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 School

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.

University Professor / 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)

(NB: The doctoral dissertation has to be submitted in four copies bound with the original signatures and the ten theses with the photocopied signatures to the Office of the PhD School concerned.

CONTENTS

LIST OF ABBREVIATIONS

Abbreviation	Full Term
AI	Allelic Imbalance
ALT	Alternative allele
ASE	Allele-Specific Expression
ChIP-Seq	Chromatin Immunoprecipitation Sequencing
Cis-elements	Cis-regulatory Elements
DEG	Differentially Expressed Gene
DNA-Seq	DNA Sequencing
DNM	De Novo Mutation
eQTL	Expression Quantitative Trait Loci
FDR	False Discovery Rate
GATK	Genome Analysis Toolkit
GFF	General Feature Format
GOMF	Gene Ontology Molecular Function
GTF	General Transfer Format
GWAS	Genome-Wide Association Studies
HET	Heterozygous
HOM	Homozygous
INDEL	Insertion/Deletion
mRNA	Messenger RNA
REF	Reference Allele
RNA	Ribo Nucleic Acid
RNAP	RNA Polymerase
RNA-Seq	RNA Sequencing
rSNP	Regulatory Single Nucleotide Polymorphism
SAM	Sequence Alignment Map
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeats
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TSS	Transcription Starting Site
WGS	Whole Genome Sequencing

1 INTRODUCTION

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 the genotype-phenotype interactions remains an essential quest for biologists. Gene expression itself is a multidimensional product that is controlled and regulated by genetics, epigenetics and environmental factors. Gene expression is also a multi-level regulated event, and regulation on the transcriptional level plays a great role in it. Transcriptional regulation involves interaction between the RNA Polymerase (RNAP) and regulatory elements; heterozygous variants in these regulatory elements can function as cis-acting elements causing alleles in diploids to be expressed unequally in a phenomenon that is known as Allele-Specific Expression (ASE). ASE phenomenon has attracted much research in for its application in the discovery of rare variants, and the impact of the variation on gene regulation across tissue development. Moreover, detecting this phenomenon itself and developing tools for its discovery have become objectives in their own right.

The influence of the variants in the cis-regulatory elements on the ASE phenomenon can be seen clearly by the alteration of the affinity between the transcription factors (TFs) and the transcription factors binding sites (TFBS). Therefore, to fully understand this influence, we should quantify the gene expression level, (for example) using mRNA analysis. However, in RNA quantification, we typically obtain the sum of the father's and the mother's allele counts at each gene. Only when a heterozygous variant is present in the transcript, we can infer the ratio for each parental allele. The problem with this approach is that not all genes have exonic variants that can be relied on to detect ASE.

ASE can also be studied by conducting expression quantitative trait loci (eQTL) experiments, and by conducting a population scale profiling utilizing data from genome-wide association studies (GWAS) collected based on two phenotypically different cohorts. This approach helps to pinpoint the genetic regions connected to the given phenotype. However, it is limited due to its high cost, higher number of individuals to be involved in the study, and the complexity of the RNA-seq (RNA Sequencing) results among other factors.

One way to tackle these issues is by conducting an intraspecific F1 hybrid study. By comparing the allele expression originated from the parents, then the presence of the Allelic Imbalance (AI) indicates a variant in cis-regulatory elements. Many studies have already utilized the F1 hybrid in many farm animals to study ASE; however, none has been conducted in rabbits. Besides, most of these studies were oriented to study the parent-of-origin phenomenon which is a special case of ASE. Moreover, these studies neglected the shared information among the progenies, which

can lead to important information about the genetic flow in the family, and instead these studies treated each progeny as a separated trio (father-mother-progeny).

Therefore, we conducted our study intending to provide an accurate characterization of ASE genes and their putative variants in the TFBS. We utilized a rabbit family consisting of genetically divergent parents, based on the previous work at our lab, and their 8 offspring, combining high-throughput data from RNA-seq and WGS of each individual. Our approach does not require a large number of individuals or a heterozygous variant to be present in the transcript. Instead, it relies on the Gene Expression levels (phenotype) supported by the Mendelian inheritance-based haplotype phasing of the variants found in any region (exonic, intronic or gene surrounding regions). Our approach is expanding the definition of the ASE phenomenon and utilizing it in a large family model by matching the patterns of haplotype with the heterozygosity of the potential regulatory variants in the TFBS. Here we also show that applying inheritance-based WGS data analysis of a larger family is a proper approach for an accurate characterization of the de novo mutations (DNMs). DNMs arise in the offspring and are absent in the parents; they are increasingly studied in farm animals with the advent of WGS. Yet, using only families with trios neglects the important potential shared information in the family and may lead to an inaccurate determination of DNMs. We leveraged our family approach to pinpoint and discover DNMs in the offspring and investigate DNMs in rabbits for the first time.

2 OBJECTIVES

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.
3. Providing a novel pipeline that categorizes gene expression level patterns in a family model.
4. Identifying putative regulatory variants within Transcription Factor Binding Sites (TFBSs) associated with Allele-Specific Expression.
5. Providing sets of genes that might be related to meat quality and quantity in rabbits as well as other farm animals.
6. Analyzing DNMs in a larger family model.

3 LITERATURE OVERVIEW

3.1. Background on gene expression

Deciphering the genotype-phenotype relationship remains an essential quest for biologists. Thus, gene expression serves as a promising area of focus, as it represents an intermediate step between DNA sequence and the observed phenotype. Moreover, gene expression itself is a multidimensional trait, or a phenotype controlled and adjusted by genetic, epigenetic, and environmental factors (Pastinen, 2010a).

The process of gene expression is defined and shaped by reading, interpreting, and conversion of the genetic information during two main processes: transcription and translation. These processes represent the flow of genetic information in the cell into biological functions (Sandra Ramírez-Clavijo et al., 2013, M. Wang et al., 2023). Transcription is a molecular process where information encoded in DNA is converted into RNA types including temporary mRNA. On the other hand, Translation, by definition, is a molecular process in which the mRNA code is decoded to synthesize a polypeptide chain by ribosomes, which are themselves RNA catalysts (Ban et al., 2000). The central dogma of molecular biology illustrates the one-way directional transfer of genetic information from nucleic acid polymers to amino acid polymers (Central Dogma, Francis Crick, 1970) i.e., transcription followed by translation, and therefore the effects on the transcriptional mechanism will influence the proteome of the cell.

3.2. Gene expression regulation

Genes in organisms do not function alone nor do they control their own expression. The gene regulatory mechanisms are complex; they include many interactive processes and stages of regulations determining the timing and the level of expression. Gene regulation is fundamentally responsible for the diversification of other biological processes that occur within the cell e.g. cell differentiation and cell development as well as environmental change adaptation (Bonnot et al., 2021). It is consequently fundamental for selective breeding for important traits.

The regulation mechanisms are multidimensional and multi-leveled controlled processes, including transcriptional and post-transcriptional (mRNA level), translational and post-translational (protein level) (Ghedira, 2018; M. Wang et al., 2023). Specific genes are expressed during organism development leading to the formation of distinguished types of cells, and a great deal of the regulation of this process happens on the transcriptional level.

3.2.1. Transcriptional regulation

In the transcription process, the information encoded in a fragment of DNA is copied into functional RNA molecules. This stage of gene expression is carried out by RNA polymerase enzyme (RNAP), and it requires intricate regulation by transcription factors.

Intrinsically acting or in response to environmental stimuli, specific genomic regions are able to control the expression level of some genes. These regulatory elements can be classified in two main categories, either as cis, i.e., the effect on the expression is of adjacent genes on the same chromosome, or as trans, when the effect happens on physically distant genes (de Souza et al., 2020a). These cis-acting regulatory elements contain unique recognition sites to which the trans-acting regulatory proteins or factors can bind in order to enhance or repress the transcription (Maston et al., 2006a).

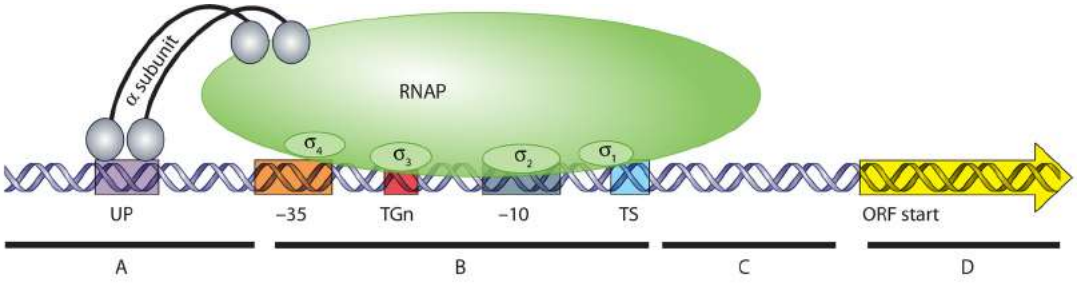
3.2.1.1 Transcriptional regulatory elements /cis-regulatory elements

Cis-regulatory elements are DNA motifs mostly in the non-coding region of the gene. These motifs are recognized by specific proteins (transcription factors), playing a crucial role in the regulation of gene transcription and overall gene expression. In Eukaryotes, RNA polymerase II is responsible for transcribing protein-coding genes. These genes have unique cis-elements that can be classified into a) the promoter region comprising the core promoter and proximal promoter elements located near the transcription start site, and b) distal regulatory elements**,** which include enhancers, silencers, insulators, and locus control regions (LCRs). These distal elements are typically enriched with TFBSs. (Cramer, 2019; Maston et al., 2006b; Pan, 2006).

A- PROMOTER

To initiate the transcription, a complex of general transcription factors (GTFs) (Kuhlman et al., 1999) assemble before the 5’ end of the gene in the transcription starting site (TSS) upstream the gene around a region known as the promoter (Menon et al., 2021). This promoter is recognized by the RNAP, which then unwinds the DNA, initiates the RNA synthesis, and moves away from the promoter until the termination of the gene and eventually releasing the DNA and RNA (Cramer, 2019; Menon et al., 2021).

The promoter sequences often contain conserved motifs referred to as core promoter elements (Figure 1)**.** The core promoter may include the TATA box(~30 bp upstream of TSS in 30 -50% of Eukaryotes), which is present in only about 27% of promoters (Menon et al., 2021), and more recently discovered downstream promoter elements (DPEs), TCT, initiators (Inr), motif ten elements (MTE), and B recognition element (BRE) (Brown, 2018; Menon et al., 2021).



Type	Mechanism	Action	TF binding			
			upstream (A)	core promoter (B)	downstream (C)	ORF (D)
repression	Steric hindrance	No RNAP binding		+		
repression	Roadblock	No transcription elongation			±	+
repression	DNA looping	No RNAP binding	+		+	±
repression	Activator modulation	Prevents activator binding	±	±		
activation	Class I	Interaction α subunit RNAP	+			
activation	Class II	Facilitates σ factor binding	+			
activation	DNA conformational change	DNA helix twist		+		
activation	Repressor modulation	Prevents repressor binding	±	±	±	±

Figure 1: General Architecture of a Promoter & Molecular mechanism of transcription modulation. Illustration of the repression and activation features. (+) refers to TF binds at this location; (+-) means there are several possible positions the TF could bind to. TS indicates the transcription start site, TGn indicates the extended 10 element, and UP points to the UP element. The Open Reading Frame (ORF) is the gene regulated by this promoter. (Menon et al., 2021)

B- Enhancers

Enhancers are cis-regulatory DNA elements, relatively short stretches of DNA in the non-coding region. These elements function in an independent manner of orientation and distance and regulate the transcription of one or more genes (Figure 2). Enhancers can control gene expression from a distance as far as 1 Mb upstream or downstream of the gene and from both orientations. Consequently, these enhancers play a crucial role in the spatiotemporal control of the cell-type development, and their function is essential for tissue development and normal cellular differentiation (Arnold & Stengel, 2023a; Bonev et al., 2017; Nord et al., 2013; Stadhouders et al., 2018). Enhancers were first observed in viruses 44 years ago in the expression of the β-globin gene and reported to enhance the expression of this gene (Banerji et al., 1981), and several papers were reported after that also in viruses before it was the first enhancer reported in Mammalian in the murine immunoglobulin locus (Banerji et al., 1983).
A normal mammalian cell would have thousands of enhancers in an active state, and the Encode project has reported 926,535 cis-regulatory elements in 2020, covering 7.9 % of the human genome, while 87% of these elements are enhancers (Abascal et al., 2020). Even with this small percentage of the enhancer sequences, 64% of non-coding traits associated with SNVs were reported within the enhancers, indicating a great deal of importance of these elements in the

genomic studies related to diseases and other traits (Hnisz et al., 2013; X. Liu et al., 2024). Cis-regulatory elements can work individually or cluster in a larger group, forming what is called the super-enhancers, to regulate gene expression networks and control the cell identity (Arnold & Stengel, 2023b; Lv et al., 2024; Tonekaboni et al., 2019).

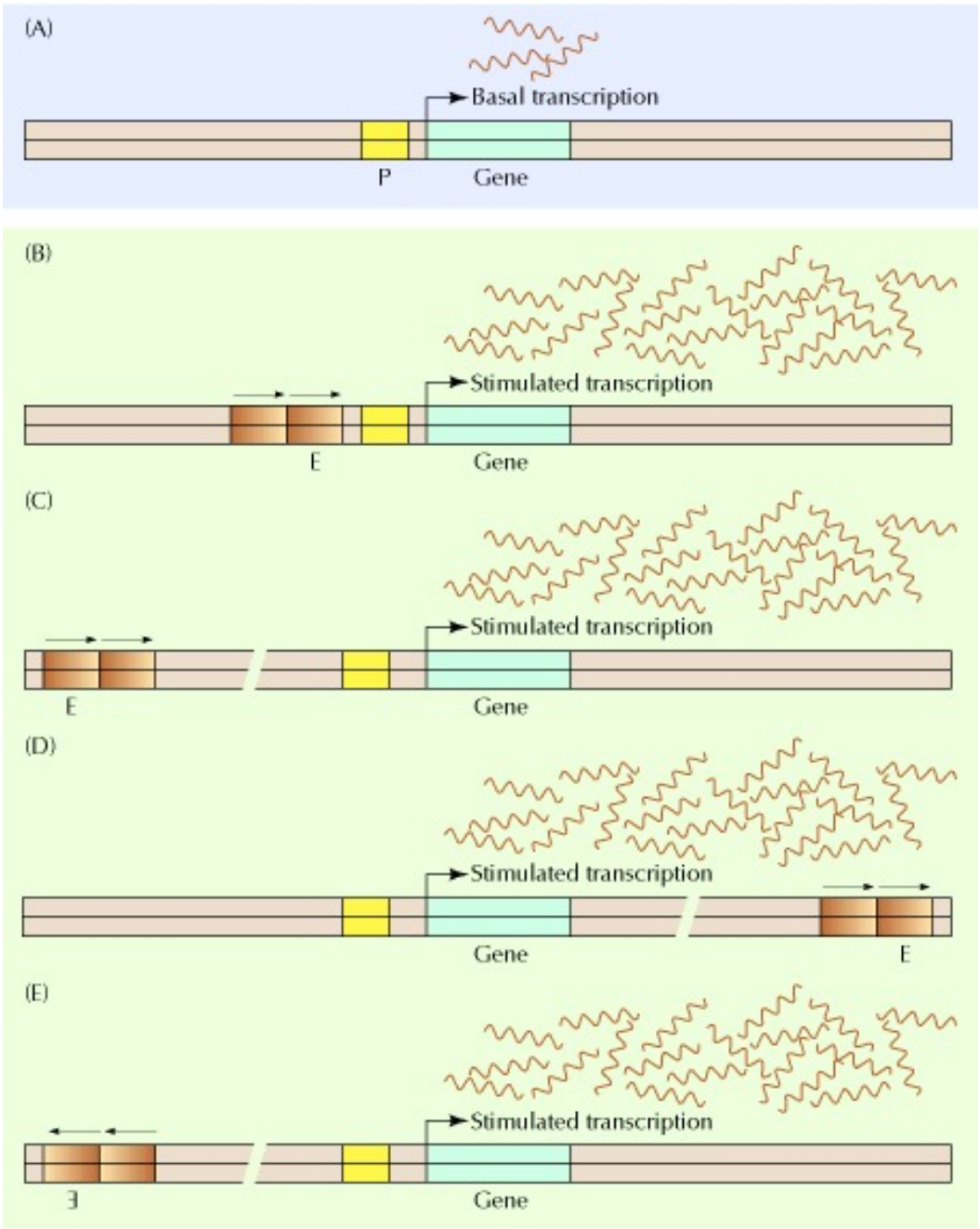


Figure 2: Enhancer effects on the transcription (A). The enhancer E stimulates transcription, its activity is independent of the location or distance from the promoter (B &C), either upstream and downstream from the transcription start site (C & D) or the orientation (E). Cooper GM. Sunderland (MA): Sinauer Associates; 2000.

Both cis and trans factors play crucial roles in shaping phenotypic variation by modulating gene expression (Gompel et al., 2005a; Muráni et al., 2009; Tian et al., 2018). However, these two classes of regulatory factors differ fundamentally in their mechanism of action**,** molecular influence, and evolutionary dynamics. Also, beneficial cis-regulatory variants are more susceptible to fixation in the next generation (Meiklejohn et al., 2014; Wray, 2007). The effect of trans is more dominant, while the combinations of the cis-elements tend to act additively and synergistically with larger quantitative effects to mediate the level of expression of the same target gene. In diploids, cis-regulatory elements are DNA sequences located on the same chromosome and regulate the nearby genes by causing changes in the transcription initiation or expression rate or even in the transcript stability. They tend to regulate the gene expression in an Allele-Specific manner, that a heterozygous variant in the transcript will have a strong imbalanced level of expression for both alleles. On the other hand, trans-regulatory elements—for instance, transcription factors, co-regulators, or non-coding RNAs—are usually encoded at genomic locations far away from the genes they target or regulate (Arnold & Stengel, 2023a; Gompel et al., 2005b; McManus et al., 2010a; Meiklejohn et al., 2014; Wray, 2007). Despite its necessity, the core promoter is not a common point of regulation. Cis-acting regulatory elements form a recognition site to which the trans-acting regulatory elements can bind, and these trans-regulatory elements are the transcription factors, and the recognition sites are the TFBSs.

3.2.1.2 Transcription Factors (TF) and Transcription Factors Binding Sites (TFBS)

To see the complete picture of the gene regulation process, Transcription Factors emerge as key mediators between cell signaling and the regulation process. These sequence-specific proteins are essential to regulate gene expression. TFs bind to the regulatory regions, such as the promoter and the enhancer, based on the likelihood of the presence and high affinity with the regulatory elements. Each TF is typically thought to recognize a set of DNA sequences or motifs, these motifs are typically 6-12 bp in length (Dror et al., 2015; Inukai et al., 2017; Weidemüller et al., 2021a).

Transcription factors can activate the transcription (activators) or suppress the transcription (repressor), and the evaluation and prediction of this functionality on the enhanced activity is not trivial. Depending on the cofactor's availability or the chromatin access, some TFs function in both gene repression and activation mode (Berest et al., 2019). TFs activity can be regulated at two main points: a) regulating its abundance or its active form, and that can be done at any stage of the regulation e.g. post-translational regulation. b) The modification of their accessibility to

the TFBSs. Both ways will lead to the regulation of the given gene by the given TF. However, as soon as the TF binds to the regulatory elements, it can make the chromatin accessible for other factors to bind, or, on the contrary, prevent these cofactors from joining the transcription process (Weidemüller et al., 2021b).

Some studies were conducted in order to understand the basis underlying the histone modification and the expression of the Quantitative Trait Locus (QTL) mechanism (Cavalli et al., 2019; Kilpinen et al., 2013; McVicker et al., 2013; Weidemüller et al., 2021a). These studies concluded that an alteration in the TFBS, and consequently in the affinity of the TF binding to this site, most probably forms the initial step in later events that lead to histone modification, RNAP II occupancy, and the regulation of the gene expression across individuals. Figure 3 provide an example of a variant in the TFBS and its potential influence.

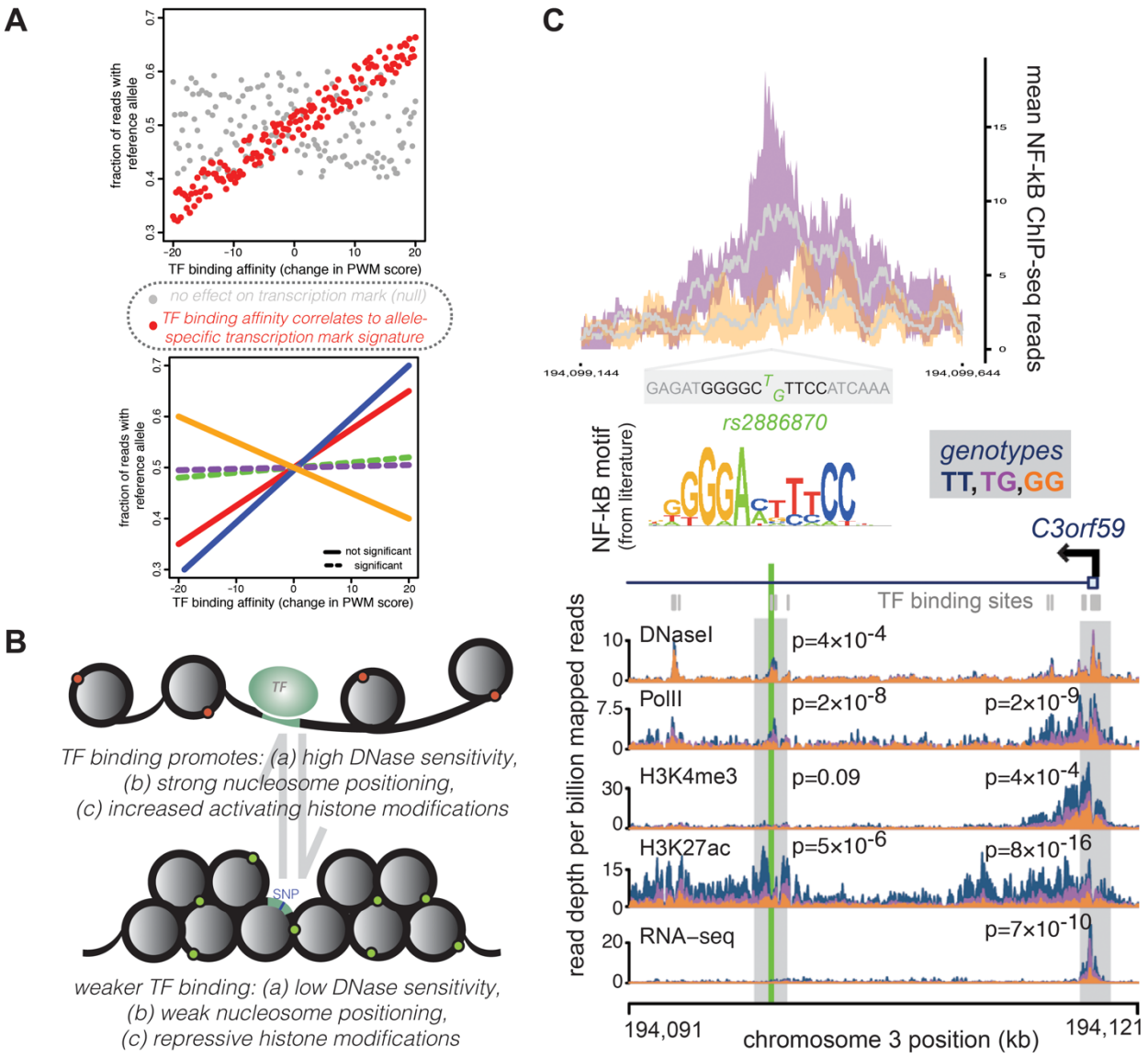


Figure 3: An example of a variant in the transcription factor binding site associated with changes in the regulatory mechanisms, leading to downstream changes. ***A) Top plot: The*** *Allele portion carrying the effective TFBS is plotted against changes in the affinity changes in the position weight matrix score. Red dots indicate a positive correlation between binding affinity and the allele-specific transcription suggesting consequences in the transcription outcomes.* ***Bottom plot:*** *Divergence between strong (solid line) and weak (dashed lines) binding affinity. Each line depicts a correlation of allelic biases observed across several histone modifications, RNAP II localization, and other genomic features.* ***B)*** *An increase in the TF binding affinity can promote open chromatin (by measuring DNase I sensitivity). Overall, SNPs that disrupt TF binding reduce the chromatin accessibility impacting the transcription activity.* ***C) NF-kB transcription factor Binding Analysis****.* ***Top:*** *Tracking the activity of NF-kB transcription factor at a specific locus, showing an SNP "rs2886870" that falls in this TFBS. ChIP-seq data show that LCLs with at least one* ***T*** *allele* ***(TT or TG)*** *match the consensus sequence motif and exhibit higher NF-κB binding compared to LCLs with the* ***GG*** *genotype (no T allele).* ***Bottom:*** *Tracking the chromatin features and RNA-seq at the locus. DNase I sensitivity indicates high chromatin accessibility. RNAPII occupancy. Histon marks for active transcription H3K4me3 and H3K27ac, and the RNA-seq (the transcriptional outputs). (Pai et al., 2015)*

The effects of cis-acting regulatory elements can be distinguished from those of trans-acting elements through allele-specific expression analysis.

3.3. Introduction to Allele-Specific Expression (ASE)

Gene expression regulating mechanisms are proper sources of phenotypic variation among individuals. For example, modulating the expression level of certain genes leads to phenotypic differences with connection to tissue differentiation and development stages in cells within identical genetic makeup (Bonasio et al., 2010). Consequently, gene expression regulation can lead to alleles being expressed unequally, contradictory to what the Mendelian inheritance model expected. To put it in another way, the messenger RNA (mRNA) abundance from paternally and maternally inherited alleles encoded in chromosomes is imbalanced (Cleary & Seoighe, 2021a; Saupe, 2012). This phenomenon where the paternal and maternal inherited alleles are differentially expressed is known as allele-specific expression (ASE). Sometimes, one allele is completely silent and only one allele is expressed; this pattern is called monoallelic expression. This unbalanced expression of alleles is common throughout the mammalian genome, and it has been reported in several studies (Chamberlain et al., 2015; St. Pierre et al., 2022; Tycko, 2010) and may affect the phenotypic variation (Khansefid et al., 2018; Muráni et al., 2009; Tuch et al., 2010).

It becomes clear by now that the genetic variants can impact gene expression on different levels of regulation from impact on chromatin structure (Cavalli et al., 2019) to transcription and posttranscription leading to Allelic Imbalance Expression, as illustrated in figure 4 (Cleary & Seoighe, 2021b). For example, a variant in the TFBS can have an impact on the gene expression level of the linked allele by altering the ratio of expression on each allele and as a result causing ASE (Figure 4 C). By the same token, Alternative splicing can cause this imbalance (Amoah et al., 2021). ASE phenomenon caused by sequence variants differs from the ASE resulting from gene imprinting, which is a form of epigenetic modification where one of the alleles is differentially methylated (Bonasio et al., 2010; Crowley et al., 2015a; St. Pierre et al., 2022). In addition, ASE is a powerful tool that can be utilized to distinguish the cis and trans effect. In the absence of trans effect, TF regulates both alleles equally, but their affinity with the nearby TFBS is different and this is known as the cis effect. Many studies have used ASE analysis with a wide

variety of species to distinguish cis and trans acting elements. Table 1 lists some studies that used ASE to distinguish between cis and trans regulatory elements.

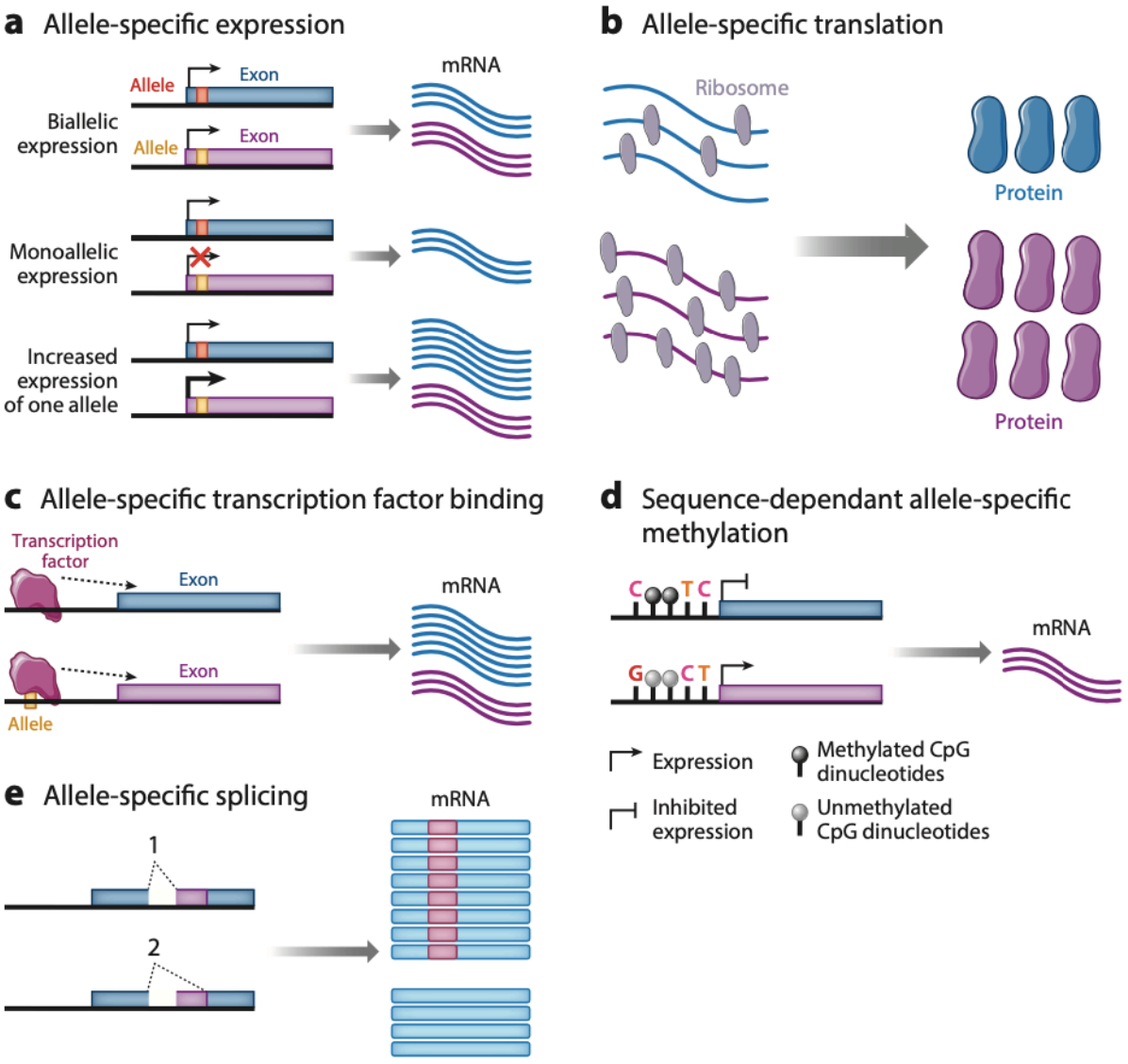


Figure 4: Illustration of Allelic Imbalance Types. a- Allele-specific Expression, the allelic imbalance caused by genetic variants. Three cases from the top: two alleles are equally expressed, only one allele is expressed (Monoallelic), or both alleles are expressed but not equally. Either this ASE is caused due to the different affinity of the Transcription factor due to a variant in the regulatory element as in ***c*** *part, or if the difference in methylation is caused by cis-acting genetic variants this also can lead to allele-specific expression* ***d****.* ***b-*** *Genetic variants can change the rate of the translation leading to Translation allelic imbalance,* ***e-*** *a variant can lead to different splicing results and it is known as Allele-Specific Splicing (Cleary & Seoighe, 2021).*

Table 1 Comparison of the previous studies that distinguished between the influence of cis and trans regulatory elements using the ASE analysis. (Q. Wang et al., 2019)

Species	Tissue	Sex	Cis	Trans	Cis and trans	Conserved and ambiguous	Method	Citation
Drosophila	**Whole fly**	**Female**	**12.4%**	**30%**	**35%**	**22.6%**	**Hierarchical statistical analyses**	(McManus et al., 2010b)
Mouse	**Liver**	**Male**	**14%**	**0.6%**	**17.4%**	**68%**	**Maximum likelihood-based approach**	(Goncalves et al., 2012)
Mouse	**Testis**	**Male**	**24%**	**9%**	**44%**	**23%**	**Hierarchical statistical analyses**	(Crowley et al., 2015b)
Coffea[a]	**Leaf**		**15.5%**	**18.5%**	**17.5%**	**48.0%**	**Hierarchical statistical analyses**	(Combes et al., 2015)
			14.5%	**18.3%**	**16.6%**	**50.6%**		
Chicken[b]	**Brain**	**Female**	**3.45%**	**3.70%**	**4.88%**	**87.99%**	**Hierarchical statistical analyses**	(Q. Wang et al., 2019)
		Male	**3.75%**	**4.86%**	**4.37%**	**87.01%**		
	Liver	**Female**	**7.41%**	**12.92%**	**16.15%**	**63.53%**		
		Male	**8.31%**	**13.93%**	**17.07%**	**60.70%**		
	Muscle	**Female**	**5.60%**	**15.80%**	**10.79%**	**67.82%**		
		Male	**4.72%**	**16.73%**	**11.58%**	**66.99%**		

ASE application in the discovery of rare variants, and the impact of the variation on gene regulation across tissue development (X. Li et al., 2017), attracted enormous number of researchers to study and detect this phenomenon itself and develop tools to infer and uncover the relationship between the regulatory variants' effects and gene expression levels. Moreover, several studies in livestock have revealed the abundance and the importance of ASE genes in the development functions, maintaining muscle tissue and meat quality (Bruscadin et al., 2021; de Souza et al., 2020b; Guillocheau et al., 2019; Y. Liu et al., 2020).

There are two principal approaches to studying ASE**:** a) through expression quantitative trait locus (eQTL) analysis and b) by analyzing genetic hybrids**,** particularly F_1 crosses.

3.3.1. Expression quantitative trait locus (eQTL) analysis

ASE is usually linked to the expression of Quantitative Trait Loci (eQTL), which is the effect of a genetic locus on gene expression. (eQTL) can work as in Cis, affecting a nearby gene, or in trans affecting unlinked genes. (eQTL) can lead to an imbalance in the alternative alleles. (Bader et al., 2015; Bruscadin et al., 2021; Pastinen, 2010b). This method is applied to infer the genetic mechanism behind ASE by conducting a population scale profiling using genome-wide analysis (GWAS) and after that by mapping the eQTLs, this approach is effective, however, it is costly, requires large samples, and involves sequencing many individuals (Kim-Hellmuth et al., 2017).

3.3.2. Allele-Specific Expression in Hybrid Studies

The other way is by conducting an interspecific hybrid F1. The presence of the Allelic Imbalance (AI) indicates a variant in cis-regulatory elements, and by comparing the allele expression originating from the parents, it is possible to infer and characterize these regulatory elements (Macias-Velasco et al., 2021).

Studies involving hybrids offer unique and powerful insights into the genetic regulation of gene expression and evolutionary processes. Beyond elucidating, this approach holds significant value in both biomedical and agricultural research—particularly in the analysis of economically important traits such as meat quality and production. The advantages of using hybrids for ASE analysis can be summarized as follows:

a) Genetic Diversity Enhancement: Hybrids offer a combination of genetic materials from two breeds or species, and that in turn leads to a greater variety of alleles (variants) and

how these alleles might interact to form gene expression variations which are essential for understanding complex traits (Bullini, 1994.).

b) Analyzing Regulatory Mechanisms: The regulatory networks that govern gene expression can be different between parental breeds/species and to understand and clarify these regulatory networks, hybrids enable researchers to study both cis and trans-acting elements (Y. Wang et al., 2019).

c) Adaptive and Conservative Biology Applications**:** Many hybrids are the results of environmental pressure response, making them models for studying evolution, revealing the alleles giving advantages under certain conditions (Bullini, 1994.).

Several studies utilized the power of intraspecific hybrids in ASE analysis in several species such as Humans (Babak et al., 2015), mice (Gregg et al.,2010.), horse vs donkey (X. Wang et al., 2013), and pigs (Lin et al., 2022). However, the majority of this research was oriented to the analysis of the parent-of-origin effect. The parent-of-Origin effect is a special case of ASE where the expression is biased and dependent on the origin of the allele. To date, there is only one recent study in pigs that leveraged the power of hybrid to characterize the ASE regulatory elements in pigs (Quan et al., 2024). The authors achieved a comprehensive set of analyses utilizing different sources of information and sequencing technologies in addition to different stages of development. However, even in this study, F1 offspring were treated as trios not considering the possible shared information among the children in F1, which can serve as a great source of knowledge to further investigate the regulatory mechanism behind the differentially expressed genes.

3.3.2. Experimental And Computational Pipelines and Considerations for ASE Analysis

3.3.2.1. SEQUENCING

Since their introduction in 2005, the advent of Next Generation Sequencing (NGS) technologies, particularly Illumina Sequencing, has revolutionized genomics (Richardson, 2010). NGS enables reads to be sequenced in a massive parallel and produces a massive amount of data. As in the NGS with a read length of 150-250 bp and providing accuracy of more than 99.8%, this technology has also revolutionized the construction of haplotypes and characterizing the genomic landscape (Garg, 2021a) However, with continuous and rapid advancement of this technology, it poses significant challenges in terms of downstream data analysis (Ho et al., 2016).

The allele-specific expression can be inferred primarily from RNA-seq data with millions of reads retrieved from an RNA sample. Whereas to discover the variants in the regulatory elements in the non-coding region, Whole Genome Sequencing (WGS) data is essential (Flanagan et al., 2024). Generating such data involves multiple steps and each step carries potential biases and different factors. The raw sequencing data is typically stored in a file format known as FASTQ. FASTQ file is used to store DNA sequence data, including sequences derived from RNA after converting to cDNA. It combines the nucleotide sequences with the corresponding base quality scores, allowing for reliability assessment simultaneously (Cock et al., 2009). FASTQ file consists of 4 lines: A sequence identifier such as the run information, the actual sequences of the base calls (A, T, C, G, and N), a separator line (+), and lastly the base quality with Phred 33 encoded.

3.3.1.2 Read Mapping and Variants Calling

Allele quantification of ASE from RNA-seq data starts with mapping the sequence reads to a transcriptome or a genome. The results of this alignment process are stored in a BAM formatted file, which is the binary version of SAM file. In RNA-seq, the proper alignment tool should satisfy the following: 1) splicing junctions' awareness 2) compatibility with pair-end reads and 3) working with strand-specific data (Baruzzo et al., 2017). Afterward, reads of each feature (gene, transcript) stored in BAM is calculated and thereafter need to be normalized. From the WGS alignment, the potential variants in the non-coding region can be discovered.
WGS offers advantages such as homogeneous coverage and higher genotyping quality. Moreover, with a decrease in the technology cost, WGS in more advanced high-scale studies, has become feasible, improving variant detection capability (including rare variants) and enhancing the quality of ASE studies (Björn N & Sahlén, 2018).

3.3.1.3 Statistical Analysis of Read Counts Normalizing (Negative Binomial Distribution):

Count-based Differentially Expressed Genes (DEG) is a fundamental approach in comparative sequencing analysis, such as data retrieved from RNA-seq. This approach raises the need for robust statistical methods to evaluate differences among experiments. When data from several individuals is available, there will be a need to qualify the difference between experiments. In this particular case, the aim is to normalize the RNA-seq read counts across the individuals. Most of the statistics-based studies aiming to analyze DEGs implement the negative binomial distribution as a replacement to the Bayesian model in order to achieve this aim such as WASP

(Van De Geijn et al., 2015), and RASQUAL (Kumasaka et al., 2016) and others as it is reviewed here (Cleary & Seoighe, 2021b).

In most cases, the goal of DEG analysis is to provide a list of genes that show statistically significant differences between conditions, typically after correcting for multiple testing and ranking by p-value. The reliability of these results depends on several factors, including the experimental design, the total number of samples, and the number of replicates per condition. The DESeq2 model (Love et al., 2014a) is a widely used and robust statistical framework designed with major aim to analyze DEGs that can be utilized for ASE detection. DESeq2 uses the negative binomial distribution to model RNA-seq count data accurately, considering experimental design and the sensitivity of RNA-seq experimental counts. It also supports complex experimental designs, allowing for reliable identification of differentially expressed genes (DEGs). DESeq2 also provides normalized read counts along with the log2FC and their adjusted p value (adj-p) value at each gene.

3.3.1.4 Haplotype Phasing Information

In diploids, haplotypes are the collection of alleles originated from multiple loci on the same chromosome that was inherited from each parent. Haplotype phasing is the process of assigning each haplotype to one of the chromosomes coming from each parent, it provides the complete description of the genome, and it has several applications in genetics and genomics (Garg, 2021b).

Haplotype phasing can help in understanding the gene function when a mutation causing the Allelic Imbalance is in the Cis-regulatory element (Lo, 2010). Since cis-regulatory elements are located in close proximity to the genes they regulate, it is possible to associate variants within TFBSs with specific haplotypes. This is only feasible when DNA sequence data is available for multiple related individuals. In such cases, the expression pattern of a gene can co-segregate with a regulatory variant on the same haplotype, as this small genomic region is less likely to be broken or get disturbed by recombination events—assuming a relatively uniform probability of recombination along the chromosome. This enables the extension of haplotypes in both directions (upstream and downstream) from the gene toward candidate TFBSs, providing an extra resolution for identifying putative regulatory SNPs that may contribute to ASE or other regulatory effects (Cleary & Seoighe, 2021b).

Haplotype phasing can be laboratory-based approaches or computational-based approaches. Several software has been created to for the haplotype imputation. However, most of this software treat family as a trio phasing problem (two parents and one child) thereby, ignoring the information coming from other siblings, or these software work with reference panels and population data rather than performing genetic phasing in families (Choi et al., 2018). Therefore, when genomic sequences from several related individuals are available, haplotype inference can be achieved through the inherited stretches of Heterozygous variants analysis. This analysis leverages shared chromosomal segments inherited from a common ancestor. This method is particularly applicable in family-based studies, and it is explained in detail in the Materials and Methods section. However, one of the phasing approach limitations is the assignment of de novo mutations (DNMs), where a mutation is present only in the offspring and absent in the parents, to its original parental chromosome.

3.4. De Novo Mutations

Apart from the inheritance of half of the genome from each parent, individuals are also born with novel variants, in a small number (Keightley et al., 2014). The induction of germline variants, or DNMs, controls the pace of genome evolution and is itself a key evolving parameter. DNMs are significant sources of genomic innovation and contribute to inherited traits and diseases; therefore, precise knowledge and understanding of spontaneous mutations are crucial for gaining advanced insights into key questions in diversity and evolution. DNMs are also a major cause of severe, early-onset genetic disorders such as intellectual disability, autism spectrum disorder, and other disorders. The occurrence of new mutations in each generation explains why these reproductively lethal disorders continue to occur in the population. (Acuna-Hidalgo et al., 2016; Lynch et al., 2016; Sakumi, 2019).

The advancement of next-generation sequencing provides a novel tool for the detection of germline mutations directly based on the comparisons of sequences among generations (Scally & Durbin, 2012). A few mechanisms were reported as DNMs induction causes in the human germline. For example, non-allelic recombination (Stankiewicz & Lupski, 2002) replication infidelity (Arana & Kunkel, 2010) and genomic damage (Friedberg, 2003). Consequently, a high interest is shown in understanding the DNMs occurrence frequency in humans (Kohailan et al., 2022; Werling et al., 2018) as well as in domesticated animals (Azevedo et al., 2024). A recent study reported the germline-mutation rate in several vertebrates (Bergeron et al., 2023). The authors of the study used 68 species to discover DNMs and determine the mutation rate. However, all the used species were trios (parents and one offspring), besides, rabbits were not used in the study.

Today, the most widely used methods for downstream variant analysis rely heavily on variant calling based on short-read sequencing (Seah et al., 2023) Although the average depth of affordable genome sequencing is getting larger over time, there are still significant challenges in the variant calling accuracy. Copy number variations, repetitive elements and low complexity regions also cause badly mapped reads (Guan & Sung, 2016), which in turn affect the variant calling and in special cases, like DNMs detection, a high level of fidelity is crucial.
In this thesis, I also demonstrate the impact of using a larger nuclear family on the accuracy of detecting DNMs over the trio approach. We show how increasing the number of utilized offspring in this experiment led to a significant drop in the false positive discovered DNMs.

4 MATERIALS AND METHODS

4.1 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, a commercial line selected for meat production due to its large body weight over more than 20 generations, and the father is a Thuringer rabbit (1.5 to 2kg), and they had 8 offspring. Artificial insemination was used instead of natural mating due to the large difference in size and avoiding complications. The collected semen was used for artificial insemination of the Hycole XXL female. The trial took place at a small-scale commercial rabbit farm that produces meat. The animals were kept under standard livestock production conditions. All animal procedures were conducted in compliance with ethical guidelines and approved protocols. Prior to the collection of the tissue, animals were euthanized using mechanical stunning and were rendered unconscious, ensuring no pain was experienced during decapitation, complying with standard procedures used in commercial rabbit meat production and compliant with Hungarian animal welfare regulations. No additional chemical anesthetics were administered. The experiment and samples collection were carried on at the University of Veterinary in Budapest. The study involves both RNA-seq and Whole Genome Sequencing (WGS) for the purpose of exploring genetic and transcriptomic variants, Differentially Expressed Genes (DEG), and implementing the results in characterizing Allele-Specific Expression (ASE) and potential variants in the cis-regulatory elements laying in the non-coding regions. A total of four biological replicates were obtained from each individual, three replicates from the back and one from the thigh. WGS were also obtained from the same individuals. RNA and DNA extraction was performed at the Institute of Genetics and Biotechnology by the Applied Wildlife and Farm Animal Genomics Group.

4.1.1 Library Preparation and RNA Sequencing

To obtain global transcriptome data, high throughput mRNA sequencing analysis was performed on the Illumina sequencing platform. Total RNA sample quality was checked by Agilent BioAnalyzer using the Eukaryotic Total RNA Nano Kit according to the manufacturer's protocol. Samples with RNA integrity number (RIN) value >7 were accepted for the library preparation process. RNA-Seq libraries were prepared from total RNA using an Ultra II RNA Sample Prep kit (New England BioLabs) according to the manufacturer's protocol. Briefly, poly-

A RNAs were captured by oligo-dT conjugated magnetic beads then the mRNAs were eluted and fragmented at 94-Celsius degrees. First-strand cDNA was generated by random priming reverse transcription and after the second-strand synthesis step, double-stranded cDNA was generated. After repairing ends, A-tailing, and adapter ligation steps adapter-ligated fragments were amplified in enrichment PCR and finally sequencing libraries were generated. Sequencing runs were executed on Illumina NextSeq 500 instrument using paired-end 150 cycles sequencing.

4.1.2. Whole Genome Sequencing (WGS)

The Oryctolagus cuniculus family sequencing and Library preparation were done by Novogene, following the standard whole genome sequencing protocol (paired-end, 150bp read length), on Illumina NovaSeq 6000 system to an average depth of 35.4 (± 6.37).

4.1.3. Quality Control (QC)

Quality Control Check for raw FASTQ files was achieved using fastQC (Andrews, 2010). For the duplication rate check-in RNA-seq, the fastp tool was used, as fastQC is more oriented towards single reads and WGS (S. Chen, 2023)

4.1.4. Generating an Annotation File for Orycun3.0 Reference Genome

Orycun3.0 was the chosen reference genome and an annotation file is needed. Gene Transfer Format (GTF) was generated since this Reference did not have an official annotation. GTF holds information about the genes, and it is broader than the general feature format (GFF). To obtain the annotation, liftoff v1.6.3 software (Shumate & Salzberg, 2021) was used, leveraging the Orycun2.0 official genome as a reference to lift the genes from along with its annotation (Ensembl GTF109). The output is a GFF file, and it was converted into a GTF file using gffread v0.12.4 (Pertea & Pertea, 2020)

4.2. Workflow Design

The workflow was built in a way that enabled us to acquire the powerful potential coming from combining the RNA-seq and WGS data. Variants (SNV and INDELs) were called from both RNA-seq and WGS. Variants from WGS are particularly important for non-coding regions. The read counts matrix was obtained from RNA-seq after mapping to the reference genome (OryCun3.0). Figure 5 represents an overview of the workflow combining the main pipelines used in this study.

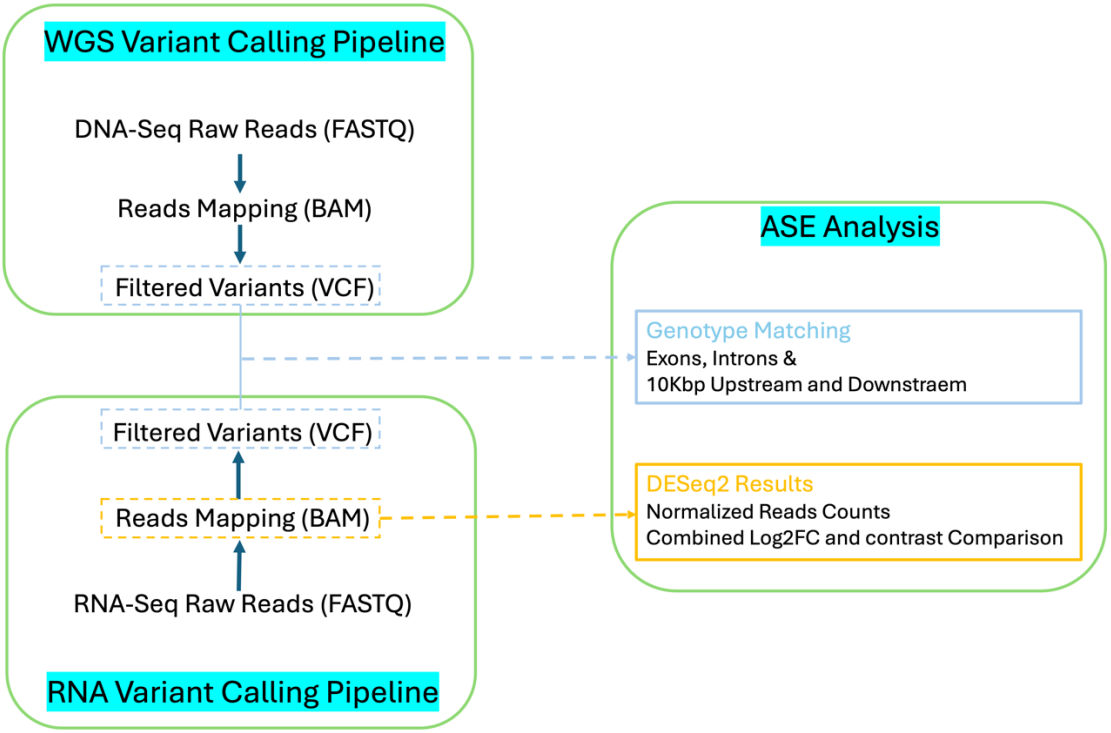


Figure 5: An overview of the pipeline design, harnessing the benefits of using RNA-seq and WGS in ASE analysis. Variants were called from both types of sequences. Read counts were obtained from RNA-seq BAM files to perform ASE downstream analysis and match the results with the variants' genotypes in coding and non-coding regions.

4.2.1. RNA-seq pipeline (Figure 6).

4.2.1.1. Read mapping

We have pair-end reads of 150 bp length from muscle tissue, and for ASE detection, it is necessary to determine the locations where each read was generated relative to the reference genome, therefore the reads were aligned to the OryCun3.0 genome. RNA-seq mapping requires a splicing-aware aligner; therefore, the reads were aligned using STAR 2.7.1a (Dobin et al., 2013) to the genome. First, by generating an index for the genome using the arguments (--sjdbGTFfil and --sjdbOverhang 50), then aligning with arguments (–quantMode GeneCounts). With this option, STAR counts the reads number for each gene while mapping. In the paired-end reads, both ends were checked for overlaps. Also, this option requires an annotation file (GTF) during the genome index generation or during the mapping.

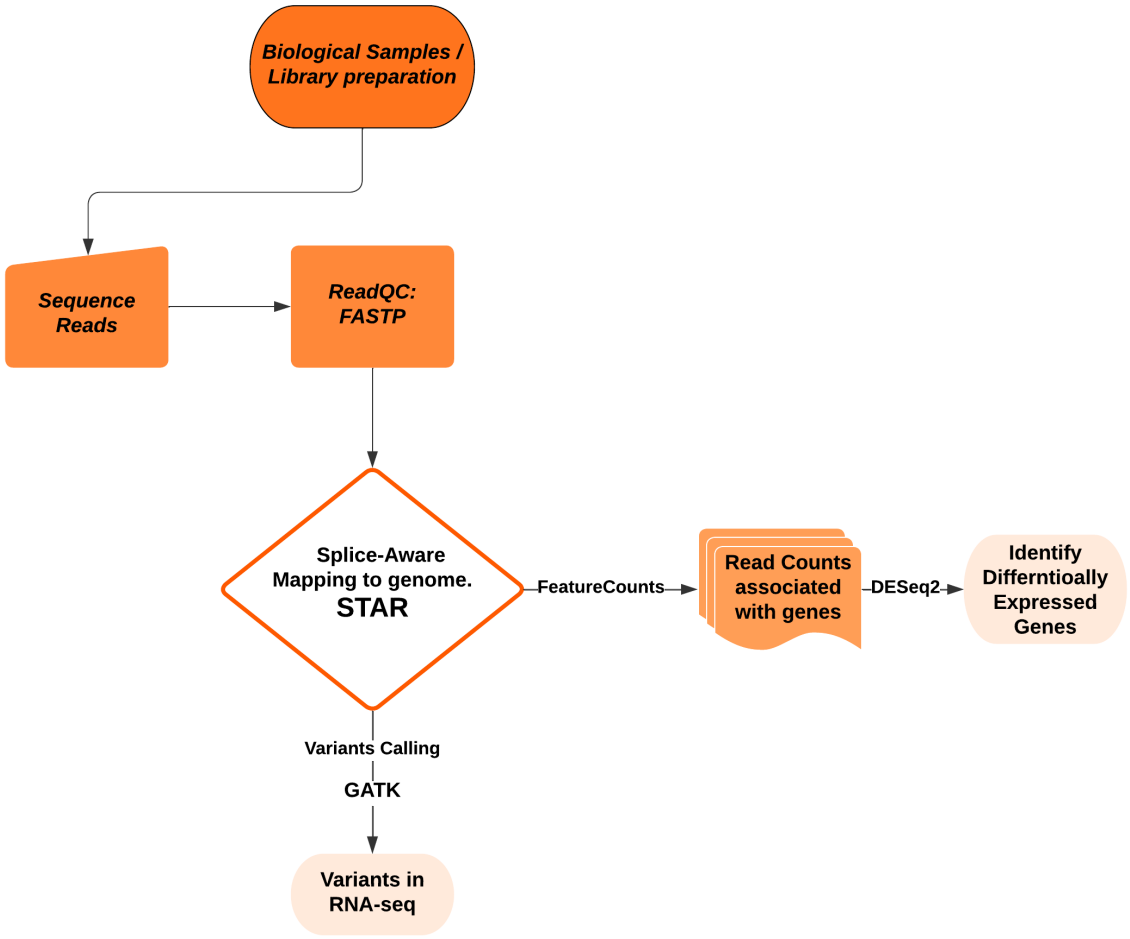


Figure 6: An overview of the RNA-seq pipeline. Initially from the raw sequence and quality control steps, reads were mapped to STAR. Reads count from BAM files used for DEG identification. Variants were called using GATK

4.2.1.2. Read Counts and Differentially Expressed Genes (DEG) Analysis in parents

In order to compare gene expression levels across the rabbit family, gene-level counts were analyzed using FeatureCounts v2.0.1 (Liao et al., 2014), which is a suitable and efficient tool for the general purpose to get sequence reads assigned to the genomic features from BAM files. FeatureCounts was used with optional arguments paired ended (-p).

Based on the read count matrix, DEG analysis was performed between the parents (3 mother replicates VS. 4 father replicates) using the integrated platform for (DEG) iDEP 2.0 (Ge et al., 2018a), with the default parameters, and using (mother vs. father) contract for the DEG section. PCA was produced by the transformed data using EdgeR within the same platform. Similarly to QC, a heatmap was produced after DEG using DESeq2 for the given gene set between the parents for all the genes that exhibit DE between the parents. Gene Ontology Molecular Function (GOMF) was also conducted in the DEG2 section using DESeq2 with default parameters (FDR

cutoff 0.1, minimum log fold change), VolcanoPlot was created using R package Enhanced Volcano Plot (Blighe, 2018).

4.2.1.3. Variants identification

Variants were called from both RNA-seq and WGS. The Genome Analysis Toolkit (GATK) best-practice pipeline was used for the Variant Discovery in High-Throughput Sequencing Data (gatk-package-4.4.0.0-local.jar), although the pipeline slightly differs between RNA-seq and WGS.

For RNA-seq, variant calling was performed on merged BAM files (4 replicates) from each individual after STAR mapping. Reads with more than 2 bp soft-clipped in either end were filtered out, then the best practices GATK (McKenna et al., 2010) and the following steps were performed Markduplicates, Split'N'Trim, base recalibration. Afterward, HaplotypeCaller was utilized to generate gVCF using the quality threshold of Q 20.0, followed by combining gVCFs and genotyping the final VCF file that contains the complete set of variants for the entire family (DePristo et al., 2011a). Additionally, variants that have no-call (./.) at any individual were filtered out, as well as for variants that violate the Mendelian Laws of inheritance. The complete pipeline can be found on the following GitHub repository https://github.com/Maher199/ASE-in-a-family.

4.2.2. Whole Genome Sequencing Pipeline (Figure 7)

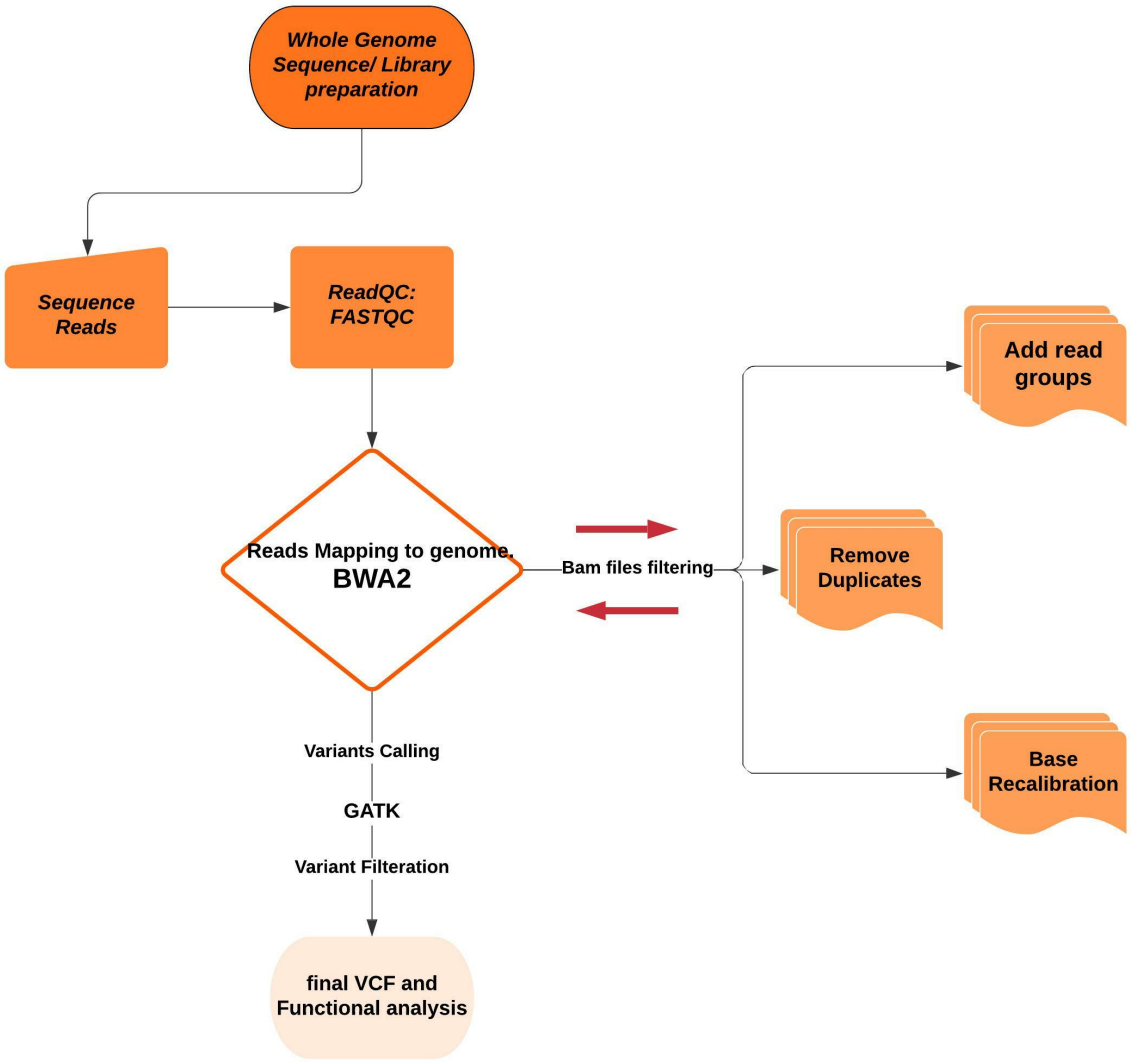


Figure 7: An overview of the WGS analysis pipeline. Starting from the raw sequence and quality control steps, reads were aligned using BWA2. Variants were called and filtered using GATK

4.2.2.1. Read mapping

On the other hand, WGS reads mapping was achieved using bwa-mem2 v2.2.1 (Md et al., 2019). Samtools v1.12 was used to sort and convert the reads to the BAM version (Danecek et al., 2021).

4.2.2.2. Variants Calling

Variant calling from the WGS data began with the mapped reads, followed by applying the following GATK tools (DePristo et al., 2011b) in order: MarkDuplicates, BaseRecalibrator, ApplyBQSR, HaplotypeCaller (-ERC GVCF) with default arguments, CombineGVCFs, and GenotypeGVCFs. After these steps, variants were selected from the combined VCF by their types (SNP, INDEL, MIXED) using SelectVariants, followed by hard filtering the variants by

their types. For SNPs the following filters were used: ("QD < 2.0", "QD2", "QUAL < 30.0", "QUAL30", "SOR > 3.0", "SOR3", "FS > 60.0", "FS60", "MQ < 40.0", "MQ40", "MQRankSum < -12.5", "MQRankSum-12.5", "ReadPosRankSum < -8.0", "ReadPosRankSum-8").
For the INDELs and MIXED variants, the following filters were used ("QD < 2.0", "QD2", "QUAL < 30.0", "QUAL30", "FS > 200.0", "FS200", "ReadPosRankSum < -20.0", "ReadPosRankSum-20").
Finally, the filtered VCFs were merged again (MergeVcfs), zipped, and indexed, variants that violate the Mendelian laws of inheritance were filtered out along with the no-call (./.) loci. The WGS pipeline can be found on the following GitHub repository: https://github.com/Maher199/ASE-in-a-family.

4.3 ASE Analysis in the Entire Family

The mainstream of ASE analysis was performed in the R environment. The read count matrix obtained from FeatureCounts for all biological replicates (after removing mother_2 due to quality control issues, as detailed in Appendix A3) was used as input to get the normalized read counts from DESeq2 v1.42.0, in R (Love et al., 2014b). DESeq2 utilizes methods for the purpose of testing for differential expression by using negative binomial generalized linear models, besides the dispersion estimation, DESeq2 calculates logarithmic fold changes while incorporating data-driven prior distributions. Also adjusted p-values (padj) was calculated using the Benjamini-Hochberg method for False Discovery Rate (FDR) control, as implemented in DESeq2. Genes with less than 10 reads at 3 samples were filtered out (`keep <- rowSums(counts(dds_full) >= 10) >= 3`). DESeq2 was further utilized to estimate contrasts, where each contrast is a linear combination of the estimation of log2 fold changes. Based on the experimental design, contrasts were created for each individual relative to the father: {mother vs. father}, {child_1 vs. father}, {child_2 vs. father}, {child_3 vs. father}, {child_4 vs. father}, {child_5 vs. father}, {child_6 vs. father}, {child_7 vs. father}, {child_8 vs father}. The resulting log2fold changes and padj-values were then merged for each individual in one large dataset. The final dataset contains log2 fold changes and padj-values comparing (in contrast) with the father, allowing us to examine and compare the entire family at one run for the downstream analyses. From this point, the interesting cases were extracted, where at least one individual has a [|log2FC| > 1] and [padj-value < 0.05], and the analysis for ASE was performed.

4.3.1. Pinpointing ASE cases (genes)

The last dataset (selected interesting cases) was used to identify the best and most reliable genes that demonstrate allele-specific expression. Our purpose from this stage is to define the ASE cases. The following criteria were set in a Python script (https://github.com/Maher199/ASE-in-a-family):

In this study, we refer to High Expression with (H), Low Expression with (L), and Moderate Expression with (M), and the order is (Mother _ Father). All the comparisons are relative to the father as it is the baseline of Log2FC:

H_L: This pattern is defined when the mother's log2FC is > 1, and all children's expression levels fall between those of the father and the mother. This also requires that the children's log2FC values span between [-0.2, mother + 0.2] to encompass cases of moderate and dominant expression where children's expression is close to either parent

L_H: (mother's Log2FC < -1), and children [mother - 0.2, 0.2].

H_M / M_L: (mother's Log2FC ≥ 0.8). Children should be in two groups, (group1's Log2FC ≥ 0.8), and (group2's Log2FC ⊆ [- 0.4, 0.4], and at least one of the children is > 1 with p-adj value < 0.05. The idea is that group 1 should be similar to the mother (H) and group 2 similar to the father (M). None of the group's sets can be empty. The thresholds used are experimental, after visual check and evaluation.

L_M / M_H: (mother's Log2FC ≤ - 0.8), (group1's Log2FC ≤ - 0.8), and (group2's Log2FC ⊆ [- 0.4, 04]), one child at least is ≤ - 1 with p-adj value < 0.05.

M_M: (mother's Log2FC ⊆ [- 0.4, 0.4]). Children can be H, M, or L, and at least two groups should not be Empty. At least one child's |log2FC| ≥ 1 with p-adj value < 0.05.

Allele Ratio was also calculated from the RNA-seq variants present in the VCF file, with the purpose of validating the hypothesized haplotype (the gene expression level).

4.3.2. Filtering Genes from the last dataset based on the Exons coverage consistency:

Pysam package https://github.com/pysam-developers/pysam, was used in addition to a custom Python script that can be found on the designated repository(https://github.com/Maher199/ASE-in-a-family), to retain only the genes with at least one exon that has more than 5 reads coverage at each bp in that exon, to ensure the consistency of the coverage along that exon.

4.3.3. Finding the relevant variants in the Genome and in the RNA-Seq:

According to our hypothesis, the phenotype differences across the family are attributed to variants in the TFBS (cis-regulatory regions), where the phenotype is the gene expression itself

in this case. Therefore, we searched in the intronic region and 10kbp of the gene surrounding region for the variants that follow the same order of the phenotype.

By applying separated criteria for each expression pattern, using custom Python scripts, the variants that follow the same pattern as the phenotype of the family at each gene were defined.

H_L & L_H Cases: Variants in the intronic region and 10kb Flanking regions should be as follows: Parents should be homozygous but different (Ref or ALT), and all the offspring should be Heterozygous.

H_M & L_M Cases: One parent is Heterozygous (HET), and the other is Homozygous (HOM). The Offspring should span both categories, following the same pattern as predicted in the haplotype.

M_M Case: Both parents should be heterozygous, and the offspring can be in any category, with respect to the phenotype pattern predicted based on the Log2FC.

By the same token, the same criteria were applied to extract the interesting variants from inside the exons from RNA-Seq and DNA-Seq.

All the previous information regarding each case for every gene was combined and organized in a dataset reflecting each case. We utilized the human ChIPSummitDB database (Czipa et al., 2020), a ChIP-seq-based database of human transcription factor binding sites developed at the University of Debrecen, to predict conserved TFBSs in rabbits. These TFBSs were transformed to Orycun3.0 genome coordinates. The human and rabbit fasta files were first converted to a 2bit format. Using this compact genome representation and using blat, each human chromosome was aligned to the rabbit genome. The following parameters were used: -tileSize=12 -fastMap -minIdentity=98 -noHead -minScore=100. The psl format file's results were converted to chain files with the axtChain command. We set the linearGap parameter to medium according to UCSC best practice. After that, chainMergeSort and chainSplit commands were used. For chainSplit, set-lump=50 was set. Afterward all the chain files were concatenated using the standard 'cat' command in Linux and sorted with chainSort. Next, chromInfo files were generated from the 2bit files using twoBitInfo command. Using the chromInfo files and the sorted chain files, a net file was generated using the chainNet command. The final step was selecting the correct alignable regions from the net file. The netChainSubset command was used for this purpose, and the final liftOver file was utilized to map human chromosome positions to the rabbit genome. Finally, the statistics (e.g., counts of genes in each ASE category, percentages of matched variants, etc.) for all cases and genes in their respective categories were calculated.

4.4. Haplotype phasing

Depending on the block stretches of Homozygosity which are uninterrupted by a recombination event, 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 for the program to work properly and give accurate phasing is to have heterozygous (HET) variants at one parent and homozygous (HOM) at the other parent in the given region, if not, the region should be extended to ensure the condition is met. Additionally, the program outputs a bar plot illustrating the haplotype of each child.

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

1- Iterating through the given region and separating the variants (one parent is HET and the other is HOM) into two tables.
2- Iterating each row in each table separately, the program checks for the grouping, the children that are HET are in the first group, and HOM is in another group, and creates a dictionary for the grouping. The key is the group, and the item is the number of occurrences of this discovered grouping.
3- The reported grouping at each haplotype is the grouping that shows abundance. i.e., the key that shows the highest number of its items.
4- Optionally, it plots the bar plot.

Figure 8 shows the mother haplotype phasing in the children, phasing the father haplotype follows the same principle.

In one table, all the variants that are HET in the mother and HOM in the father were classified. Followed by iterating through the rows and defining the children that have similar genotypes as the mother. Next, grouping was defined, and incidences were counted.

In this case, the first 4 variants have the grouping of children 1,2, and 3 as the mother and this represents the abundance of the grouping; therefore, this is the grouping (A) which will be reported in the final phasing of the mother's haplotype.

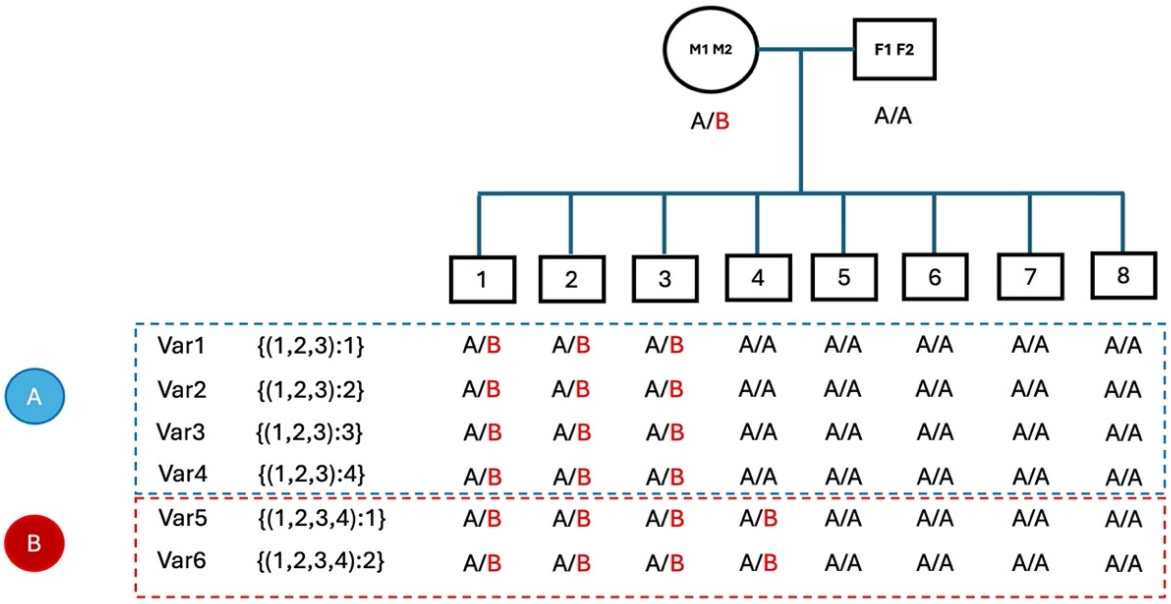


Figure 8: Haplotype phasing logic used by phase_M.py. Illustration of the mother's haplotype phasing in the children. Two possible phasing scenarios (A & B), they show a dictionary example where {(key):value}. The A option has four variants with the same group (block), and this is more than the variants in option B. The groupings of the children are the dictionary's keys, while the number of incidents of conforming variants is the item's value.

4.5. De Novo Mutations (DNMs) Discovery

In order to accurately discover the novel variants in the offspring, few stringent criteria should be applied. The following criteria were set in a python script (Discover_DNMs.py) (https://github.com/Maher199/Discover_DNMs) and applied on the WGS VCF file obtained by the above explained pipeline:

- 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 should be presented.
- The variant should be present at the given child only, while the rest of the children should be homozygous for the reference, just like the parents.
- The minimum coverage threshold used is 15 and the maximum is 45 to limit the false variant calls due to inefficient reads.
- PL > 450. PL parameter in VCF file is the normalized likelihoods of the possible genotypes.
- At least (50%) of the minimum depth (7-8) reads, should support the alternative allele at the DNM.

4.5.1. Filtering DNMs

To ensure more accuracy and exclude repetitive regions from the analysis, RepeatMasker version 4.1.5 (Smit, 2013) was used to identify and mask the repetitive elements that carry the risk of false positively reported DNMs. Then all the overlapping variants (DNMs) reported from the python script were filtered-out in these repetitive regions.

5 RESULTS AND DISCUSSION

The research design was built with the aim of Characterizing the Allele-Specific Expression (ASE) phenomenon in a family, inspecting for potential cis-regulatory elements that control this phenomenon, and analyzing Differentially Expressed Genes (DEG) between the parents. To analyze and discover the genes that exhibit ASE, we utilized a combination of RNA-seq and WGS data. We sequenced a rabbit family consisting of the parents (the mother as Hycole, a meat-producing rabbit, and the father as a Thuringer pet animal) and their eight offspring. For each individual, we had four biological replicates for the RNA-seq, three from the back and one from the thigh. WGS data were obtained from the same individuals. For sequencing, we used the Illumina 2x150bp paired-end technology. The number of reads obtained was 16-28 million in the 40 samples. On average, 87% of them were aligned with the OryCun 3.0 reference genome (Appendix A2). After the quality checking, we noticed that one of the four mother samples generally showed higher expression values and had less correlation with the rest of the replicates compared to the father samples. Therefore, we excluded this sample from the mother–father gene expression comparison and kept it for the ASE comparisons (Appendix A3). Firstly, I will present and discuss the results related to the DEGs between the parents and then discuss the ASE results.

5.1. Gene expression differences between the two parents:

Our primary aim was to detect and explore ASE in rabbit muscle tissue. For this purpose, we chose two breeds that seem to be genetically divergent from each other as was studied and reported by our laboratory (Fekete et al., 2025). Therefore, the father was a Thuringer rabbit, while the mother was a meat-producing (Hycole) breed, which had undergone a rigorous selection process to improve the quantity and quality of the meat. Considering this, we expect a significant difference in the expression level of the muscle-based genes between the two parents. To examine this hypothesis, firstly we generated a count matrix on the gene level for both the mother and the father. Next, we conducted a full differential gene expression (DGE) analysis on the iDEP2.0 platform using the DESeq2 method (Ge et al., 2018b). The results are summarized in Figure 9. Altogether, we identified 773 differentially expressed genes (DEGs) between the parents, where 410 genes were upregulated and 363 were downregulated in the mother (Figure 9 A). In Figure 9 B, we highlighted the most significant (DEGs) in an Enhanced Volcano Plot. Among them, the ENSOCUG00000021647 gene has the highest significance. Interestingly, this is a novel gene located within an intron of the ENSOCUG00000002686 (FAM162A) gene. Its

human corresponding ortholog, CORO2B, is involved in various cellular processes, including cytoskeletal organization and signal transduction. Figure 9 C illustrates the distribution of parent samples according to their gene's expression profiles, suggesting a distinct biological expression or differences in the phenotype between the parents.

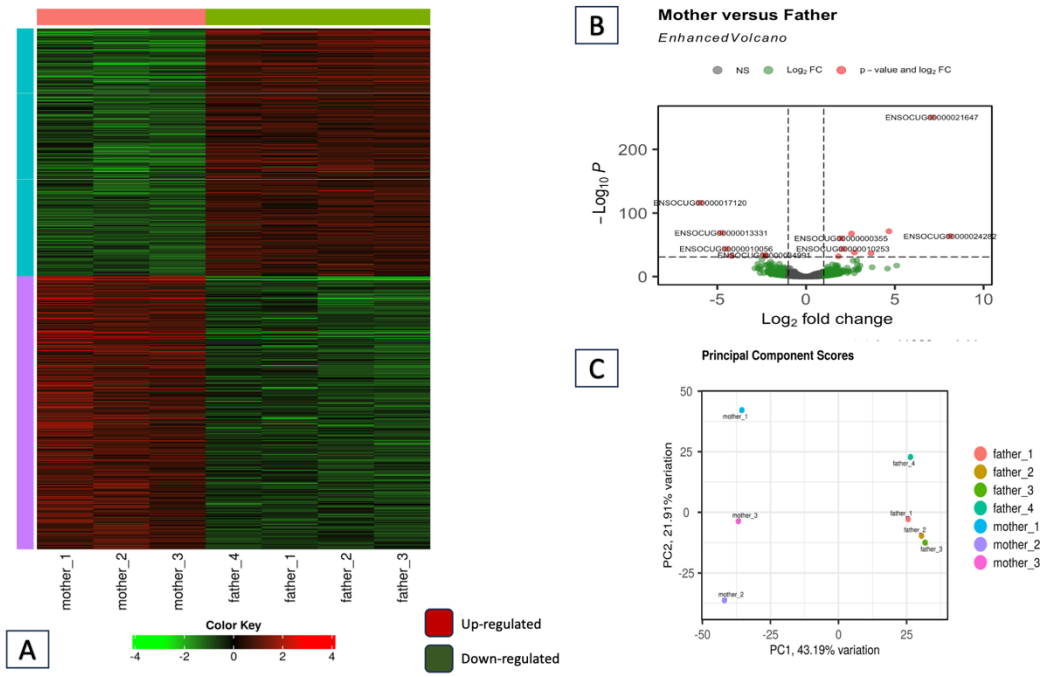


Figure 9: DEGs analysis between the parents. A- Heatmap represents all DEG set between the parents, red: up-regulated, green: down-regulated in the mother. B- Enhanced Volcano plot created in R (Blighe, 2018) and highlights the top DEG, cutoffs used: log2FC=1, p-value=0.001. C- Principal Component Scores for the divergent parents' samples, transformed data with EdgeR (Y. Chen et al., 2025), showing the clustered samples by parent in PCA1 and PCA2.

The highlighted genes in the plot (Figure 9 B) are listed in Table 2. The complete list of up and down-regulated genes is included along with the Gene Ontology (GO) enrichment analysis and can be found on the designated repository on GitHub (https://github.com/Maher199/ASE-in-a-family).

Table 2 Up and Down-regulated genes that are highlighted on the above volcano plot. Gene names are provided when found as well as the Gene Ontology (GO) in Biological Process Terms

Group	Ensembl ID	Gene Name	GO Biological Process
Up-regulated	ENSOCUG00000021647	/	oxidation-reduction process (GO:0055114)

Up-regulated	ENSOCUG00000024282	PLEKHB2	
Up-regulated	ENSOCUG00000021375	RPL12	translation (GO:0006412)
Up-regulated	ENSOCUG00000023448	GSTM2	
Up-regulated	ENSOCUG00000000355	CNTFR	positive regulation of cell population proliferation (GO:0008284); skeletal muscle organ development (GO:0060538)
Up-regulated	ENSOCUG00000010253	TMEM209	
Down-regulated	ENSOCUG00000017120	PDK4	phosphorylation (GO:0016310)
Down-regulated	ENSOCUG00000010056	CYP1A2	oxidation-reduction process (GO:0055114)
Down-regulated	ENSOCUG00000024691	/	ATP synthesis coupled proton transport (GO:0015986)
Down-regulated	ENSOCUG00000034991	UCP3	response to superoxide (GO:0000303)
Down-regulated	ENSOCUG00000013331	GPX1	response to oxidative stress (GO:0006979)

Using the same iDEP 2.0 platform, we conducted the GO Enrichment analysis on these two gene lists (up and down-regulated genes). The Gene Ontology Molecular Functions (GOMF) results indicate an up-regulation of several possible pathways that might be correlated to the phenotype difference between the parents. These pathways include, but not limited to, a growth factor binding pathway and Platelet-derived growth factor binding, which could be directly related to muscle growth and meat production by stimulating the cell tissue of the organism to grow or proliferate, along with other pathways such as 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. The following (Table 3) lists the top GO Enrichment pathways divided as up/down-regulated groups.

Table 3: 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

5.2. Discussion on DEGs in the parents

Our GOMF results indicate an up-regulation of several pathways that might be correlated to the phenotype difference between the parents. The gene set provides a growth factor binding pathway and Platelet-derived growth factor binding, which might be in a direct connection to muscle growth and meat production by stimulating the cell tissue of the organism to growth or proliferation, along with other pathways such as extracellular matrix components, skeletal system development, glycosaminoglycan binding, collagen binding, and fibronectin binding. These pathways, which are related to development and growth, are essential for several biological processes, such as tissue regeneration and embryonic development, besides the functions of the organs. Among these significant pathways we are highlighting some of the reported genes, in the growth factor binding pathways, important genes were identified such as the collagen types of alphas (COL1A2, COL1A1, COL3A1) that are highly enriched in Glycine and proline, besides the insulin-like growth factor binding protein IGFBP genes (IGFBP2, IGFBP5, IGFBP6), and (KLB) which has a positive cell population proliferation regulation. Also, the platelet-derived growth factor receptor alpha (PDGFRA), (ACVR1B), (MYO7A) and the aldehyde oxidase (AOX1) were also found to be up-regulated in the mother. This AOX1 gene was reported in previous studies related to meat and muscle growth in farm animals (Guillocheau et al., 2019; D. Liu et al., 2023), and it is responsible for encoding a homodimeric protein, and this gene plays a critical role in muscle development in cattle. Along with other genes such as KCTD15, UCPs, SESN1, MGP, PLBD.

On the down-regulation side, several factors that work in collaboration and counteract the growth and the accumulation of protein and fat in the wild rabbit were identified. These mainly include, but are not limited to, the catalytic activity that aims to modify the proteins and metabolize fatty acids and increase the oxidation-reduction process. It is also possible to regulate the temperature of the body. For example, Catalase (CAT) was found to be down-regulated which is released in response to stress oxidation and reduction, as well as acyl-CoA oxidase 2 (ACOX2) which is responsible for the metabolic process of fatty acids. Moreover, down-regulation of the estrogen 2-hydroxylase activity.

Considering that muscle development is a complex process governed by different pathways up and down-regulated genes, this analysis might provide insight into the pathways and important genes that regulate muscle development, muscle disease, and tissue development. Overall, the phenotype difference between the father and the mother can result from the interactions between several factors. These factors are working together and sometimes in a counterpart. From one side, the upregulated genes in the mother are oriented to increase the development of the connective tissues and macro proteins, enhanced by the growth factor genes and collagen binding pathways. On the other side, the down-regulated genes in the mother tend to target the genes that accelerate and catalyze the lipids using energy from the hydrolysis of ATP.

5.3. Allele-Specific Expression Characterization

ASE, where parental alleles are expressed at different levels in the offspring—can arise as a consequence of variation in either cis- or trans-acting regulatory elements. ASE can be detected Either by an expression quantitative trait loci (eQTL) analysis across large populations, which associates genetic variants with differences in gene expression, or utilizing hybrid-based studies, such as crosses between genetically distinct breeds, where ASE is assessed in the F_1 offspring. We designed our pipeline based on the following hypothesis: a heterozygous locus in a cis-regulatory elements, where the transcription factor is binding, is necessary to induce ASE. In the conventional approach, heterozygous loci in the transcripts are needed to detect and count for ASE. However, we designed a new experimental setup suitable for detecting ASE in a family without the need for the heterozygous sites in the children's transcripts. However, having any heterozygous variant in the transcript can serve as a validation focal point to our hypothesis.

5.3.1. The experiment design overview

The family-model approach consists of the parents and eight offspring. Considering the gene expression as the phenotype, we sought to match the phenotype with the variant genotypes in the

hypothetical driving TFBS, based on the configuration of intermediate inheritance, we propose classifying expression levels into three categories: High (H), Low (L), and Moderate (M). Figure 10 illustrates these possible matching scenarios in the family-model approach.

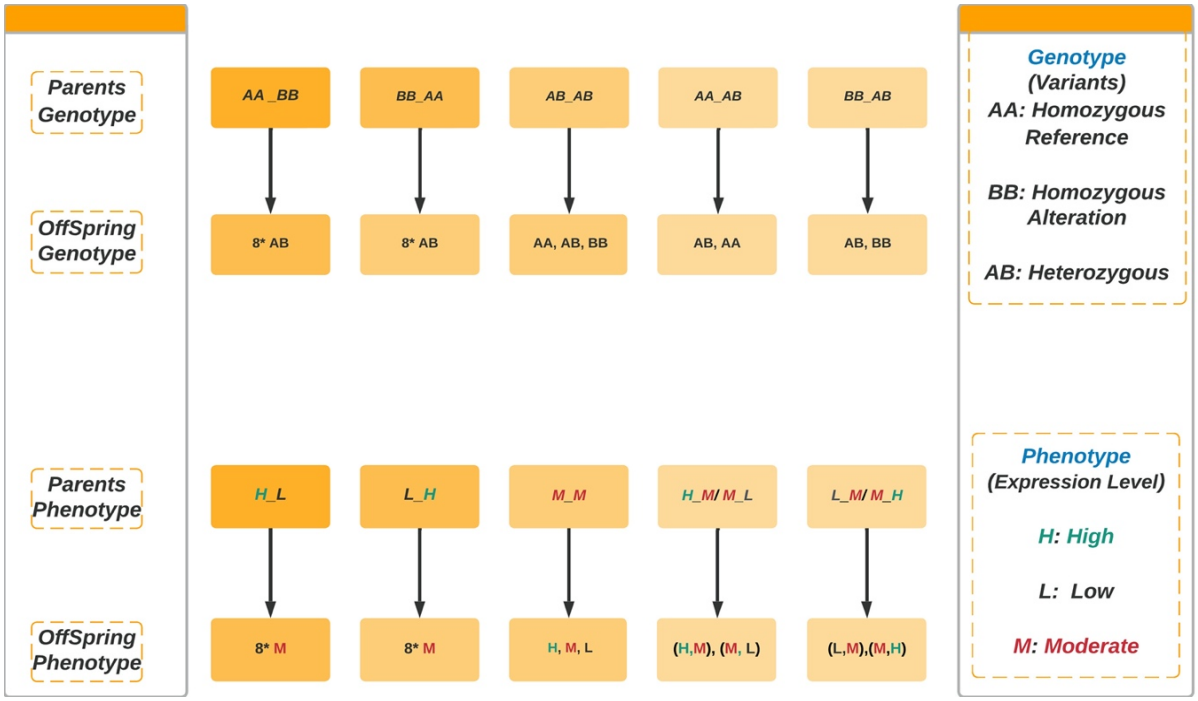


Figure 10: Possible combination of Genotype and Phenotype inherited patterns with the conformation of the Mendelian Laws. Color-coded as the expression level (H: green, L: black, and M: red)

To ensure more variants in the family, we chose the two parents which are genetically far to have as many variants as possible. Table 4 presents the numbers of variants found at each individual.

Table 4: Number of Variants found at each individual in the family (Heterozygous & Homozygous Alteration). Considering the Orycun 3.0 the reference genome as A.

Sample	Heterozygous (AB)	Homozygous_Alteration (BB)
mother	9517501	18607460
child_1	11897605	17375538
child_2	11845501	17376134
child_3	11847617	17373343
child_4	11807813	17412656
child_5	11690283	17493245
child_6	11874573	17393958

child_7	11752351	17434102
child_8	11790412	17428400
father	10416784	18170458

Accordingly, the mother Hycole animal had 9.5 million heterozygous and 18.6 homozygous ALT variants, while the Thuringer father animal had 10.4 million heterozygous and 1.6 homozygous ALT variants. More importantly, there are 3.3 million sites where one of the parents is HOM ALT, while the other one is HOM REF. Similarly, there are about 13 million sites where one of the parents is heterozygous, while the other one is homozygous. Finally, there are about 2.9 million sites where both parents are heterozygous, as it is concluded in the Table 5.

Table 5: Summary of WGS variant counts in parents considering all possible combinations of the Homozygosity and Heterozygosity between the parents considering the Reference genome Orycun3.0 as A.

Reference	Mother	Father	Variant Counts
A	AA	BB	1 612 167
A	BB	AA	1 693 817
A	BB	BB	13 833 792
A	AB	AB	2 896 419
A	AA	AB	4 198 928
A	AB	AA	3 657 972
A	BB	AB	2 765 210
A	AB	BB	2 439 608

Table 5 shows that there is indeed a significant difference between the two breeds used as parents in this family experimental design.

5.3.2. 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 figure 10, we will 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. It is worth mentioning that the cases (H_L, H_M, and M_L) are what can

be observed during the comparison between the parents' transcription when the mother has a higher level of expression, while the opposite can be found in the cases (L_H, L_M, and M_H). In fact, the (L_M and M_H) are phenotypically the same and can only be distinguished after genotype analysis and assigning the M phenotype to the heterozygous allele. The same is true for (H_M and M_L). In cases with the M phenotype, we hypothesize a heterozygous genotype in a regulatory region.

The main novelty of our family-based approach is that we can compare the expression levels across the entire family in one run. Therefore, the expression can be assigned into categories according to which children exhibit a similar level of expression to which parent and if the children are assigned to more than one group (other than the M), then we can safely hypothesize that one or both parents have both the H and L alleles in the regulatory region, which is responsible for the ASE. In other words, if one or both parents are heterozygous in a locus that is responsible for the ASE, that trait will segregate in the children. Otherwise, if one parent is H and the other is L, then all the offspring must have an M level of expression. Naturally, with eight children, we have a better chance to see more expression patterns comparing to families with only one child (Trio). According to this above-mentioned logic, similarly to the parents, we determined the gene expression levels of all eight children. We then compared everything vs. the father as the baseline in terms of log2FC. Figure 11 shows examples of all the possible expression patterns in the family. Utilizing all this expression information and based on the intermediate inheritance, the possible ASE categories can be determined by applying the following criteria:

1. The two parents have significantly different expression levels $|log2FC| > 1$ and p-adj value < 0.05. If all the children show medium (M) expression levels between the parents, then these genes can be assigned into the H_L or L_H categories (Figure 11; A, B) depending on which parent's expression was higher (the first letter refers to the mother).

2. Again, parents have different expression levels, but the individual children's expression levels are distributed into two categories, which align with the parents' two expression levels, and at least one individual has $|log2FC| > 1$ and p-adj value < 0.05. We can assume that one of the parents is under heterozygous regulation (M). (H_M or L_M) and (M_H or M_L) cases fall into this category (Figure. 11; C, D).

3. There is also a special case where both parents' expression levels are similar ($Log2FC \subseteq [-0.4,0.4]$), but complying with the Mendelian law of segregation, the expression levels of the children fall into three categories (Figure 11, E); some will be above (H), some will

be below (L), and some will be similar (M) to the parents' expression level. Although eight children are not enough for deep statistical analysis, the ratio of the three cases will comply to some extent with Mendel's law of segregation in F2 (1:2:1). Here, the second category with the two times ratio is for the children with a similar expression level to what is observed in the parents.

4. As expected, there are many cases where the parents' expression levels are similar, but the children do not segregate. In these cases, we can assume that there is no ASE (Figure 11, F).

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. 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 6). The plots, scripts, normalized read counts, and predicted ASE types are available in the repository (https://github.com/Maher199/ASE-in-a-family). Notably, the majority of 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 6: 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**

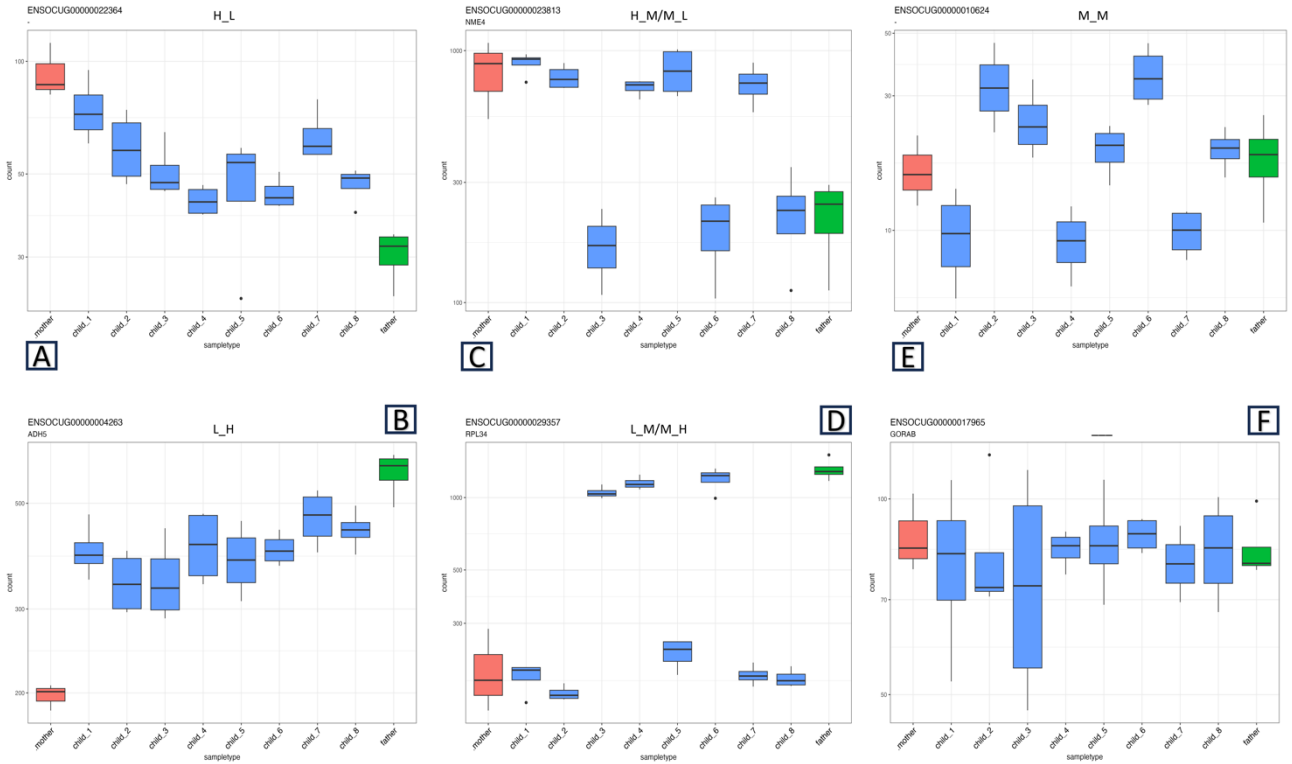


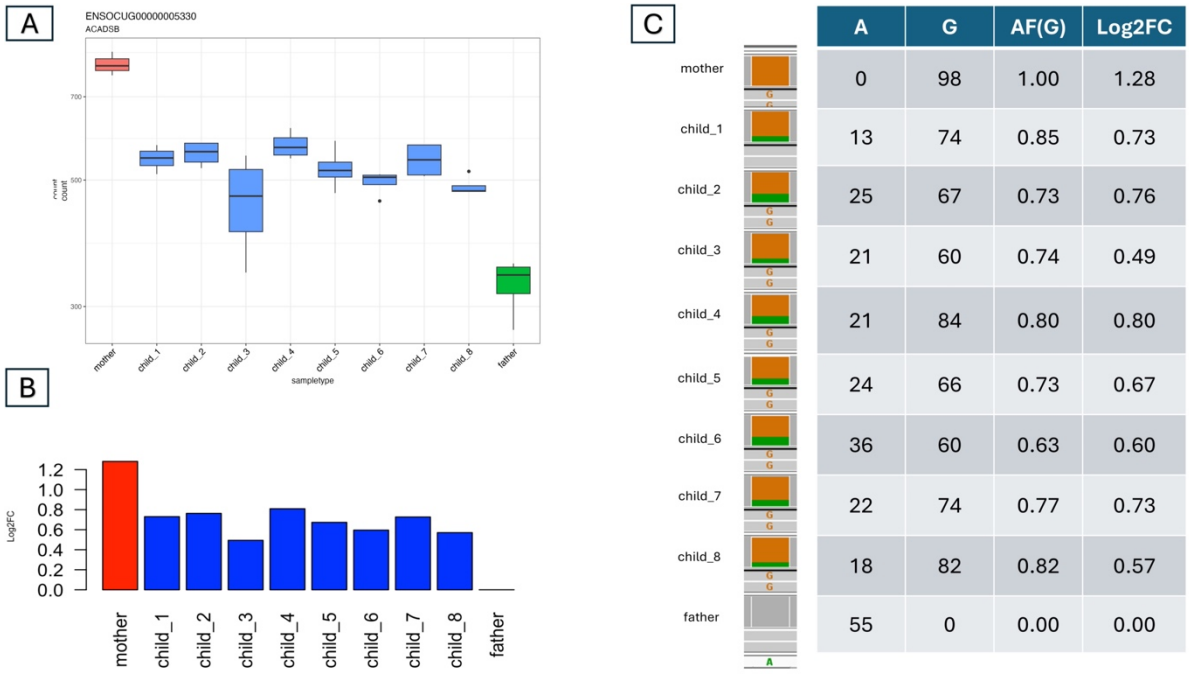
Figure 11: Inheritance Patterns of Gene Expression Across the Family. Cases examples of possible gene expression in our family model. The X-axis lists the family members, and the Y-axis shows the RNA-seq normalized read counts by DESeq2. A & B: (H_L) and (L_H) cases respectively, where the parents are 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.

5.3.3. 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 (in our case the father). However, in the traditional approach involving crossing two breeds where the RNA-seq is available for the entire family, the ASE can be detected only if at least one heterozygous site can be found in the transcript (exonic region). In order to demonstrate the feasibility of our approach and advantage over the conventional approach, we carried out the analysis using the traditional method, so we can ensure that our approach includes all the cases from the traditional approach and also other cases that lack the variants in the transcripts. We can count the reads bearing one or the other allele in these cases at each heterozygous variant in the transcript. We can also do the same approach by counting alleles reads at the eight children and compare the results considering the heterozygous variant found in the transcript. After variants calling from RNA-seq, we checked the allele ratio at each heterozygous variant in the exonic region of the ASE genes discovered by our method to validate our bulk expression-based predictions. We only considered variants that conform the expression patterns of the family. We found that not all genes have variants in RNA-seq, we found about 19% of the expressed genes (2440/12659). We found 42% of genes in H_L have variants in RNA-seq that conform to the expression, 51% in L_H, 13% in L_M or M_H, 6.7% in H_M or M_L, and only 0.3% genes in M_M to be reliable and conform the expression. The results are summarized in Table 6 and the entire gene set analysis along with the Allele ratio at every variant can be found in the tables library on following repository (https://github.com/Maher199/ASE-in-a-family).

Figure 12 demonstrates a simple case of the acyl-CoA dehydrogenase short/branched chain (ACADSB) gene from the predicted H_L cases. Here, the expression data is based on normalized read counts (Figure 12 A) and the father-based log2fold changes (Figure 12 B). Based on these counts and log2FC, it is clearly shown that, indeed, all eight children have intermediate expression levels (M) between the mother (H) and the father (L). Although this result clearly indicates the ASE, we have also examined the allele ratios in the children in order to compare the findings with the conventional approach. Figure 12 C shows the results of Allele ratio at one of the heterozygous variants found in the transcripts of this gene. As it is shown in the Figure 12 C, the mother is homozygous (GG) for the alternative allele (HOM_ALT), the father is homozygous (AA) for the reference allele (HOM_REF), and all children are heterozygous (HET) (AG). The allele frequency of the G allele in the mother is 1.0 and in the father is 0.0, and children are between 0.63 and 0.85 indicating ASE (0.5 would have indicated

equal expression at both alleles). Allele frequency correlates quite well with the measured Log2FC (Figure 12 C).



A	G	AF(G)	Log2FC
0	98	1.00	1.28
13	74	0.85	0.73
25	67	0.73	0.76
21	60	0.74	0.49
21	84	0.80	0.80
24	66	0.73	0.67
36	60	0.63	0.60
22	74	0.77	0.73
18	82	0.82	0.57
55	0	0.00	0.00

*Figure 12: Comprehensive representation highlights the expression Profile of an H_L Gene (ACADSB) Across Family Member, exhibiting all three expression types (H, M, and L). A: RNA-seq normalized read counts. The X- axis represents family members, and Y-axis represents the normalized read counts. Mother is **H**, father is **L**, and all the offspring are **M**. B: Log2 Fold Change (Log2FC). Expression quantified of the family member in relative to the father. The mother is H (> 1 Log2FC). C: Heterozygous variant identified in the exon (the left bar is an IGV screenshot): the mother is **HOM_ALT** (G/G), the father is **HOM_REF** (A/A), and all children are **HET** (G/A). The table lists the allele read counts at this variant. The read count ratio (allele frequency) is calculated, with the Log2FC values for this gene provided in the final column.*

In this study, we also implemented an additional approach to support phenotype validation through haplotype phasing. We developed an algorithm that resolve and phase offspring haplotypes by identifying contiguous stretches of DNA variants inherited from each parent. The custom script, Phase_M.py (https://github.com/Maher199/ASE-in-a-family), was applied to the relevant genomic regions allowing to generate the phasing panels presented in the following figures.

During the analysis of RNA-seq variants, we found that different combinations of variants can exist in many genes, especially if they have long introns and/or many exons. Figure 13 shows an example for BDH2 gene which is a lipid deposition-related genes reported in (L. Wang et al., 2021). Based on the gene expression data of the parents and offspring (panel A), this BDH2 gene is initially assigned into the (L_M or M_H) category. According to our hypothesis, this implies that one of the parents must be heterozygous in a site where the other parent is homozygous.

Surprisingly, when we inspected the overlapping variants, we found examples for different types of combination, which involves at least one heterozygous variation at the parents (Figure 13 D). In this panel, the first column represents an IGV screenshot of a variation where both parents are heterozygous, and the alleles segregate in the F2 in the 1:6:1 ratio. This variation cannot explain the observed expression pattern because the expression level is different in the parents. In the second column of panel D, the mother is HET (A/G), while the father is HOM (G/G). From this, we would have inferred that this gene might be in M_H category. However, when we compared the variation pattern of the children to the expression pattern, we found that they did not match. Finally, the last column of Panel D shows an exact match to the observed expression pattern and thus confirms our ASE prediction. Here, the mother is HOM_REF (AA), and the father is HET (A/G). The eight children are HOM, HOM, HET, HET, HOM, HET, HOM, HOM, which aligns perfectly with the observed M, M, H, H, M, H, M, M expression pattern, respectively, as well as with the haplotype phasing at this gene region (figure 13 C). This sort of analysis allows us to differentiate the actual ASE in the H_M or M_L and L_M or M_H cases respective with their expression. At this example, the BDH2 gene, and based on the mere expression analysis, it is not feasible to decide whether it is an L_M or M_H ASE. However, after scrutinizing the variation patterns, we found that the accurate match was where the mother is the homozygous one. This unanimously confirmed that this is an L_M ASE because the M parent must be heterozygous.

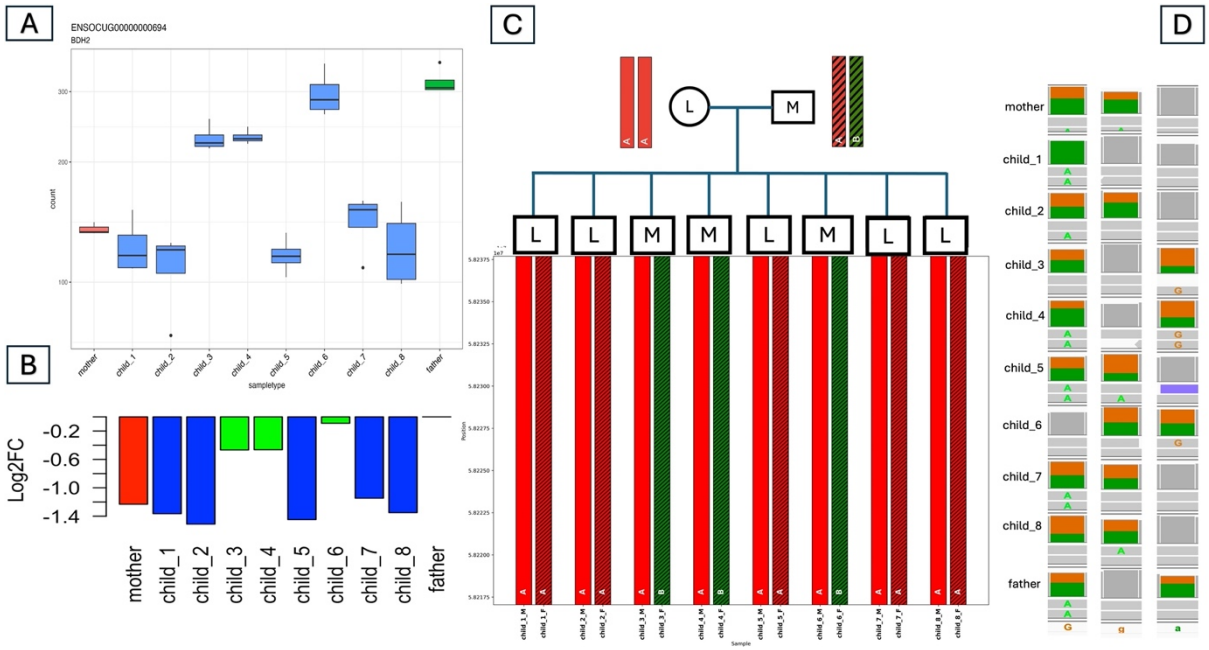


Figure 13: BDH2 gene with allele-specific expression analysis across the family members. A) normalized read counts at all family members, first is the mother in red and last is the father in green and children are in between. B) Log2FC comparison in the family relative to the father as the 0 baseline. Mother in red and children with <= - 0.8 log2FC are in Blue. C) haplotype phasing in the family, low expression (L) is homozygous (AA) and Moderate expression (M) is heterozygous (AB). D) IGV snapshots for three different genotype combination of variants in the same gene, only the last variant matches the expression pattern mother (AA), and father is (AG).

Using the same approach, we checked all predicted ASE for transcript variants. Table 4 summarized the cases numbers along with the variants discovered. Surprisingly, we found variants only at a small fraction of the predicted ASE genes. For example, out of 469 predicted M_M cases only one gene contained a heterozygous variant at both parents in the RNA-seq. On the other hand, in most cases, there was more than one variant. As in the above-described gene, the allele ratios correlate with the changes in the expression values.

5.3.4. Genotype matching with the predicted phenotype

In general, we hypothesize that a heterozygous variant (AB) in a cis-regulatory element can lead to the three distinguished expression phenotypes (H, L, and M). The homozygous genotype is responsible for the H or L and the AB for the M phenotype. This means that if we see, for example, an H_M ASE, then the genotype at the parents must be AAxAB, and the offspring should be segregated into both the AA and the AB genotype's groups. Therefore, finding variants that exhibit the same pattern as the predicted ASE type can support our original expression-based ASE predictions.

To carry out this analysis, a whole genome sequencing (WGS) was obtained for all ten animals with a ~30x coverage (Appendix A1 for QC). After the mapping and the variant calling, variants were examined at ASE-predicted genes. Three regions were considered: the transcripts (Exonic region, practically what we would see in the RNA-seq), the introns, and the gene surrounding 10 kb up and downstream regions at each gene. To see the entire analysis, how many variants support the expression-based ASE predictions at these three regions at each ASE-predicted gene, check the following repository (https://github.com/Maher199/ASE-in-a-family). The results are summarized in Table 6. For the H_L and L_H cases, the predicted ASE genes are confirmed with variants in 56.7 and 62.5%, respectively. In these cases, it means that there was at least one variation in these regions, where both parents are homozygous (REF and ALT), and all children are heterozygous (HET). This matches perfectly with the observed moderate (M) expression phenotype in the children. In the categories of (H_M or M_L) and (L_M or M_H), from the 119 and 133 predicted genes, confirming variation patterns were found only at 29.2 and 33 percent. Surprisingly, this ratio is much lower at the 469 predicted M_M cases, where only 7.3 percent of the genes had at least one confirming variation.

M_M cases are unique in that they cannot be observed without inspecting the expression levels of children, and at least one child should be homozygous. Figure 14 provides an example of a novel gene that exhibits all expression levels in the children (L, H, and M). In addition, the variants at the gene region and haplotype phasing support the expression pattern.

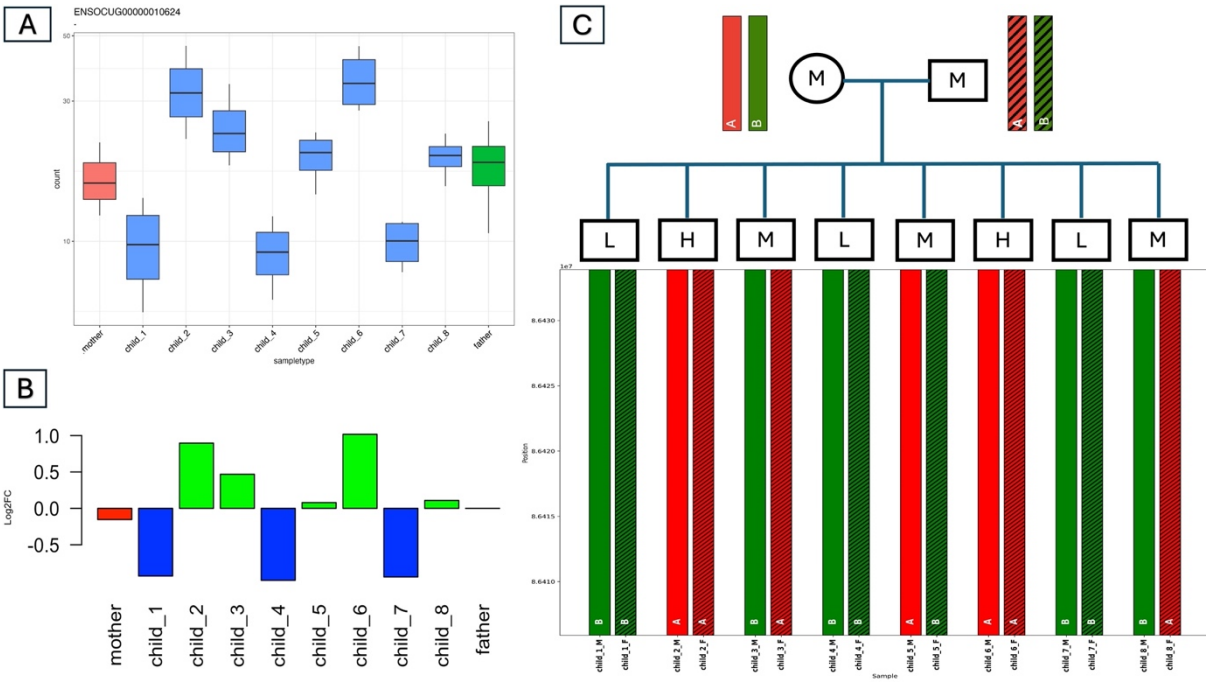


Figure 14.: A novel gene (ENSOCUG00000010624) expression analysis across the family member. A) normalized read counts at all family members, first is the mother in red and last is the father in green and children are in between. B) Log2FC comparison in the family relative to the father as the 0 baseline. Mother in red, and children <= -0.8 are in Blue and reflect the (L). c) Haplotype phasing at the gene region, heterozygosity (AB) is in accordance with the (M) expression in children (3,5&8), (AA) reflects the (H) expression in children (2&6), and (BB) reflects the (L) expression in children (1,4&7).

A continuous analysis of M_M cases led us to find some interesting cases where all the offspring exhibit Moderate expression (M) except for only one child which is rather high (H) or low (L). In some genes, only one child shows H or L while the rest of the children are M, and that can be explained as the alleles that are responsible for this distinguishing expression level, met only once in the designed family. In the following example (figure 15), the B allele, which is responsible for L expression, and coming from the mother,

matched with the other B allele coming from the father only once in the first child (BB). Interestingly, no heterozygous variant was found in this gene region. Therefore, haplotype phasing for this gene was only feasible after extending the phasing region beyond the gene borders. The resulting haplotype conforms to the expression pattern: all children are HET (AB), except for the child_1, which exhibited L expression and whose haplotype was HOM (BB).

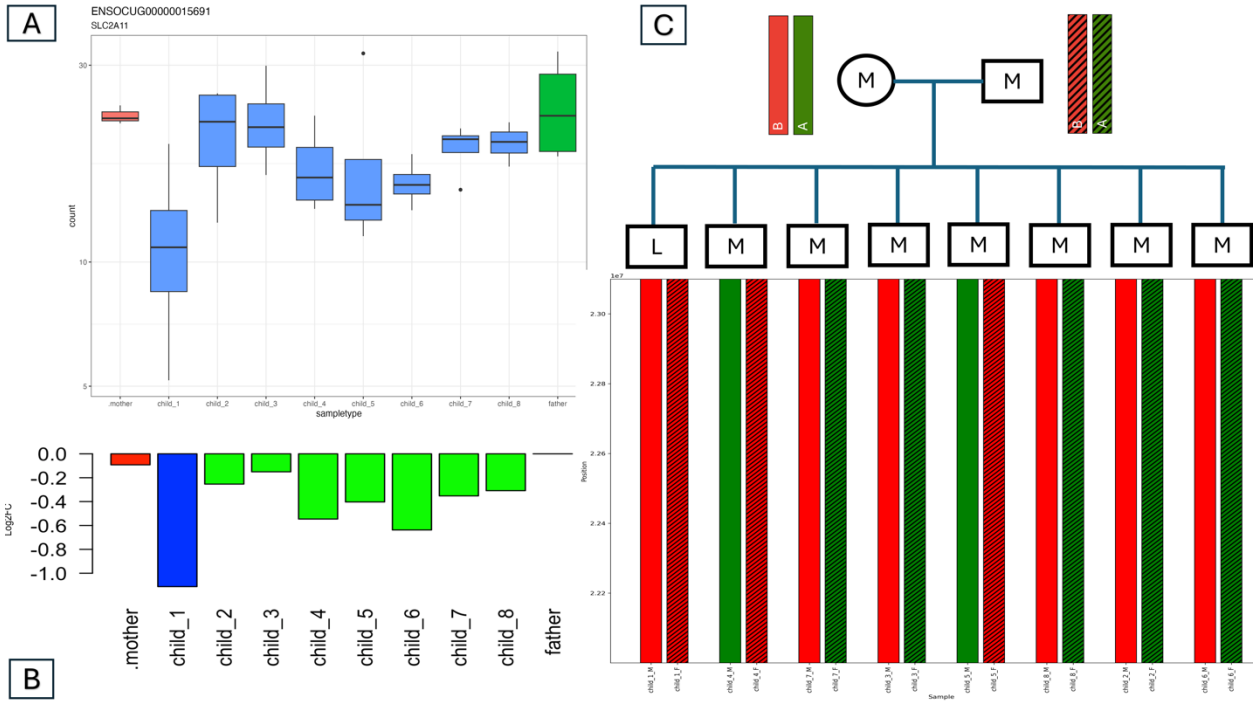


Figure 15.: SLC2A11 gene ASE analysis in the family model. A) normalized read counts at all family members, first is the mother in red and last is the father in green and children are in between. All children have an (M) expression, except for the child_1 with (L) expression. B) Log2FC comparison in the family relative to the father as the 0 baseline. Mother in red, and only child_1 with low expression (<= -0.8 log2FC). C) Extended haplotype phasing beyond the gene region. Only child_1 is Homozygous (BB) in convention with the L expression, the Heterozygosity in the rest of the children is in convention with the M expression.

5.3.5. Identification of regulatory variants

The availability of the 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 firstly essential to predict TFBSs in the rabbit genome. Since it lacks a comprehensive ChIP-seq-based TFBS genomic mapping, we utilized our human

ChIPSummitDB database (Czipa et al., 2020). In short, first, the whole genome of the human hg19 reference 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, those TFBSs were determined, which are conserved between the human and the rabbit genome. Finally, the ASE-predicted genes and their potential regulatory elements (intronic and 10kb surrounding regions) were searched for variants in the predicted TFBSs as summarized in Table 7. The entire dataset for TFBS found is available in the designated repository (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.

Table 7.: Summary of potential Transcription Factor Binding Sites (TFBS) analysis. The table lists the number of the Transcription Factors (TF) having Binding Sites overlapping with variants having the same patterns as the corresponding expression in both the Intronic region and in the 10Kb surrounding regions. The table also shows the number of genes that potentially these TF are regulating in both regions

	TF					
	H_L	L_H	M_M	H_M/M_L	L_M/H_M	Sum
Suurounding_region	24	43	11	5	14	97
Interons	33	83	0	2	7	125
	Genes					
Suurounding_region	11	22	1	3	6	43
Interons	11	30	0	1	5	47

Figure 16 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. The expression levels at this gene can be divided into two groups, the first group contains (the mother and children 3, 6, and 8) and has lower gene expression levels than the other group (father and children 1,2,4,5 and 7). 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 16, 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. This also means that this is an L_M ASE case. Interestingly, the ALT allele at the father turns the TFBS into an AGGTCAcGGTCA nuclear hormone receptor site, which now contains a perfect first-half site. We hypothesize that this variation also causes the observed higher expression, as seen in the father and the child 1,2,4,5 and 7.

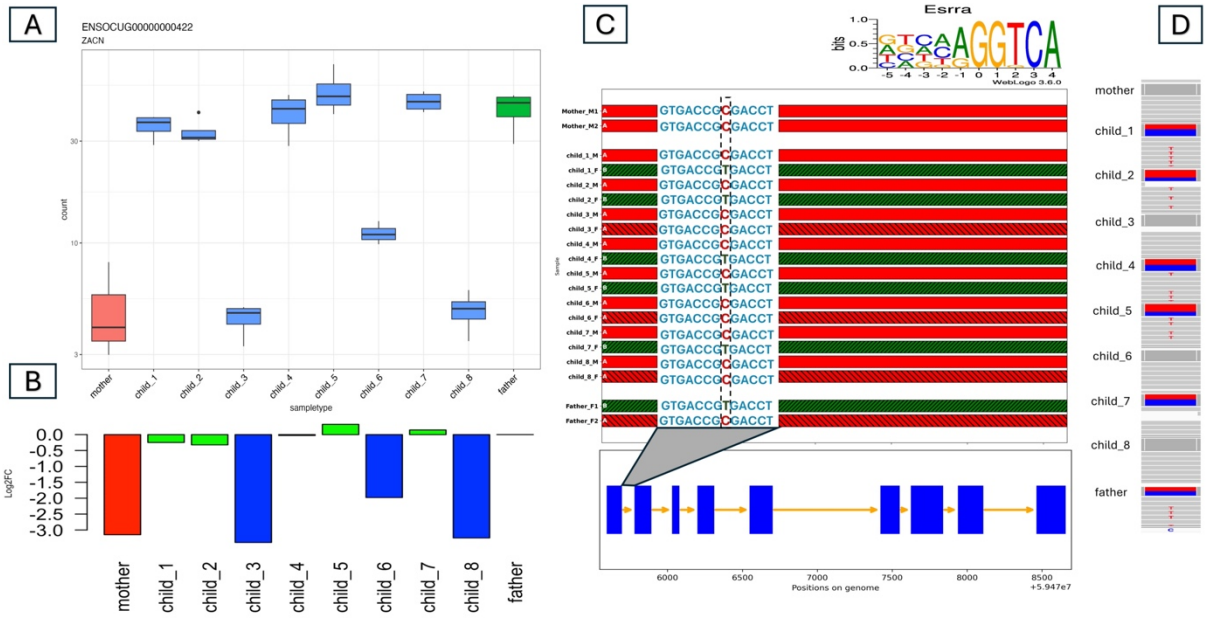


Figure 16: ZACN Gene Analysis Demonstrating an Example of L_M Expression Pattern Across Family Members. A: Normalized RNA-seq read counts with color-coded members - mother (red), father (green) and children (blue). The mother and three children (3,6, and 8) exhibit Low (L) expressions, while the father and the remaining children have Moderate (M) expressions. B: Log2 Fold Change (Log2FC) relative to father. The L-expressing individuals display Lof2FC 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 Y-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).

5.4. Discussion on ASE characterization in the family model

Allele-specific expression (ASE) occurs when the maternal and paternal alleles of a gene are unequally expressed. The presence of ASE implies that the gene is subject to differential regulatory influence between the maternal and paternal alleles, and most commonly due to a variant in cis-acting regulatory elements. These differences can affect transcription initiation, mRNA stability and abundance in an allele-specific manner.

Most key ASE findings rely on quantifying overlapping reads on a variation in a transcribed region, usually utilizing RNA sequencing samples from an F1 offspring after crossing two genetically distant parents (Lin et al., 2023; Quan et al., 2024).

However, the limitation of this approach is the necessity of at least one heterozygous site in the exonic region. Nevertheless, not all genes have exonic variants; we found about 19% of the expressed genes (2440/12659) do not have any reliable variant in RNA-seq. Additionally, the ExAC project, which genotyped 91,000 exomes, reported that 99% of variants had a frequency of less than 1%, and 50% were singletons (Lek et al., 2016). Approximately 30% of genes analyzed in the 1000 Human Genomes Project contain only one heterozygous exonic variant (Zou et al., 2024). Moreover, when the gene expression is very low, the heterozygous variants might be mistakenly called homozygous. The other common approach, expression quantitative trait loci (eQTL) mapping, which is typically performed in a population involving both RNA and WGS sequencing of a large number of individuals to identify genetic variants whose genotype patterns (for example: AA, AB, and BB) are statistically in line with the given gene's expression level (Bruscadin et al., 2022). To the best of our knowledge, rabbits have not previously been used in ASE studies. We selected this species because we found at our laboratory that the two breeds under investigation are genetically divergent and harbor a relatively high number of heterozygous variants (Fekete et al., 2025). In this work, we demonstrated that a family-based approach can detect ASE even with the absence of pre-existing variants within the transcribed regions. moreover, our method also helps in predicting putative cis-regulatory variants, as it does not require resequencing a large number of individuals.

We developed a novel pipeline for discovering and understanding the ASE phenomenon. The pipeline utilizes a combination of the WGS analysis and the RNA-seq analysis in a large family. The pipeline does not necessarily require heterozygous variants in the gene body. Instead, it relies on the gene expression (phenotype) supported by the Mendelian inheritance-based haplotype phasing of the variants that exhibit the same pattern as the gene expression on the family model. Accordingly, the gene expression can be confirmed by variants that can be either transcribed (and thus detected from the RNA-seq or WGS reads) or in the non-coding regions (intronic or the surrounding intergenic), the latter can be detected from the WGS reads only.

Based on our ChIPSummitDB, a developed human database of consensus TFBS catalogs, we determined the rabbit-conserved TFBSs. This database was developed at the University of Debrecen (Czipa et al., 2020). Cis-element effects were hypothesized in cases where a genetic variant exhibits the same inheritance pattern as a nearby gene that is in ASE status. These potential TFBSs represent promising targets for future studies focusing on meat quantity and quality in farm animals, especially in rabbit studies.

Assuming a regulatory variant causing high, low, and medium expression levels, we can theoretically expect seven different combinations of the three expected gene expression levels (L, H, and M). However, relying only on the expression data, only six distinguishable expression patterns (5+1) can be observed (Figure 11). In our family-based model, examples for all these categories were represented, categorized and reported. Among them, the most straightforward cases are where both parents are homozygous for opposite regulatory alleles (H_L or L_H), and they produce offspring with clearly intermediate expression levels that suggest regulation in cis. The M_M category, in which both parents heterozygous and exhibit similar levels of expression, is understudied due to its subtle phenotypic observation. Yet M_M still can cause highly different expression levels in children after segregation, i.e., M_M cases can provide insight into the functional impact and regulatory quality of specific alleles.

Accurately defining cis-regulatory elements with a potential influence on gene expression presents considerable challenges. Moreover, distinguishing the impact of these elements from that of the trans-element' effect is yet to be accomplished (Y. Li et al., 2015). Our family-based setup offers an extra layer of accuracy in pinpointing potentially acting cis-elements without sequencing large number of individuals, by

matching the predicted phenotypic expression patterns with the corresponding genotypic patterns. Establishing the expression patterns of the genes within a family provides an easy, yet powerful, way to identify the matching variants nearby. The current approach is a foundational step to pinpoint strong individual candidates before moving to more complex multi-variant analyses. By leveraging the conserved TFBSs, we were able to predict putative cis-acting variants. On average, 21% of predicted ASE genes were found to be potentially regulated by cis-acting elements. This suggests that the expression differences of the remaining 79% of genes can be attributed to trans-regulatory mechanisms such as variation in the abundance or activity level of a regulatory factor (e.g., a transcription factor) acting from a distant location or due to a complex multi-factor regulation that cannot be ambiguously attributed to a particular source. These findings show consistency with previous studies that showed that cis-elements in mice regulate 12-24% of genes (Crowley et al., 2015a; Goncalves et al., 2012).

H_L and L_H demonstrate the highest percentage among all expression categories potentially regulated by cis elements with 56% and 62%, respectively). On the other hand, M_M categorized genes have the lowest rate of matched variants in the TFBS with 7.3% (Table 6 summary). This observation might be attributed to the complexity of these scenarios, as well as fewer heterozygous variants shared by both parents. Here, it is important to note that in these special cases, the given gene has the same expression level in both parents, yet the offspring exhibit two or three different expression levels. Apart from the trivial cis-regulated M-M ASE cases, the underlying causes of this expression divergence are likely multifactorial involving complex interactions or stochastic effects and therefore need further investigation. It is important to note that among the 773 differentially expressed genes identified between the parents, 307 genes exhibit ASE.

5.5. De Novo Mutation Discovery and Filtration

The experiment setup consists of two genetically divergent breeding parents (mother (Hycole) as a meat producing animal and the father is a Thuringer) with their eight offspring. This experimental setup with Whole Genome Sequencing (WGS) for each individual helped us to investigate the de novo mutations (DNMs) events.

To discover DNMs in the offspring, the variant should only be present at the given child and not at the parents. However, and as a result of the sequencing errors that might lead to a positive DNM identifications, we assumed that this variant should not be also present in other siblings. Based on this idea, we developed an algorithm with strict filtering in a python script (https://github.com/Maher199/Discover_DNMs) to demonstrate the power of using a larger family in detecting DNMs. The script can be executed on a trio family i.e., the parents and only one child, as well as a larger family with several children. The script also classifies the DNMs into Single Nucleotide Variants (SNVs) and Insertions and Deletions (INDELs).

With the purpose of understanding the influence of the number of analyzed children in the pipeline, we created a combination of all possible trios and started adding siblings gradually and ran the script on the family with all combinations. Figure 17 illustrates the decline in the number of DNMs reported after adding more siblings keeping only the stringently filtered set of DNMs especially after the addition of the 3rd sibling. An immediate drop can be noticed when at least one sibling is available. Initially, when each child was treated as an only child (trio), a high number of DNMs (an average of about 4650 per child) was reported. However, gradually adding more siblings led to a decrease in false-positive variants until the number became approximately 170 per child.

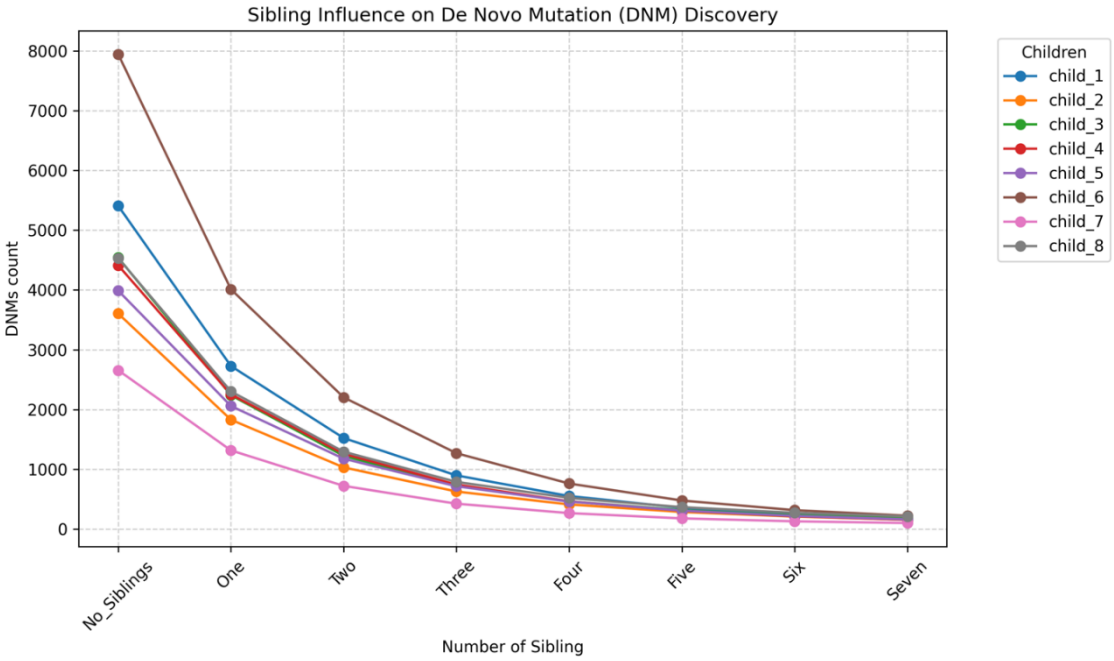


Figure 17 The impact of a larger family size on the number of reported DNMs. X-axis indicates the number of siblings starting with no sibling (i.e. trio family), while Y-axis refers to the putative DNMs reported. Lines colored by children.

After we obtained the putative DNMs based on our script leveraging the entire set of siblings, we filtered out the DNMs that fall in a repetitive region to add another layer of stringent filtering. Repetitive regions such as Short Tandem Repeats (STRs), duplications and transposons can cause misalignments especially at the short read technology. This could lead to detecting false positive DNMs. Therefore, we masked these regions to eliminate the DNMs found here. Table 8 lists the number of potential DNMs discovered using our algorithm and the remained number after removing DNMs in the masked regions. It also lists the DNMs results when no siblings were considered. On average the remained putative DNMs number is about 135 variants at each child. The latter number is divided into SNVs and INDELs as the table shows. Interestingly, for reasons yet to be determined, *Child_7* exhibits a relatively lower number of DNMs, despite all offspring and their genomic data being processed under identical conditions. More interestingly the DNMs drop in *child_6* after applying the family approach, and that is due to the false discovery rate and sequencing errors.

Table 8 Number of DNMs at each child after running the algorithm on the entire family. "After Repeat Mask" column indicates the number of DNMs that remained after eliminating the variants that fall in the masked regions. The final number of DNMs was classified as SNVs and INDELs.

Children	No siblings	DNMs	After Repeat Mask	SNVs	INDELs
Child_1	5408	175	139	51	88
Child_2	3605	156	132	53	79
Child_3	4541	174	143	51	92
Child_4	4412	154	135	55	80
Child_5	3988	162	117	43	74
Child_6	7945	222	178	76	102
Child_7	2656	100	80	40	40
Child_8	4534	210	154	87	67

5.5.1. De Novo Mutation Hotspots

In order to further investigate the behavior of the DNMs, we carried out the discovered DNMs analysis after filtering the variants in the masked region to investigate for DNMs hotspots. DNMs hotspots are regions in the genome prone to have more mutations comparing to other regions. After filtering out the mutations that fell in the masked region, we characterized the DNMs by chromosomes. Figure 18 represents these

hotspots, where these regions tend to produce more mutations than other regions on the same chromosome. The histogram on the chromosomes represents the density of the DNMs, and it is interesting how these variants are clustered, and the majority are located near the telomeres.

Figure 18 The Density Distribution of DNMs along the chromosomes. A: The entire set of DNMs at all children along all the set of chromosomes. The plot shows the density of DNMs within a 2M bp window. Created Using KaryoploteR (Gel & Serra, 2017).

5.5.2. De Novo Mutations Base Substitution

The last step in the analysis was an investigation for the distribution of the base types of substitution across the SNVs. Base substitution in the SNVs such as (C =>T or its complementary G => A) etc. is usually classified into two groups. Transition (Ts) which includes (purine => purine or pyrimidine => pyrimidine) and Transversion (Tv) including (purine => pyrimidine or pyrimidine => purine). Figure 19 illustrates the base substitutions count for the DNMs discovered. The substitution categorized as it is colored by the type of substitution (Ts or Tv). We found that the most frequent type of base substitution in our discovered DNMs was (Ts) while the least were (Tv) with the ratio (Ts/Tv) of ~2.8.

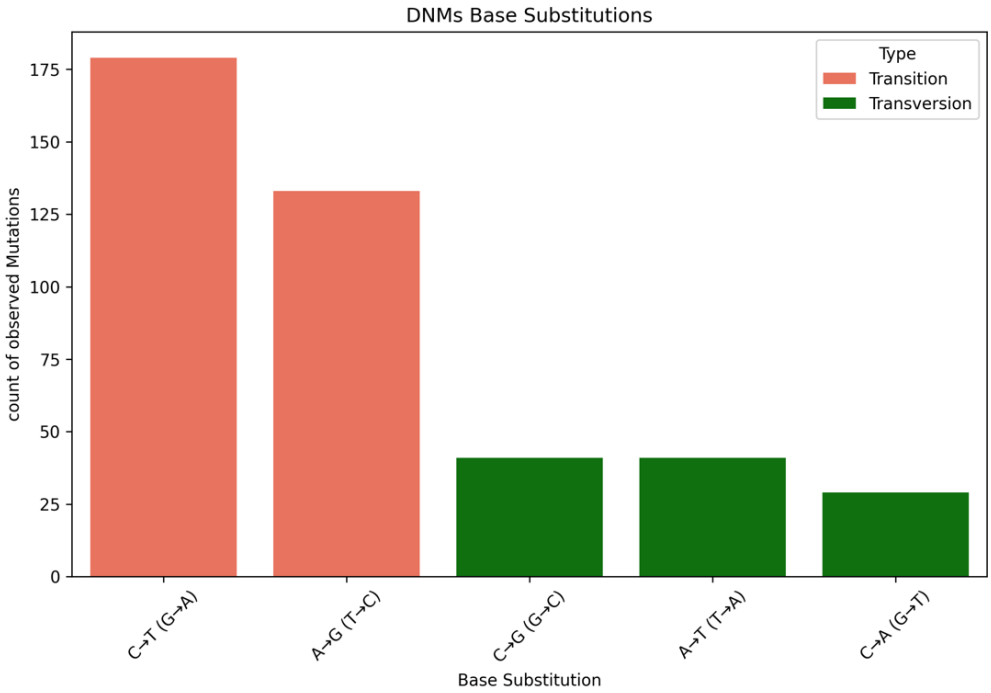


Figure 19 Distribution of SNV Base Substitutions: Transition vs. Transversion. X-axis is base substitution type (and its complementary) categorized by the substitution type (red for Transition and green for Transversion). Y-axis is the occurrence count of observed Single Nucleotide DNMs.

5.6. De Novo Mutations Discussion

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, rabbit was 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.

Despite the fact that the mutation is a random event, its occurrence and distribution along the genome is not completely random and there will be some sites with higher probability of mutations (Supek & Lehner, 2019). These sites are what is referred to as DNM hotspots and it is where the new mutations cluster, where these mutations tend to be closer to each other

comparing with other sites (Keightley et al., 2014). We found these DNM hotspots, and we reported their occurrence especially in sites closer to the telomeres. However, further investigations are necessary in this area. Here we also identified the DNMs base substitution types the Transitions (Ts) and the Transversions (Tv). We found the ratio of the (Ts/Tv) is about 2.8. It is consistent with several previous studies that investigated the same ratio in humans (Conrad et al., 2011), and in monkeys (Bergeron et al., 2021). This ratio is important since it might indicate evolutionary or environmental signatures.

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.

6 CONCLUSIONS AND RECOMMENDATIONS

Allele-Specific Expression occurs when the maternal and paternal copies of genes are unequally expressed. The presence of ASE implies that the given gene is subject to genetically determined, differential regulatory control between the maternal and paternal alleles, often due to cis-acting variants, although trans effects can also contribute. ASE can be utilized to differentiate the cis-acting elements influence from the trans mechanism. ASE also is an important phenomenon in detecting rare variants, that may not be captured by eQTL studies and can influence the phenotype of several diseases and traits. ASE has been reported in several studies on farm animals related to muscle growth and meat quality and production.

Most key ASE findings rely on counting the ratio of the overlapping reads at a variant in a transcribed region (Lin et al., 2023; Quan et al., 2024). The limitation of this approach is the necessity of at least one heterozygous site to be present in the transcribed region and, therefore, sequenced region. However, not all genes have Exonic variants, we found about 19% of the expressed genes (2440/12659) do not have any reliable variant in RNA-seq. In this study, we expanded the detection of ASE beyond the need for a heterozygous variant in the transcript. By employing a family-based approach one can detect ASE even in the absence of variants in the exonic regions of the gene and instead comparing the expression level of the mRNA at each individual of the family to infer ASE.

The importance of eQTLs studies is well recognized, however it also comes with some drawbacks. eQTLs are genomic regions containing regulatory DNA variation that shape gene expression. This mechanism is an important way DNA variants connected to complex traits by altering the expression of critical genes. Many studies utilized eQTLs to investigate the relationship between gene expression, as influenced by eQTLs, and genetically complex growth traits (Renganaath & Albert, 2023). However, our method helps identifying putative causal regulatory variants in the TFBS without resequencing a large number of individuals as it is considered the disadvantage of conducting eQTLs studies.

We provided a novel analytical pipeline that can potentially pinpoint putative causal variants in the TFBS that might underlie the ASE at a specific region by matching the gene expression pattern across the family with the genotype pattern of the variant in the potential TFBS. This also adds an extra layer of filtering out the false variants from the analysis when they do not comply the family members grouping.

On average, we found that 21% of predicted ASE genes were potentially regulated by cis-acting elements, suggesting that the remaining 79% of genes may have been caused by trans-regulatory factors or by a complex multi-layered regulatory mechanism (interaction between cis and trans regulatory elements). Our findings are consistent with previous studies that estimated the cis-regulatory elements in mice regulate 12-24% of genes.

In the current study, we applied the introduced pipeline on a hybrid rabbit family and eventually detected 913 genes that exhibit ASE and comply with the intermediate law of inheritance. These genes were assigned to their categories of expression patterns suggested in this study. These ASE genes suggest the involvement of cis-regulatory elements in the nearby regions. Using the whole genome alignment between the human H19 and the rabbit OryCun 3.0 reference genomes, we mapped the conserved consensus transcription factor binding site sets from our ChIP-SummitDB to the rabbit reference. To find the potential regulatory SNPs, we intersected these conserved TFBSs with the variations found in and around the ASE genes and showed the same inheritance pattern than the ASE. This analysis resulted in 222 conserved TFBS potentially regulating 90 ASE genes.

To our knowledge, at the time of the analysis, this study utilized Rabbits for the first time to explore and characterize the ASE phenomenon. We also conducted DEGs between the divergent parents and provided some instances of genes that might be related to meat production in rabbits. Some of these reported genes were novel genes while others were reported in previous studies in pigs and other farm animals. Among the 773 differentially expressed genes identified between the parents, 307 genes exhibit ASE.

We believe that this pipeline can be further fine-tuned by integrating long reads technology and implementing a reciprocal crossing to eliminate the parent-of-origin influence. Future work could also involve using machine learning models that integrate multiple variant effects. In addition, TFBS variants that were reported in this study can serve as a base-reference and valuable resource for future investigations into cis-regulatory elements regulating muscle development and growth traits in rabbits.

De novo mutations are new genetic changes that occur spontaneously in an individual and are not inherited from either parent. DNMs were also investigated in rabbits for the first time. We developed a versatile Python script to identify DNMs, highlighting how extended family data enhances analytical precision. We believe that adopting long-read sequencing technologies will significantly improve DNM detection, particularly in repetitive genomic regions. On average, the final DNMs number was about 135 variants at each child after introducing the eight

offspring. However, further research is essential to better understand DNM hotspots and their functional consequences.

7 NEW SCIENTIFIC RESULTS

1. Utilizing rabbits for characterizing 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 amount in farm animals.
4. Developed and reported a 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 the DNMs hotspots.

8 SUMMARY

The precise identification of regulatory elements, especially cis-elements, that control gene expression remains a challenging quest in molecular biology. Allele-specific expression (ASE), which is a widespread phenomenon, can be utilized to infer the existence of heterozygous variants within the cis-regulatory elements, particularly in the transcription factor binding site (TFBS). The requirement of a heterozygous variant to be present in the transcript and the vast number of individuals necessary to conduct an expression quantitative trait locus (eQTL) analysis contribute to the complexity of resolving the ASE phenomenon. In this study, we introduce a novel pipeline aiming to characterize the ASE phenomenon based on a family model. The pipeline utilizes a combination of RNA-seq and WGS data, and it is not limited by the presence of heterozygous variants in the gene's exonic region but instead relies on the intermediate level of gene expression matched with haplotype phasing-based of variants at any region (gene region or surrounding regions). We applied the pipeline on a hybrid rabbit family, coming from two genetically divergent parents to study ASE and report potential variants in the conserved TFBS. We identified 913 genes that exhibit ASE, with respect to their category of expression patterns, and reported 222 conserved TFBS potentially regulating 90 genes. We found that 21% of these ASE are potentially regulated by cis-acting elements. Our approach offers a comprehensive pipeline for characterizing ASE genes in a large family enabling both the identification of ASE genes and pinpointing of putative cis-acting regulatory elements. In addition to providing a detailed ASE analysis, we identified genes that might be responsible for the increased meat production and quality in the Hycole rabbit breed. We further utilized extended family structures to enhance the detection of de novo mutations (DNMs), assessing how the inclusion of additional siblings influences analytical accuracy. Eventually, we provided a python script that is capable of identifying DNMs in a family with any size. DNM hotspots, i.e. the genomic regions that are more prone to produce these mutations, were analysed and reported in rabbits for the first time.

ÖSSZEFOGLALÓ

A génexpressziót szabályozó elemek, különösen a cis-szabályozó elemek pontos azonosítása továbbra is jelentős kihívást jelent a molekuláris biológiában. Az allél-specifikus expresszió (ASE) egy széles körben elterjedt jelenség, amely lehetőséget nyújt a heterozigóta variánsok jelenlétének kimutatására cis-szabályozó régiókban, különösen a transzkripciós faktor kötőhelyeken (TFBS). Az ASE vizsgálatának összetettségét fokozza, hogy a heterozigóta variánsnak meg kell jelennie a transzkriptumban, valamint, hogy nagyszámú egyed szükséges az eQTL-elemzés elvégzéséhez.

Ebben a tanulmányban egy új pipeline-t mutatunk be, amely a családi modell alapján jellemzi az ASE-jelenséget. A pipeline RNA-seq és WGS adatok kombinációját használja, és nem korlátozódik a gén exonos régióiban található heterozigóta variánsokra, hanem a génexpresszió közepes szintjét és haplotípusfázis alapján történő variánsokkal való megfeleltetést alkalmaz bármely régióban (gén vagy környező régiók).

A módszert egy hibrid nyúlcsaládon alkalmaztuk, amely két genetikailag eltérő szülő keresztezéséből származik, az ASE tanulmányozása és a konzervált TFBS-ekben található potenciális variánsok feltárása céljából. Összesen 913 ASE-t mutató gént azonosítottunk, amelyek különböző expressziós mintázatokkal rendelkeznek, és 222 konzervált TFBS-t jelentettünk, amelyek várhatóan 90 gén expresszióját szabályozzák. Megállapítottuk, hogy ezen ASE-gének 21%-át cis-hatású elemek szabályozzák.

Módszerünk egy átfogó pipeline-t kínál nagy családokon belüli ASE gének jellemzésére, amely lehetővé teszi az ASE-gének azonosítását és a putatív cis-szabályozó elemek feltérképezését. Az ASE részletes elemzésén túlmenően potenciális géneket is azonosítottunk, amelyek a Hycole nyúlfajta húsminőségének és termelésének növekedéséért felelősek.

Továbbá kiterjesztett családszerkezetet használtunk a de novo mutációk (DNM-ek) kimutatásának javítására, vizsgálva, hogyan befolyásolja a további testvérek bevonása az elemzés pontosságát. Végül egy olyan Python szkriptet biztosítottunk, amely bármilyen méretű család esetén képes DNM-ek azonosítására. A DNM-hotspotokat, vagyis azokat a genomi régiókat, amelyek hajlamosabbak mutációk kialakulására, elsőként elemeztük és jelentettük nyúlban.

الملخص

تُعدّ عملية التحديد الدقيق للعناصر التنظيمية، وبالأخص العناصر الموجودة في نفس الموضع الجيني التي تتحكم في التعبير الجيني، من التحديات المستمرة في علم الأحياء الجزيئي. يمكن الاستفادة من ظاهرة التعبير الخاص بالأليل، وهي ظاهرة واسعة الانتشار، للاستدلال على وجود متغيرات غير متماثلة الزيجوت ضمن العناصر التنظيمية الموضعية، وخاصة مواقع ارتباط عوامل النسخ. تتطلب دراسة هذه الظاهرة وجود متغيّر غير متماثل في النسخة المنسوخة من الجين، كما تتطلّب عدداً كبيراً من الأفراد لإجراء تحليل العلاقة الكميّة لتعبير الجين، مما يزيد من تعقيد دراسة هذه الظاهرة.

في هذه الدراسة، نقدم نموذجاً تحليليّاً مبتكراً يهدف إلى توصيف ظاهرة استناداً إلى نموذج عائليز يستخدم هذا الإطار أو النوذج بيانات التسلسل الكامل للحمض النووي وبيانات تسلسل الحمض النووي الريبي، ولايقتصر فقط على وجود متغيرات غير متماثلة في المناطق الإكزونية من الجينات، بل يعتمد على مستوى التعبير الجيني الوسيط المرتبط بتحديد الطور للمتغيرات في أي منطقة (ضمن الجين أو المناطق المحيطة به).

طبقنا هذا الأنبوب التحليلي على عائلة أرانب هجينة ناتجة عن أبوين مختلفين وراثيّاً لدراسة ظاهرة التعبير الخاص بالأليل جيناً تظهر تعبير 913وتحديد المتغيرات المحتملة ضمن مواقع ارتباط عوامل النسخ المحفوظة تطوريّاً. تمكنّا من تحديد 21 % جيناً. ووجدنا أنّ 90 موقعاً محفوظاً لعوامل النسخ يحتمل أنها تنظم 222خاصاً بالأليل، وفقاً لأنماط تعبيرها، وبلغنّا عن من هذه الجينات التي تظهر التعبير الخاص بالأليل يتمّ تنظيمها بواسطة عناصر تنظيميّة موضعيّة.

يوفر هذا النهج أو الإطار تحليلاً شاملاً لتوصيف جينات التعبير الخاص ضمن عائلة كبيرة، مما يتيح تحديد الجينات الخاضعة لتأثير هذا التعبير الخاص، وكذلك العناصر التنظيميّة الموضعيّة المحتملة التي تتحكم في تعبيرها. بالإضافة إلى التحليل التفصيلي لهذه الظاهرة، حددنا جينات محتملة مسؤولة عن زيادة انتاج اللحم في سلالة أرانت هاكول.

كما استخدمنا بنى عائلية موسعة لتحسين اكتشاف الطفرات الجديدة، وقمنا بتقييم كيفية تأثير ادخال اخوة اضافيين على دقة التحليل. في النهاية، قدمنا سكريبت بلغة بايثون قادر على تحديد هذه الطفرات الحديثة ضمن عائلة بأي حجم. وقد قمنا بتحليل مناطق الطفرات الساخنة، أي المناطق الجينومية الأكثر عرضة لهذه الطفرات، ولأول مرة تم الإبلاغ عنها في الأرانب.

Acknowledgment

This work was funded by the Hungarian Government Organization NRDI (National Research, Development, and Innovation Office: https://nkfih.gov.hu/about-the-office) through the grant NKFI/2017-1.3.1-VKE-2017-00026 and NKFI/OTKA K 132814. Project no. TKP2021-NKTA-34 awarded to Dr. Endre Barta. The resulting scientific articles were funded by the University of Debrecen.

I am grateful to the Tempus Foundation and the Stipendium Hungaricum Scholarship for financially supporting my master's and PhD studies in Hungary.

I would like to extend my thanks to the following people who have helped me undertake this research.

First and foremost, my deepest thanks and appreciation go to my supervisor, Dr. Endre Barta, who introduced me to the fields of bioinformatics and genomics, and who has supported me through years of knowledge transfer and close supervision of my work.

To my colleagues at the Genetics and Genomics Group, with whom I worked during this PhD research.

To the logistic support of my work at the Genetics and Biotechnology Institute, Ms. Beatrix Pethőné Rétháti, Ms. Ágnes Pégerné.

Special thanks go to Ms. Tassy Zsuzsanna for her continuous support during my years in Hungary.

To my friends and the music group for being there for me.

Finally, to my family and dear parents and siblings.

APPENDICES

A1 Appendix: Bibliography

Abascal, F., Acosta, R., Addleman, N. J., Adrian, J., Afzal, V., Aken, B., Akiyama, J. A., Jammal, O. Al, Amrhein, H., Anderson, S. M., Andrews, G. R., Antoshechkin, I., Ardlie, K. G., Armstrong, J., Astley, M., Banerjee, B., Barkal, A. A., Barnes, I. H. A., Barozzi, I., … Zimmerman, J. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, *583*(7818), 699–710. https://doi.org/10.1038/s41586-020-2493-4

Acuna-Hidalgo, R., Veltman, J. A., & Hoischen, A. (2016). New insights into the generation and role of de novo mutations in health and disease. In *Genome Biology* (Vol. 17, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-016-1110-1

Amoah, K., Hsiao, Y. H. E., Bahn, J. H., Sun, Y., Burghard, C., Tan, B. X., Yang, E. W., & Xiao, X. (2021). Allele-specific alternative splicing and its functional genetic variants in human tissues. *Genome Research*, *31*(3), 359–371. https://doi.org/10.1101/GR.265637.120

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

Arana, M. E., & Kunkel, T. A. (2010). Mutator phenotypes due to DNA replication infidelity. In *Seminars in Cancer Biology* (Vol. 20, Issue 5, pp. 304–311). https://doi.org/10.1016/j.semcancer.2010.10.003

Arnold, M., & Stengel, K. R. (2023). Emerging insights into enhancer biology and function. In Transcription (Vol. 14, Issues 1–2, pp. 68–87). Taylor and Francis Ltd. https://doi.org/10.1080/21541264.2023.2222032

Azevedo, L., Amaro, A. P., Niza-Ribeiro, J., & Lopes-Marques, M. (2024). Naturally occurring genetic diseases caused by de novo variants in domestic animals. In *Animal Genetics* (Vol. 55, Issue 3, pp. 319–327). John Wiley and Sons Inc. https://doi.org/10.1111/age.13403

Babak, T., Deveale, B., Tsang, E. K., Zhou, Y., Li, X., Smith, K. S., Kukurba, K. R., Zhang, R., Li, J. B., Van Der Kooy, D., Montgomery, S. B., & Fraser, H. B. (2015). Genetic

conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nature Genetics*, *47*(5), 544–549. https://doi.org/10.1038/ng.3274

Bader, D. M., Wilkening, S., Lin, G., Tekkedil, M. M., Dietrich, K., Steinmetz, L. M., & Gagneur, J. (2015). Negative feedback buffers effects of regulatory variants. *Molecular Systems Biology*, *11*(1). https://doi.org/10.15252/msb.20145844

Ban, N., Nissen, P., Hansen, J., Moore, P. B., & Steitz, T. A. (n.d.). *The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 Å Resolution*. http://science.sciencemag.org/

Banerji, J., Olson, L., & Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, *33*(3), 729–740. https://doi.org/10.1016/0092-8674(83)90015-6

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2), 299–308. https://doi.org/10.1016/0092-8674(81)90413-X

Baruzzo, G., Hayer, K. E., Kim, E. J., DI Camillo, B., Fitzgerald, G. A., & Grant, G. R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods*, *14*(2), 135–139. https://doi.org/10.1038/nmeth.4106

Berest, I., Arnold, C., Reyes-Palomares, A., Palla, G., Rasmussen, K. D., Giles, H., Bruch, P. M., Huber, W., Dietrich, S., Helin, K., & Zaugg, J. B. (2019). Quantification of Differential Transcription Factor Activity and Multiomics-Based Classification into Activators and Repressors: diffTF. *Cell Reports*, *29*(10), 3147-3159.e12. https://doi.org/10.1016/j.celrep.2019.10.106

Bergeron, L. A., Besenbacher, S., Bakker, J., Zheng, J., Li, P., Pacheco, G., Sinding, M. H. S., Kamilari, M., Gilbert, M. T. P., Schierup, M. H., & Zhang, G. (2021). The germline mutational process in rhesus macaque and its implications for phylogenetic dating. *GigaScience*, *10*(5). https://doi.org/10.1093/gigascience/giab029

Bergeron, L. A., Besenbacher, S., Zheng, J., Li, P., Bertelsen, M. F., Quintard, B., Hoffman, J. I., Li, Z., St. Leger, J., Shao, C., Stiller, J., Gilbert, M. T. P., Schierup, M. H., & Zhang, G. (2023). Evolution of the germline mutation rate across vertebrates. *Nature*, *615*(7951), 285–291. https://doi.org/10.1038/s41586-023-05752-y

Björn N, & Sahlén. (2018). Next Generation Sequencing & Applications Comparison of Variant Calls from Whole Genome and Whole Exome Sequencing Data Using Matched Samples. *Next Generat Sequenc & Applic*, *5*, 1. https://doi.org/10.4172/2469-9853.1000154

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

Bonasio, R., Tu, S., & Reinberg, D. (2010). Molecular signals of epigenetic states. In *Science* (Vol. 330, Issue 6004, pp. 612–616). https://doi.org/10.1126/science.1191078

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J. P., Tanay, A., & Cavalli, G. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell*, *171*(3), 557-572.e24. https://doi.org/10.1016/j.cell.2017.09.043

Bonnot, T., Blair, E. J., Cordingley, S. J., & Nagel, D. H. (2021). Circadian coordination of cellular processes and abiotic stress responses. In *Current Opinion in Plant Biology* (Vol. 64). Elsevier Ltd. https://doi.org/10.1016/j.pbi.2021.102133

book_enhancers. (n.d.).

Bookshelf, N., Anaya, ;, Shoenfeld, J. M., & Rojas-Villarraga, Y. (2013). *A service of the National Library of Medicine, National Institutes of Health*. El Rosario University Press. https://www.ncbi.nlm.nih.gov/books/NBK459456/

Brown, J. C. (2018). Control of human gene expression: High abundance of divergent transcription in genes containing both INR and BRE elements in the core promoter. *PLoS ONE*, *13*(8). https://doi.org/10.1371/journal.pone.0202927

Bruscadin, J. J., Cardoso, T. F., da Silva Diniz, W. J., de Souza, M. M., Afonso, J., Vieira, D., Malheiros, J., Andrade, B. G. N., Petrini, J., Ferraz, J. B. S., Zerlotini, A., Mourão, G. B., Coutinho, L. L., & de Almeida Regitano, L. C. (2022). Differential Allele-Specific Expression Revealed Functional Variants and Candidate Genes Related to Meat Quality Traits in B. indicus Muscle. *Genes*, *13*(12). https://doi.org/10.3390/genes13122336

Bruscadin, J. J., de Souza, M. M., de Oliveira, K. S., Rocha, M. I. P., Afonso, J., Cardoso, T. F., Zerlotini, A., Coutinho, L. L., Niciura, S. C. M., & de Almeida Regitano, L. C. (2021). Muscle allele-specific expression QTLs may affect meat quality traits in Bos indicus. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-86782-2

Bullini, L. (n.d.). *Origin and e animal hybrid species mass production of animal proteins)*.

Cavalli, M., Baltzer, N., Umer, H. M., Grau, J., Lemnian, I., Pan, G., Wallerman, O., Spalinskas, R., Sahlén, P., Grosse, I., Komorowski, J., & Wadelius, C. (2019). Allele specific chromatin signals, 3D interactions, and motif predictions for immune and B cell related diseases. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-39633-0

central dogma. (n.d.).

Chamberlain, A. J., Vander Jagt, C. J., Hayes, B. J., Khansefid, M., Marett, L. C., Millen, C. A., Nguyen, T. T. T., & Goddard, M. E. (2015). Extensive variation between tissues in allele specific expression in an outbred mammal. *BMC Genomics*, *16*(1). https://doi.org/10.1186/s12864-015-2174-0

Chen, S. (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *IMeta*, *2*(2), e107. https://doi.org/https://doi.org/10.1002/imt2.107

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

Choi, Y., Chan, A. P., Kirkness, E., Telenti, A., & Schork, N. J. (2018). Comparison of phasing strategies for whole human genomes. *PLoS Genetics*, *14*(4). https://doi.org/10.1371/journal.pgen.1007308

Cleary, S., & Seoighe, C. (2021). Annual Review of Biomedical Data Science Perspectives on Allele-Specific Expression. Annu. Rev. Biomed. Data Sci, 2021, 101–122. https://doi.org/10.1146/annurev-biodatasci-021621

Cleary, S., & Seoighe, C. (2025). Perspectives on Allele-Specific Expression. *Annual Review of Biomedical Data Science Downloaded from Www.Annualreviews.Org. Guest*, *28*, 6. https://doi.org/10.1146/annurev-biodatasci-021621

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Research*, *38*(6), 1767–1771. https://doi.org/10.1093/nar/gkp1137

Combes, M. C., Hueber, Y., Dereeper, A., Rialle, S., Herrera, J. C., & Lashermes, P. (2015). Regulatory divergence between parental alleles determines gene expression patterns in hybrids. *Genome Biology and Evolution*, *7*(4), 1110–1121. https://doi.org/10.1093/gbe/evv057

Conrad, D. F., Keebler, J. E. M., Depristo, M. A., Lindsay, S. J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C. L., Torroja, C., Garimella, K. V., Zilversmit, M., Cartwright, R., Rouleau, G. A., Daly, M., Stone, E. A., Hurles, M. E., & Awadalla, P. (2011). Variation in genome-wide mutation rates within and between human families. *Nature Genetics*, *43*(7), 712–714. https://doi.org/10.1038/ng.862

Cooper, Geoffrey M. 2000. *The Cell: A Molecular Approach*. 2nd ed. Sunderland, MA: Sinauer Associates.

Cramer, P. (2019). Organization and regulation of gene transcription. In *Nature* (Vol. 573, Issue 7772, pp. 45–54). Nature Publishing Group. https://doi.org/10.1038/s41586-019-1517-4

Crowley, J. J., Zhabotynsky, V., Sun, W., Huang, S., Pakatci, I. K., Kim, Y., Wang, J. R., Morgan, A. P., Calaway, J. D., Aylor, D. L., Yun, Z., Bell, T. A., Buus, R. J., Calaway, M. E., Didion, J. P., Gooch, T. J., Hansen, S. D., Robinson, N. N., Shaw, G. D., … De Villena, F. P. M. (2015). Analyses of allele-specific gene expression in highly divergent mouse crosses identifies pervasive allelic imbalance. Nature Genetics, 47(4), 353–360. https://doi.org/10.1038/ng.3222

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

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

de Souza, M. M., Zerlotini, A., Rocha, M. I. P., Bruscadin, J. J., Diniz, W. J. da S., Cardoso, T. F., Cesar, A. S. M., Afonso, J., Andrade, B. G. N., Mudadu, M. de A., Mokry, F. B., Tizioto, P. C., de Oliveira, P. S. N., Niciura, S. C. M., Coutinho, L. L., & Regitano, L. C. de A. (2020a). Allele-specific expression is widespread in Bos indicus muscle and affects meat quality candidate genes. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-67089-0

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., & Daly, M. J. (2011a). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 43(5), 491–498. https://doi.org/10.1038/ng.806

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner.

Bioinformatics (Oxford, England), *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dror, I., Golan, T., Levy, C., Rohs, R., & Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Research*, *25*(9), 1268–1280. https://doi.org/10.1101/gr.184671.114

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

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

Flanagan, J. T., Nam, K., & Lee, S. (2024). *ENCODE guided WGS analysis can identify trait associated regulatory regions driven by rare-variants*. https://doi.org/10.1101/2024.11.06.24316407

Friedberg, E. C. (2003). DNA damage and repair. In *feature 436 NATURE* | (Vol. 421). www.nature.com/nature

Garg, S. (2021). Computational methods for chromosome-scale haplotype reconstruction. In Genome Biology (Vol. 22, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-021-02328-9

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

Gel, B., & Serra, E. (2017). KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics*, *33*(19), 3088–3090. https://doi.org/10.1093/bioinformatics/btx346

Ghedira, K. (2018). Introductory Chapter: A Brief Overview of Transcriptional and Post-transcriptional Regulation. In *Transcriptional and Post-transcriptional Regulation*. InTech. https://doi.org/10.5772/intechopen.79753

Gompel, N., Prud'homme, B., Wittkopp, P. J., Kassner, V. A., & Carroll, S. B. (2005). Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. www.nature.com/nature

Goncalves, A., Leigh-Brown, S., Thybert, D., Stefflova, K., Turro, E., Flicek, P., Brazma, A., Odom, D. T., & Marioni, J. C. (2012). Extensive compensatory cis-trans regulation

in the evolution of mouse gene expression. *Genome Research*, *22*(12), 2376–2384. https://doi.org/10.1101/gr.142281.112

Gregg, C., Zhang, J., Weissbourd, B., Luo, S., Schroth, G. P., Haig, D., & Dulac, C. (n.d.). *High Resolution Analysis of Parent-of-Origin Allelic Expression in the Mouse Brain*.

Guan, P., & Sung, W. K. (2016). Structural variation detection using next-generation sequencing data: A comparative technical review. In *Methods* (Vol. 102, pp. 36–49). Academic Press Inc. https://doi.org/10.1016/j.ymeth.2016.01.020

Guillocheau, G. M., El Hou, A., Meersseman, C., Esquerré, D., Rebours, E., Letaief, R., Simao, M., Hypolite, N., Bourneuf, E., Bruneau, N., Vaiman, A., Vander Jagt, C. J., Chamberlain, A. J., & Rocha, D. (2019). Survey of allele specific expression in bovine muscle. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-40781-6

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., & Young, R. A. (2013). XSuper-enhancers in the control of cell identity and disease. *Cell*, *155*(4), 934. https://doi.org/10.1016/j.cell.2013.09.053

Ho, C. K., Cui, X., Grubner, S., Larson, C. A., Wei, Y., & Flook, P. K. (2016). Whole-genome sequencing analysis using next-generation sequencing data. *Current Protocols in Essential Laboratory Techniques*, *2016*, 11.5.1-11.5.20. https://doi.org/10.1002/cpet.2

Inukai, S., Kock, K. H., & Bulyk, M. L. (2017). Transcription factor–DNA binding: beyond binding site motifs. In *Current Opinion in Genetics and Development* (Vol. 43, pp. 110–119). Elsevier Ltd. https://doi.org/10.1016/j.gde.2017.02.007

Keightley, P. D., Ness, R. W., Halligan, D. L., & Haddrill, P. R. (2014). Estimation of the spontaneous mutation rate per nucleotide site in a Drosophila melanogaster full-sib family. *Genetics*, *196*(1), 313–320. https://doi.org/10.1534/genetics.113.158758

Khansefid, M., Pryce, J. E., Bolormaa, S., Chen, Y., Millen, C. A., Chamberlain, A. J., Vander Jagt, C. J., & Goddard, M. E. (2018). Comparing allele specific expression and local expression quantitative trait loci and the influence of gene expression on complex trait variation in cattle. *BMC Genomics*, *19*(1). https://doi.org/10.1186/s12864-018-5181-0

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N. I., Yurovsky, A., Lappalainen, T., Romano-Palumbo, L., Planchon, A., Bielser, D., Bryois, J., Padioleau, I., Udin, G., Thurnheer, S., … Dermitzakis, E. T. (2013). Coordinated

effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science*, *342*(6159), 744–747. https://doi.org/10.1126/science.1242463

Kim-Hellmuth, S., Bechheim, M., Pütz, B., Mohammadi, P., Nédélec, Y., Giangreco, N., Becker, J., Kaiser, V., Fricker, N., Beier, E., Boor, P., Castel, S. E., Nöthen, M. M., Barreiro, L. B., Pickrell, J. K., Müller-Myhsok, B., Lappalainen, T., Schumacher, J., & Hornung, V. (2017). Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-00366-1

Kohailan, M., Aamer, W., Syed, N., Padmajeya, S., Hussein, S., Sayed, A., Janardhanan, J., Palaniswamy, S., El hajj, N., Al-Shabeeb Akil, A., & Fakhro, K. A. (2022). Patterns and distribution of de novo mutations in multiplex Middle Eastern families. *Journal of Human Genetics*, *67*(10), 579–588. https://doi.org/10.1038/s10038-022-01054-9

Kuhlman, T. C., Cho, H., Reinberg, D., & Hernandez, A. N. (1999). The General Transcription Factors IIA, IIB, IIF, and IIE Are Required for RNA Polymerase II Transcription from the Human U1 Small Nuclear RNA Promoter. In *MOLECULAR AND CELLULAR BIOLOGY* (Vol. 19, Issue 3). http://mcb.asm.org/

Kumasaka, N., Knights, A. J., & Gaffney, D. J. (2016). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics*, *48*(2), 206–213. https://doi.org/10.1038/ng.3467

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., … Williams, A. L. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Li, X., Kim, Y., Tsang, E. K., Davis, J. R., Damani, F. N., Chiang, C., Hess, G. T., Zappala, Z., Strober, B. J., Scott, A. J., Li, A., Ganna, A., Bassik, M. C., Merker, J. D., Aguet, F., Ardlie, K. G., Cummings, B. B., Gelfand, E. T., Getz, G., … Montgomery, S. B. (2017). The impact of rare variation on gene expression across tissues. *Nature*, *550*(7675), 239–243. https://doi.org/10.1038/nature24267

Li, Y., Chen, C. yu, Kaye, A. M., & Wasserman, W. W. (2015). The identification of cis-regulatory elements: A review from a machine learning perspective. In *BioSystems* (Vol. 138, pp. 6–17). Elsevier Ireland Ltd. https://doi.org/10.1016/j.biosystems.2015.10.002

Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

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

Liu, D., Zhang, H., Yang, Y., Liu, T., Guo, Z., Fan, W., Wang, Z., Yang, X., Zhang, B., Liu, H., Tang, H., Yu, D., Yu, S., Gai, K., Mou, Q., Cao, J., Hu, J., Tang, J., Hou, S., & Zhou, Z. (2023). Metabolome-Based Genome-Wide Association Study of Duck Meat Leads to Novel Genetic and Biochemical Insights. *Advanced Science*, *10*(18). https://doi.org/10.1002/advs.202300148

Liu, X., Chen, M., Qu, X., Liu, W., Dou, Y., Liu, Q., Shi, D., Jiang, M., & Li, H. (2024). Cis-Regulatory Elements in Mammals. In *International Journal of Molecular Sciences* (Vol. 25, Issue 1). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/ijms25010343

Liu, Y., Liu, X., Zheng, Z., Ma, T., Liu, Y., Long, H., Cheng, H., Fang, M., Gong, J., Li, X., Zhao, S., & Xu, X. (2020). Genome-wide analysis of expression QTL (eQTL) and allele-specific expression (ASE) in pig muscle identifies candidate genes for meat quality traits. *Genetics Selection Evolution*, *52*(1). https://doi.org/10.1186/s12711-020-00579-x

Lo, C. (n.d.). *Algorithms for Haplotype Phasing*.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12). https://doi.org/10.1186/s13059-014-0550-8

Lv, J., Maher, K. A., Veluchamy, A., Kim, Y., Dong, L., Ju, B., Valentine, V., Valentine, M., Burden, S., Easton, J., Pounds, S. B., & Abraham, B. J. (2024). *Topology-informed regulatory element collections coordinate cell identity gene expression programs*. https://doi.org/10.1101/2024.02.01.578210

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, *17*(11), 704–714. https://doi.org/10.1038/nrg.2016.104

Macias-Velasco, J. F., St. Pierre, C. L., Wayhart, J. P., Yin, L., Spears, L., Miranda, M. A., Carson, C., Funai, K., Cheverud, J. M., Semenkovich, C. F., & Lawson, H. A. (2021).

Parent-of-origin effects propagate through networks to shape metabolic traits. https://doi.org/10.1101/2021.08.10.455860

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. In Annual Review of Genomics and Human Genetics (Vol. 7, pp. 29–59). https://doi.org/10.1146/annurev.genom.7.080505.115623

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McManus, C. J., Coolon, J. D., Duff, M. O., Eipper-Mains, J., Graveley, B. R., & Wittkopp, P. J. (2010). Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Research, 20(6), 816–825. https://doi.org/10.1101/gr.102491.109

McVicker, G., Van De Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., & Pritchard, J. K. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science*, *342*(6159), 747–749. https://doi.org/10.1126/science.1242429

Md, V., Misra, S., Li, H., & Aluru, S. (2019). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. *Proceedings - 2019 IEEE 33rd International Parallel and Distributed Processing Symposium, IPDPS 2019*, 314–324. https://doi.org/10.1109/IPDPS.2019.00041

Meiklejohn, C. D., Coolon, J. D., Hartl, D. L., & Wittkopp, P. J. (2014). The roles of cis- and trans-regulation in the evolution of regulatory incompatibilities and sexually dimorphic gene expression. *Genome Research*, *24*(1), 84–95. https://doi.org/10.1101/gr.156414.113

Menon, S., Piramanayakam, S., & Agarwal, G. (2021). COMPUTATIONAL IDENTIFICATION OF PROMOTER REGIONS IN PROKARYOTES AND EUKARYOTES. *EPRA International Journal of Agriculture and Rural Economic Research (ARER)-Peer-Reviewed Journal*. https://doi.org/10.36713/epra0813

Muráni, E., Ponsuksili, S., Srikanchai, T., Maak, S., & Wimmers, K. (2009). Expression of the porcine adrenergic receptor beta 2 gene in longissimus dorsi muscle is affected by cis-regulatory DNA variation. *Animal Genetics*, *40*(1), 80–89. https://doi.org/10.1111/j.1365-2052.2008.01811.x

Nord, A. S., Blow, M. J., Attanasio, C., Akiyama, J. A., Holt, A., Hosseini, R., Phouanenavong, S., Plajzer-Frick, I., Shoukry, M., Afzal, V., Rubenstein, J. L. R., Rubin, E. M., Pennacchio, L. A., & Visel, A. (2013). Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell*, *155*(7), 1521–1531. https://doi.org/10.1016/j.cell.2013.11.033

Pai, A. A., Pritchard, J. K., & Gilad, Y. (2015). The Genetic and Mechanistic Basis for Variation in Gene Regulation. *PLoS Genetics*, *11*(1), e1004857. https://doi.org/10.1371/journal.pgen.1004857

Pan, Y. (2006). Advances in the Discovery of cis-Regulatory Elements. In *Current Bioinformatics* (Vol. 1).

Pastinen, T. (2010). Genome-wide allele-specific analysis: Insights into regulatory variation. In Nature Reviews Genetics (Vol. 11, Issue 8, pp. 533–538). https://doi.org/10.1038/nrg2815

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000Research*, *9*(304), 1–20.

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

Richardson, P. (2010). Special issue: Next generation DNA sequencing. In *Genes* (Vol. 1, Issue 3, pp. 385–387). https://doi.org/10.3390/genes1030385

Sakumi, K. (2019). Germline mutation: De novo mutation in reproductive lineage cells. *Genes and Genetic Systems*, *94*(1), 3–12. https://doi.org/10.1266/ggs.18-00055

Saupe, S. J. (2012). A fungal gene reinforces Mendel's laws by counteracting genetic cheating. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 109, Issue 30, pp. 11900–11901). https://doi.org/10.1073/pnas.1209748109

Scally, A., & Durbin, R. (2012). Revising the human mutation rate: Implications for understand ing human evolution. *Nature Reviews Genetics*, *13*(10), 745–753. https://doi.org/10.1038/nrg3295

Seah, Y. M., Stewart, M. K., Hoogestraat, D., Ryder, M., Cookson, B. T., Salipante, S. J., & Hoffman, N. G. (2023). In Silico Evaluation of Variant Calling Methods for Bacterial Whole-Genome Sequencing Assays. *Journal of Clinical Microbiology*, *61*(8). https://doi.org/10.1128/jcm.01842-22

Shumate, A., & Salzberg, S. L. (2021). Liftoff: Accurate mapping of gene annotations. *Bioinformatics*, *37*(12), 1639–1643. https://doi.org/10.1093/bioinformatics/btaa1016

Smit, A. H. R. & G. P. (2013). RepeatMasker Open-4.0. <*http://Www.Repeatmasker.Org*>.

St. Pierre, C. L., Macias-Velasco, J. F., Wayhart, J. P., Yin, L., Semenkovich, C. F., & Lawson, H. A. (2022). Genetic, epigenetic, and environmental mechanisms govern allele-specific gene expression. *Genome Research*, *32*(6), 1042–1057. https://doi.org/10.1101/gr.276193.121

Stadhouders, R., Vidal, E., Serra, F., Di Stefano, B., Le Dily, F., Quilez, J., Gomez, A., Collombet, S., Berenguer, C., Cuartero, Y., Hecht, J., Filion, G. J., Beato, M., Marti-Renom, M. A., & Graf, T. (2018). Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature Genetics*, *50*(2), 238–249. https://doi.org/10.1038/s41588-017-0030-7

Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends in Genetics*, *18*(2), 74–82. https://doi.org/10.1016/S0168-9525(02)02592-1

Supek, F., & Lehner, B. (2019). Scales and mechanisms of somatic mutation rate variation across the human genome. In *DNA Repair* (Vol. 81). Elsevier B.V. https://doi.org/10.1016/j.dnarep.2019.102647

Tian, L., Khan, A., Ning, Z., Yuan, K., Zhang, C., Lou, H., Yuan, Y., & Xu, S. (2018). Genome-wide comparison of allele-specific gene expression between African and European populations. *Human Molecular Genetics*, *27*(6), 1067–1077. https://doi.org/10.1093/hmg/ddy027

Tonekaboni, S. A. M., Mazrooei, P., Kofia, V., Haibe-Kains, B., & Lupien, M. (2019). Identifying clusters of cis-regulatory elements underpinning TAD structures and lineage-specific regulatory networks. *Genome Research*, *29*(10), 1733–1743. https://doi.org/10.1101/gr.248658.119

Tuch, B. B., Laborde, R. R., Xu, X., Gu, J., Chung, C. B., Monighetti, C. K., Stanley, S. J., Olsen, K. D., Kasperbauer, J. L., Moore, E. J., Broomer, A. J., Tan, R., Brzoska, P. M., Muller, M. W., Siddiqui, A. S., Asmann, Y. W., Sun, Y., Kuersten, S., Barker, M. A., … Smith, D. I. (2010). Tumor transcriptome sequencing reveals allelic expression

imbalances associated with copy number alterations. *PLoS ONE*, *5*(2). https://doi.org/10.1371/journal.pone.0009317

Tycko, B. (2010). Allele-specific DNA methylation: Beyond imprinting. *Human Molecular Genetics*, *19*(R2). https://doi.org/10.1093/hmg/ddq376

Van De Geijn, B., Mcvicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: Allele-specific software for robust molecular quantitative trait locus discovery. In *Nature Methods* (Vol. 12, Issue 11, pp. 1061–1063). Nature Publishing Group. https://doi.org/10.1038/nmeth.3582

Wang, L., Zhang, Y., Zhang, B., Zhong, H., Lu, Y., & Zhang, H. (2021). Candidate gene screening for lipid deposition using combined transcriptomic and proteomic data from Nanyang black pigs. *BMC Genomics*, *22*(1). https://doi.org/10.1186/s12864-021-07764-2

Wang, M., Li, Q., & Liu, L. (2023). Factors and Methods for the Detection of Gene Expression Regulation. In *Biomolecules* (Vol. 13, Issue 2). MDPI. https://doi.org/10.3390/biom13020304

Wang, Q., Jia, Y., Wang, Y., Jiang, Z., Zhou, X., Zhang, Z., Nie, C., Li, J., Yang, N., & Qu, L. (2019). Evolution of cis- And trans-regulatory divergence in the chicken genome between two contrasting breeds analyzed using three tissue types at one-day-old. *BMC Genomics*, *20*(1). https://doi.org/10.1186/s12864-019-6342-5

Wang, X., Miller, D. C., Harman, R., Antczak, D. F., & Clark, A. G. (2013). Paternally expressed genes predominate in the placenta. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(26), 10705–10710. https://doi.org/10.1073/pnas.1308998110

Wang, Y., Gao, S., Zhao, Y., Chen, W. H., Shao, J. J., Wang, N. N., Li, M., Zhou, G. X., Wang, L., Shen, W. J., Xu, J. T., Deng, W. D., Wang, W., Chen, Y. L., & Jiang, Y. (2019). Allele-specific expression and alternative splicing in horse×donkey and cattle×yak hybrids. *Zoological Research*, *40*(4), 293–304. https://doi.org/10.24272/j.issn.2095-8137.2019.042

Weidemüller, P., Kholmatov, M., Petsalaki, E., & Zaugg, J. B. (2021). Transcription factors: Bridge between cell signaling and gene regulation. In Proteomics (Vol. 21, Issues 23–24). John Wiley and Sons Inc. https://doi.org/10.1002/pmic.202000034

Werling, D. M., Brand, H., An, J.-Y., Stone, M. R., Zhu, L., Glessner, J. T., Collins, R. L., Dong, S., Layer, R. M., Markenscoff-Papadimitriou, E., Farrell, A., Schwartz, G. B., Wang, H. Z., Currall, B. B., Zhao, X., Dea, J., Duhn, C., Erdman, C. A., Gilson, M. C.,

… Sanders, S. J. (2018). An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nature Genetics*, *50*(5), 727–736. https://doi.org/10.1038/s41588-018-0107-y

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. In *Nature Reviews Genetics* (Vol. 8, Issue 3, pp. 206–216). https://doi.org/10.1038/nrg2063

Zou, X., Gomez, Z. W., Reddy, T. E., Allen, A. S., & Majoros, W. H. (2024). *Bayesian Estimation of Allele-Specific Expression in the Presence of Phasing Uncertainty*. https://doi.org/10.1101/2024.08.09.607371

A2 Appendix: Additional files for RNA-seq and WGS Quality Control (QC) in the samples.

Table 9: RNA-Seq Quality Control (QC) obtained using fastQC and fastp. The table shows the total number of reads and the numbers of mapped and unmapped reads using STAR, in addition to the read duplication rate, categorised by the family samples and replicate

Category	total reads	Mapped to too many loci	Unmapped: too short	Unmapped: other	Duplication_rate
mother_1	38381096	1156043	1518259	21114	0.146357
mother_2	38846324	1027015	1383248	23313	0.117831
mother_3	36299170	1017956	1229151	21787	0.138375
mother_4	42087416	1087564	1521512	25253	0.145757
child_1_1	41583570	1034475	1707830	20802	0.139262
child_1_2	43972960	1275997	1536596	21983	0.167627
child_1_3	38884312	1091727	1443586	25258	0.147575
child_1_4	45694084	1204653	1622798	20571	0.164227
child_2_1	51489446	1200736	2029281	23148	0.156707
child_2_2	50327762	1401227	2036737	27659	0.149304
child_2_3	32583968	890798	1122846	14667	0.14205
child_2_4	43254396	952095	1702790	25931	0.108353
child_3_1	52241822	1214732	2173379	20898	0.158788
child_3_2	37080678	1054444	1353293	16684	0.142487
child_3_3	39969660	1129820	1694868	23984	0.171802
child_3_4	45322108	1092539	1557167	15866	0.155916
child_4_1	60733864	1347163	2710323	27346	0.159823
child_4_2	37811570	1033849	1287280	17013	0.144524
child_4_3	40467826	1071513	1387164	20251	0.154389
child_4_4	53892492	1382441	1741345	21565	0.165384
child_5_1	57456166	1585141	3555163	43110	0.18012
child_5_2	38828660	1043689	1469810	21358	0.133971
child_5_3	39128752	985706	1272057	21527	0.136031
child_5_4	41587546	1073381	1702445	20787	0.137911

child_6_1	52684230	1249994	2000966	34227	0.159954
child_6_2	28529394	818843	1009196	19984	0.135152
child_6_3	45918934	1293793	1289436	29827	0.181987
child_6_4	41146768	1058339	1818539	32915	0.138749
child_7_1	55506222	1269924	2441647	27746	0.159736
child_7_2	35551978	977630	1298523	15987	0.14852
child_7_3	42142082	1080585	1396770	16854	0.17218
child_7_4	46077270	1111660	2432776	18430	0.160997
child_8_1	55579624	1338088	2026016	27792	0.165088
child_8_2	37018486	1007595	1310603	22214	0.148081
child_8_3	44010782	1135202	1722999	22005	0.169718
child_8_4	48490872	1236423	2054804	24231	0.164248
father_1	46751208	977322	1722372	23370	0.142802
father_2	36324032	948045	1319233	19988	0.14273
father_3	51066896	1276787	2407417	30668	0.176983
father_4	38578938	973318	1768620	19287	0.145773

Table 10: WGS read Quality Control using fastQC. It shows the total number of reads as well as the duplicated rates and mapping details using bwa2

Sample	Unique Reads	Duplicate Reads	% duplicated reads	Total Reads(M)	Total Passed QC(M)	Mapped(M)	%mapped
mother_WGS_1	219155612	55899814	25,51	558,93	558,93	555,67	99,42
mother_WGS_2	221014444	54040982	24,45				
child_1_WGS_1	283283889	91433958	32,28	761,67	761,67	757,21	99,42
child_1_WGS_2	286976438	87741409	30,57				
child_2_WGS_1	244358002	75179217	30,77	649,37	649,37	645,61	99,42
child_2_WGS_2	247843107	71694112	28,93				
child_3_WGS_1	265659488	80614937	30,35	703,43	703,43	699,48	99,44
child_3_WGS_2	268530782	77743643	28,95				

child_4_WGS_1	262391915	85819605	32,71	707,80	707,80	703,34	99,37
child_4_WGS_2	265700589	82510931	31,05				
child_5_WGS_1	246960891	73452758	29,74	651,38	651,38	647,68	99,43
child_5_WGS_2	249361419	71052230	28,49				
child_6_WGS_1	389915527	154779412	39,7	1107,08	1107,08	1099,85	99,35
child_6_WGS_2	395310745	149384194	37,79				
child_7_WGS_1	222286045	46692068	21,01	546,57	546,57	543,24	99,39
child_7_WGS_2	224168488	44809625	19,99				
child_8_WGS_1	261706415	78976670	30,18	692,64	692,64	688,30	99,37
child_8_WGS_2	265149187	75533898	28,49				
father_WGS_1	227177492	100373520	44,18	665,89	665,89	653,80	98,19
father_WGS_2	228637016	98913996	43,26				

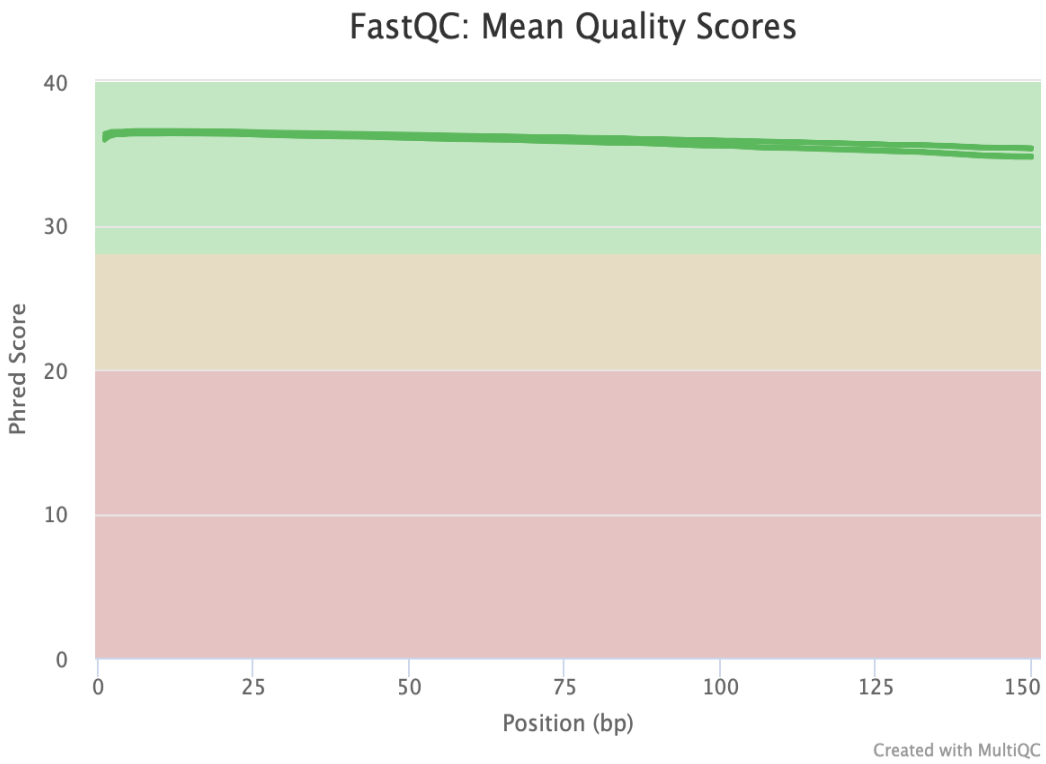


Figure 20: Mean Sequence Quality Scores for all samples using multiQC (Ewels et al., 2016) . The distribution of mean quality value at each base of the reads for all reads.

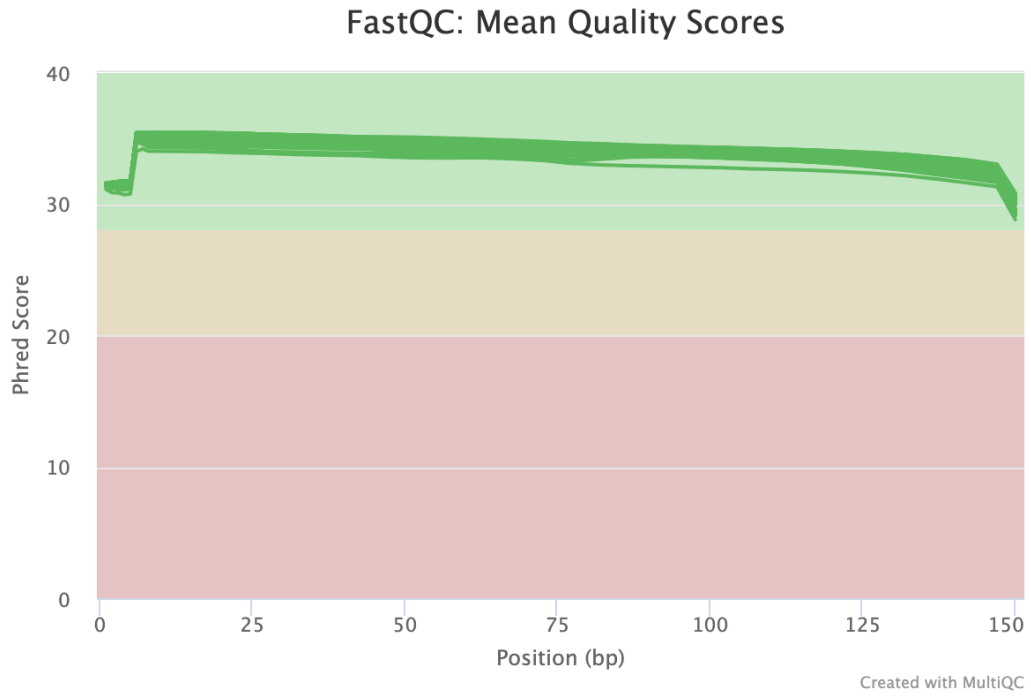


Figure 21: Mean Sequence Quality Scores for all RNA-Seq samples using multiQC. The distribution of mean quality value at each base of the reads for all reads.

A3 Appendix: Mother samples Quality Control and the elimination of mother_2 sample.

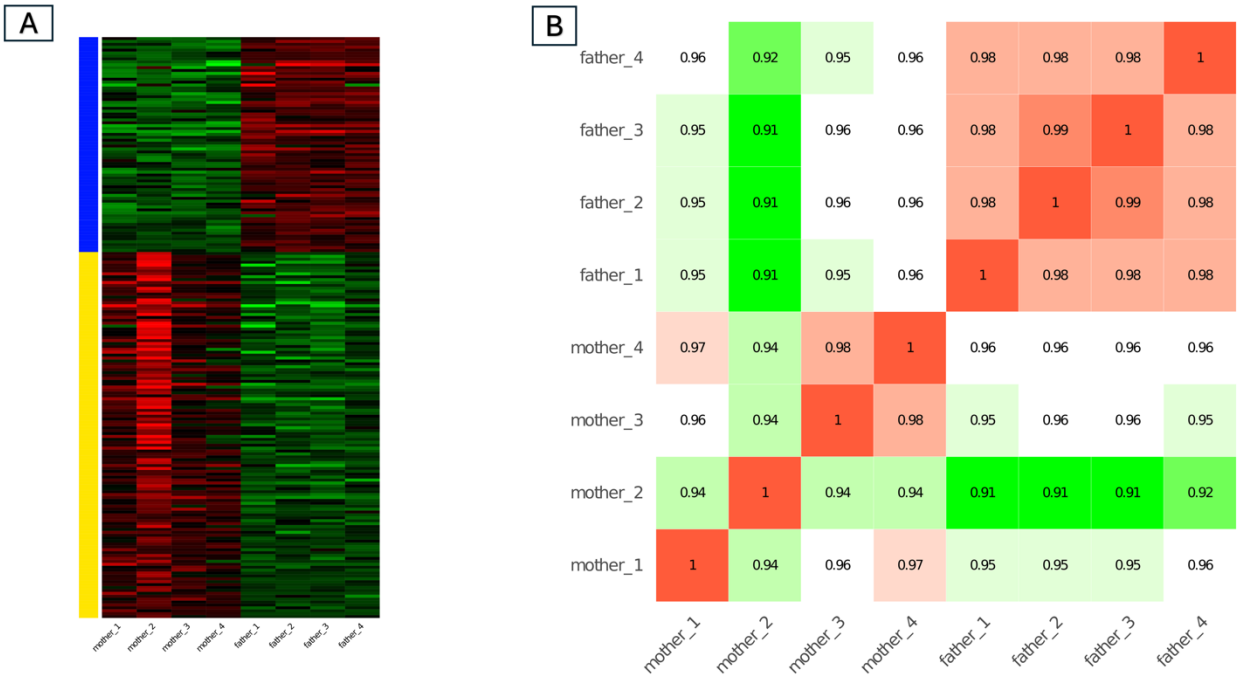


Figure 22: Quality Check for the mother replicates. A) Heatmap for the DEGs between the parents before the mother_2 removal, red color means up-regulated, and the more the color, the higher the expression, green color on the hand refers to the down-regulated genes. B) Correlation Matrix for the parents, a linear relationship among samples. The correlation matrix was achieved for each replicate in the parents using cor and corrplot functions in R, after the Variance Stabilizing Transformation (VST)

A4 Appendix: R Scripts Written for this Study Analysis: (the complete set of scripts can be found at the following repository: https://github.com/Maher199/ASE-in-a-family)

title: "ASE cases with DESeq2"
author: "Maher Alnajjar"
date: "2025-03-01"
Read Parameters in R Code

Introduction

This is a documentation for ASE cases definition in R, using mainly the DESeq2 package to create contrast of expression for the family members relative to the father in terms of log2FC

Load Required Libraries

```
suppressMessages(library(DESeq2))
suppressMessages(library(plyr))
suppressMessages(library(dplyr))
suppressMessages(library(tidyverse))
suppressMessages(library(ggplot2))
```

```
if (!file.exists(filename.count)) {
  stop(paste("Error: File not found -", filename.count))
}

txt = read.table(filename.count, header=TRUE)

txt_mother_2 <- txt %>% select(-mother_2)

data <- txt_mother_2 %>% as.data.frame()

sampletype <- factor(c(rep(".mother",3), rep("child_1", 4), rep("child_2",
4), rep("child_3", 4), rep("child_4", 4), rep("child_5", 4), rep("child_6",
4), rep("child_7", 4), rep("child_8", 4), rep("father", 4) ))

meta_full <- data.frame(sampletype, row.names = colnames(data))
```

```r
dds_full <- DESeqDataSetFromMatrix(data, colData = meta_full, design = ~
sampletype)

keep <- rowSums(counts(dds_full) >= 10) >= 3

dds_full <- dds_full[keep,]


dds_full <- DESeq(dds_full)

# Defining levels to compare ()
children_levels <- c("child_1", "child_2", "child_3", "child_4", "child_5",
"child_6", "child_7", "child_8",".mother")

# Creating an empty list for results
results_list <- list()

# Looping through each individual relatively to father
for (child_level in children_levels) {
  # Define the contrast
  contrast <- c("sampletype", child_level, "father")

  ### Run DESeq analysis
  child_results <- results(dds_full, contrast = contrast)

  ## Store
  results_list[[child_level]] <- child_results
}

convert_to_df <- function(result, child_id) {
  df <- data.frame(
    Gene = rownames(result),
    log2FoldChange = result$log2FoldChange,
    padj = result$padj,
    Child = child_id
  )
```

```
  return(df)
}

child_1_df <- convert_to_df(results_list$child_1, "child_1")
child_2_df <- convert_to_df(results_list$child_2, "child_2")
child_3_df <- convert_to_df(results_list$child_3, "child_3")
child_4_df <- convert_to_df(results_list$child_4, "child_4")
child_5_df <- convert_to_df(results_list$child_5, "child_5")
child_6_df <- convert_to_df(results_list$child_6, "child_6")
child_7_df <- convert_to_df(results_list$child_7, "child_7")
child_8_df <- convert_to_df(results_list$child_8, "child_8")
mother_df <- convert_to_df(results_list$.mother, ".mother")

### Combining all DFs as one data frame
combined_df <- bind_rows(
  child_1_df,
  child_2_df,
  child_3_df,
  child_4_df,
  child_5_df,
  child_6_df,
  child_7_df,
  child_8_df,
  mother_df
)


### Combine and rename

Log2FC_FULL <- Log2FC_FULL%>%
  rename(
    .mother_log2fc = log2FoldChange_.mother,
    .mother_padj = padj_.mother,
    child_1_log2fc = log2FoldChange_child_1,
    child_1_padj = padj_child_1,
    child_2_log2fc = log2FoldChange_child_2,
    child_2_padj = padj_child_2,
```

```
    child_3_log2fc = log2FoldChange_child_3,
    child_3_padj = padj_child_3,
    child_4_log2fc = log2FoldChange_child_4,
    child_4_padj = padj_child_4,
    child_5_log2fc = log2FoldChange_child_5,
    child_5_padj = padj_child_5,
    child_6_log2fc = log2FoldChange_child_6,
    child_6_padj = padj_child_6,
    child_7_log2fc = log2FoldChange_child_7,
    child_7_padj = padj_child_7,
    child_8_log2fc = log2FoldChange_child_8,
    child_8_padj = padj_child_8
  ) %>%
  select(Gene, .mother_log2fc, .mother_padj, starts_with("child"))

write.table(Log2FC_FULL, filename.output_full, sep='\t',quote = TRUE)
################
# Get normalised counts from Deseq2
##############

normalized_counts_full <- counts(dds_full, normalized = TRUE)

write.table(normalized_counts_full, filename.output_normalized,
sep='\t',quote = TRUE)

## Filter the rows and keep only the potential ASE (|log2fc| >1 )
selected_rows <- Log2FC_FULL %>%
  filter(if_any(ends_with("_log2fc"), ~ abs(.) >= 1))

## Output the potential ASE for downstream analysis
write.table(selected_rows, filename.output_ASE, sep='\t',quote = TRUE)

## 1- Count Plot, gene_symbol can be added

log2_plot_file_path <- file.path("./", paste(gene, "_log2_plot.png", sep =
""))
```

```
plotCounts(dds_full, gene = gene, intgroup = "sampletype", returnData =
TRUE) %>%
    ggplot() +
    aes(sampletype, count) +
    geom_boxplot(aes(fill = ifelse(grepl("^child_", sampletype), "children",
sampletype))) +
    scale_y_log10() +
    theme_bw() +
    theme(legend.position = "none", # This removes the legend
          axis.text.x = element_text(angle = 45, hjust = 1, size = 12, color
= "black")) +
    ggtitle(gene,symbol)


## 2- Log2FC bar plot

data <- selected_rows[selected_rows$Gene == "ENSOCUG000000000", 2:10]

 # Reorder the data to have 'mother' first
data <- data[, c(9, 1:8)]  # 'mother' is the 9th column

data$father <- 0.0
 # Convert to a numeric vector
data_values <- as.numeric(data)

 # Define the names for the bars (column names)
names(data_values) <- colnames(data)

 #Define the colors: expression >= -0.5, red for children < -0.5
colors <- c("red", ifelse(data_values[-1] < -0.8,"blue", "green"))

png(log2_plot_file_path, width = 1550, height = 920, res = 300) #Adjusting
the resolution & size

log_plot <- barplot(data_values,
```

```
                    ylab = "Log2FC",
                    col = colors,
                    las = 2,
                    cex.main = 0.8,  # Reduce title size
                    cex.lab = 0.5,   # Reduce axis label size
                    )

dev.off()
```

A5 Appendix: List of Publications and Presentations:

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. Family-based Analysis of Allele-Specific Expression Reveals Mostly Simple, Intermediate-type of Inheritance, 05 June 2025, PREPRINT 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)

List of 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.

Additional Presentations and Posters:

Maher Alnajjar, Levente Kontra, Zsófia Fekete, Tibor Nagy, Nóra Ninausz, Viktor Stéger, Zoltán Német, Endre Barta: Allele-Specific Expression Analysis using combined RNA-seq and WGS analysis of a rabbit family. **The Evolution of Animal Genomes, in Seville (poster w23-05)**, Spain. 18-21.09.2023.

Maher Alnajjar, Zsófia Fekete, Tibor Nagy, Nóra Ninausz, Viktor Stéger, Zoltán Német, **Endre Barta**: Family-based Approach for Detecting Allele-specific Expression, the Parents' Haplotypes, and the Recombination Events in the Children. **International Plant and Animal Genome Conference** (PAG 32) in San Diego January 10-15, 2025

Agshin Sakif1, **Alnajjar Maher**, Tibor Nagy, Endre Barta. Variation-Based Binding Affinity Analysis of The Human CTCF Transcription Factor. **Hungarian Molecular Biology Conference**, Eger, Hungary, 28-30.03.2025.