tibor Nagy, Mátyás Schiller, Nóra Ninausz, Viktor Stéger, Zoltán Német, **Endre Barta.** Combined RNA-seq and WGS Analysis of a Rabbit Family Reveals a Clear Genotype Supporting Intermediate Inheritance of Gene Expression Levels at Many Genes. **The 47th FEBS Congress**, Tours, France. 08-12.07.2023.

References

Andrews, S. (2010). *FASTQC. A quality control tool for high throughput sequence data*.

Blighe, K. S. R. and M. L. (2018). *EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling*.

Chen, Y., Chen, L., Lun, A. T. L., Baldoni, P. L., & Smyth, G. K. (2025). edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Research*, *53*(2). https://doi.org/10.1093/nar/gkaf018

Czipa, E., Schiller, M., Nagy, T., Kontra, L., Steiner, L., Koller, J., Pálné-Szén, O., & Barta, E. (2020). ChIPSummitDB: A ChIP-seq-based database of human transcription factor binding sites and the topological arrangements of the proteins bound to them. *Database*, *2020*. https://doi.org/10.1093/database/baz141

Fekete, Z., Német, Z., Ninausz, N., Fehér, P., Schiller, M., Alnajjar, M., Szenes, Á., Nagy, T., Stéger, V., Kontra, L., & Barta, E. (2025). Whole-Genome Sequencing-Based Population Genetic Analysis of Wild and Domestic Rabbit Breeds. *Animals*, *15*(6), 775. https://doi.org/10.3390/ani15060775

Ge, S. X., Son, E. W., & Yao, R. (2018). iDEP: An integrated web application for differential expression and pathway analysis of RNA-Seq data. BMC Bioinformatics, 19(1), 1–24. https://doi.org/10.1186/s12859-018-2486-6

Lin, Y., Li, J., Chen, L., Bai, J., Zhang, J., Wang, Y., Liu, P., Long, K., Ge, L., Jin, L., Gu, Y., & Li, M. (2023). Allele-specific regulatory effects on the pig transcriptome. *GigaScience*, *12*. https://doi.org/10.1093/gigascience/giad076

Quan, J., Yang, M., Wang, X., Cai, G., Ding, R., Zhuang, Z., Zhou, S., Tan, S., Ruan, D., Wu, J., Zheng, E., Zhang, Z., Liu, L., Meng, F., Wu, J., Xu, C., Qiu, Y., Wang, S., Lin, M., … Wu, Z. (2024). Multi-omic characterization of allele-specific regulatory variation in hybrid pigs. *Nature Communications*, *15*(1). https://doi.org/10.1038/s41467-024-49923-5

Renganaath, K., & Albert, F. W. (2023). Trans *-eQTL hotspots shape complex traits by modulating cellular states*. https://doi.org/10.1101/2023.11.14.567054