Technische Universität München
TUM School of Medicine and Health

Decoding the targets and mechanisms of the non-coding genome

Tim Phillip Hasenbein

Vollständiger Abdruck der von der TUM School of Medicine and Health der Technischen Universität München zur Erlangung eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz: **Prof. Dr. Alessandra Moretti**

Prüfende der Dissertation:

1. **TUM Junior Fellow Dr. Daniel Andergassen**
2. **Prof. Dr. Julien Gagneur**

Die Dissertation wurde am **28.02.2025** bei der Technischen Universität München eingereicht und durch die TUM School of Medicine and Health am **15.07.2025** angenommen.

Abbreviations

Ago	Argonaute	**GMC**	German Mouse Clinic
Airn	Antisense Igf2r ncRNA	**gRNA**	Guide-RNA
ASE	Allele-specific expression	**GSEA**	Gene set enrichment analysis
ATAC-seq	ATAC sequencing	**GTEx**	Genotype-Tissue Expression
BAE	Biallelic	**GWAS**	Genome-wide association study
BL6	C57BL/6J	**H3K27ac**	Acetylated histone H3 lysine 27
bp	Base pair	**H3K27me3**	Trimethylation of histone H3 lysine 27
CAST	CAST/EiJ	**H3K3me**	Trimethylated histone H3 lysine 4
cCRES	Candidate *cis*-regulatory elements	**H3K4me1**	Monomethylated histone H3 lysine 4
CTCF	CCCTC-binding factor	**HDAC3**	Histone deacetylase 3
DEGs	Differentially expressed genes	**hetSNPs**	Heterozygous single nucleotide polymorphisms
DKO	Double knockout	**HLA**	Human leukocyte antigen
DNA	Deoxyribonucleic acid	**hnRNP**	Heterogeneous nuclear ribonucleoprotein
DNMT	DNA methyltransferase	**ICEs**	Imprinting control elements
E12.5	Embryonic day 12.5	***Igf2r***	Insulin-like growth factor 2 receptor
ENCODE	Encyclopedia of DNA Elements	**IGV**	Integrative Genomics Viewer
eQTL	Expression quantitative trait loci	**kb**	Kilobase
eRNAs	Enhancer RNAs	**LINE**	Long interspersed nuclear element
FACS	Fluorescence-activated cell sorting	**lncRNA**	Long non-coding RNA
FDR	False discovery rate	**log$_2$FC**	Log$_2$ fold change
FELASA	Federation of European Laboratory Animal Science Associations	**LS**	Linkage score
Firre	Functional intergenic repeating RNA element	**MAPQ**	Mapping quality
GATK	Genomic analysis toolkit	**MAT**	Maternal
gDMRs	Gametic differentially methylated regions	**Mb**	Megabase
GEO	Gene Expression Omnibus	**miRNAs**	Micro RNAs

Misc	Miscellaneous
ml	Milliliter
Mm	Mus musculus
mRNA	Messenger RNA
MSigDB	Molecular Signatures Database
N.S.	Not significant
Nc	Non-coding
ncRNA	Non-coding RNA
ng	Nanogram
NGS	Next-generation sequencing
PAT	Paternal
pcGenes	Protein-coding genes
PCR	Polymerase chain reaction
piRNAs	Piwi-interacting RNAs
PRC	Polycomb Repressive Complex
PREs	Pumilio Response Elements
RBPs	RNA-binding proteins
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RRDs	Repeating RNA domains
rRNAs	Ribosomal RNAs
scRNA-seq	Single-cell RNA sequencing
siRNAs	Small interfering RNAs
snoRNAs	Small nucleolar RNAs
SNP	Single-nucleotide polymorphism
snRNAs	Small nuclear RNAs
SRA	Sequence Read Archive
TADs	Topologically associating domains
TEC	To be experimentally confirmed
TKO	Triple knockout
TPM	Transcripts per million
tRNAs	Transfer RNAs
UMAP	Uniform Manifold Approximation and Projection
WT	Wildtype
Xa	Active X chromosome
XCI	X chromosome inactivation
Xi	Inactive X chromosome
XIST	Xi-specific transcript
µl	Microliter
µm	Micrometer

Abstract

Protein-coding genes (pcGenes) make up less than 2% of the human genome, while the majority consists of non-coding sequences. These non-coding regions contain millions of regulatory elements that play essential roles in regulating gene expression and cellular functioning. In recent decades, non-coding RNAs (ncRNAs) have been identified as key genomic regulators, but our understanding of their functions and mechanisms remains incomplete. Thus, the ability to predict how genetic variation in non-coding regions translates into diseases is limited. This dissertation aimed to contribute to the functional understanding of the non-coding genome by exploring experimental and computational approaches.

The first project investigated the *in vivo* contribution of the three X-linked long non-coding RNAs (lncRNAs) *Crossfirre*, *Firre*, and *Dxz4*. Prior to this study, *Crossfirre* was entirely uncharacterized, including its effect on imprinted and random X chromosome inactivation (XCI). Additionally, investigating the functional role of *Firre* and *Dxz4* in random XCI has been challenging to address due to the complexity of studying mixed cell populations where either the maternal or paternal X chromosome is inactive. Using a large cohort of genetically modified mouse models, this study uncovered the functional role of these loci at the molecular and phenotypic levels. Despite the imprinting of *Crossfirre* and the unique female-specific epigenetic characteristics of *Crossfirre*, *Firre*, and *Dxz4*, these loci were found to be dispensable for XCI. In contrast, the study identified a combined effect of *Crossfirre* and *Firre* in autosomal gene regulation. Subsequent large-scale phenotyping of triple knockout mouse models revealed multiple knockout- and sex-specific phenotypes and shed light on the *in vivo* roles of *Crossfirre*, *Firre*, and *Dxz4*. The resulting dataset provides a robust basis for further studies exploring these X-linked loci.

Given that the experimental characterization of ncRNAs is laboratory extensive, the second project aimed to computationally predict the target genes and mechanisms of *cis*-acting ncRNAs. The study identified a significant enrichment of allele-specific ncRNAs nearby allele-specific pcGenes in both mice and humans, suggesting that the allele-specific expression (ASE) pattern could predict the *cis*-acting targets of ncRNAs. This concept was translated into a novel bioinformatics framework and used to predict 397 ncRNA-to-target linkages and their mechanisms across the major mouse organs. Extending this approach to human samples, the strategy was applied to 54 tissues from nearly 1,000 individuals of the Genotype-Tissue Expression (GTEx) database. The outbred nature of the human population led to the discovery of novel linkages with each sample, resulting in 2,291 human ncRNA-to-target linkages and their predicted mode-of-action. Following extensive validation using sample-matched

expression quantitative trait loci (eQTLs), the integration of genome-wide association study (GWAS) data allowed a substantial proportion of ncRNA-overlapping risk variants to be mapped to their respective protein-coding targets. With the increasing availability of sequencing data, this strategy has the potential to elucidate the targets and mechanisms of the majority of the *cis*-acting elements of the non-coding genome.

Zusammenfassung

Protein-kodierende Gene (pcGene) machen weniger als 2% des menschlichen Genoms aus, während der überwiegende Teil aus nicht-kodierenden Sequenzen besteht. Diese nicht-kodierenden Regionen enthalten Millionen von regulatorischen Elementen, die eine wesentliche Rolle bei der Regulierung der Genexpression und der zellulären Funktion spielen. In den letzten Jahrzehnten haben sich nicht-kodierende RNAs (ncRNAs) als wichtige genomische Regulatoren herausgestellt, aber unser Verständnis ihrer Funktionen und Mechanismen ist nach wie vor unvollständig. Daher können wir nur begrenzt vorhersagen, wie sich genetische Variationen in nicht-kodierenden Regionen auf Krankheiten auswirken. Ziel dieser Dissertation war es, durch die Erforschung experimenteller und computergestützter Ansätze einen Beitrag zum funktionellen Verständnis des nicht-kodierenden Genoms zu leisten.

Das erste Projekt untersuchte die *in vivo* Funktion der drei X-chromosomalen langen nicht-kodierenden RNAs (lncRNAs) *Crossfirre*, *Firre* und *Dxz4*. Vor dieser Studie war *Crossfirre*, einschließlich seiner Wirkung auf die geprägte und zufällige X-Chromosom-Inaktivierung (XCI), nicht charakterisiert. Darüber hinaus war die Untersuchung der funktionellen Rolle von *Firre* und *Dxz4* bei der zufälligen XCI aufgrund der Komplexität der Untersuchung gemischter Zellpopulationen, in denen entweder das mütterliche oder das väterliche X-Chromosom inaktiv ist, eine Herausforderung. Mit Hilfe einer großen Kohorte von genetisch veränderten Mausmodellen hat diese Studie die funktionelle Rolle dieser Loci auf molekularer und phänotypischer Ebene aufgedeckt. Trotz der Prägung von *Crossfirre* und der einzigartigen weibchenspezifischen epigenetischen Merkmale von *Crossfirre*, *Firre* und *Dxz4* erwiesen sich diese Loci als nicht relevant für XCI. Im Gegensatz dazu wurde in der Studie ein kombinierter Effekt von *Crossfirre* und *Firre* auf die autosomale Genregulation festgestellt. Die anschließende Phänotypisierung von dreifach-knockout-Mausmodellen ergab mehrere knockout- und geschlechtsspezifische Phänotypen und gab Aufschluss über die *in vivo* Rolle von *Crossfirre*, *Firre* und *Dxz4*. Der resultierende Datensatz bietet eine solide Grundlage für weitere Studien zur Erforschung dieser X-chromosomalen Loci.

Da die experimentelle Charakterisierung von ncRNAs im Labor sehr umfangreich ist, zielte das zweite Projekt darauf ab, die Zielgene und Mechanismen von *cis*-wirkenden ncRNAs computergestützt vorherzusagen. In der Studie wurde eine signifikante Anreicherung von allele-spezifischen ncRNAs in der Nähe von allele-spezifischen pcGenen sowohl bei Mäusen als auch bei Menschen festgestellt. Dies deutet darauf hin, dass diese sogenannten *cis*-aktiven Ziele von ncRNAs anhand des allele-spezifischen Expressionsmusters vorhergesagt werden können. Dieses Konzept wurde in ein neuartiges bioinformatisches Modell überführt und zur Vorhersage von 397 ncRNA-Zielgen-Interaktionen und deren Mechanismus in den zentralen Organen der Maus verwendet. Zudem wurde der Ansatz auf menschliche Proben ausgeweitet und auf 54 verschiedene Gewebe von fast 1.000 Individuen aus der GTEx-Datenbank angewendet. Die genetische Varianz innerhalb der menschlichen Population führte zur Entdeckung neuer Verbindungen mit jeder Probe, was zu 2.291 menschlichen ncRNA-Ziel-Verbindungen und deren vorhergesagter Wirkungsweise führte. Nach einer umfassenden Validierung mit Hilfe von probenangepassten eQTLs ermöglichte die Integration von GWAS-Daten die Zuordnung eines erheblichen Anteils von ncRNA-überlappenden Risikovarianten zu ihren jeweiligen protein-kodierenden Zielgenen. Mit der zunehmenden Verfügbarkeit von Sequenzierungsdaten hat diese Strategie das Potenzial, die Ziele und Mechanismen eines Großteils der *cis*-wirkenden Elemente des nicht-kodierenden Genoms zu identifizieren.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 The non-coding genome: An overview

The human genome comprises about 20,000 pcGenes which are essential for fulfilling the fundamental processes of life[1,2]. Surprisingly, these genes make up less than 2% of the genome, while the vast majority is non-coding[3,4]. These non-coding regions have long been considered as "junk" DNA, encompassing introns and intergenic sequences without functional relevance[1,5]. However, over the past decades, it has become evident that the non-coding genome harbors the critical regulatory elements that define when and where genes are turned on and off. Additionally, 90% of the disease-associated genetic variants have been identified outside of pcGenes, highlighting the functional relevance of the non-coding genome in health and disease[1,6,7]. However, our lack of understanding of the underlying regulatory mechanisms poses a significant challenge in elucidating how non-coding variants contribute to diseases.

1.1.1 DNA regulatory elements

A key group of regulatory elements within the non-coding genome are DNA elements that guide the regulation of gene expression in a tissue- and cell-type-specific manner. These elements include four general types of regulators, including promoters, enhancers, silencers, and insulators (**Figure 1.1**)[1,8].

Promoters are short sequences of DNA adjacent to the transcription start site of the genes they regulate[9]. Here, they serve as binding sites for transcription factors and the RNA polymerase, playing critical roles in initiating transcription and determining the efficiency at which a gene is expressed[1,10]. By integrating epigenetic characteristics and distal regulatory elements, promoters act as central hubs in the gene regulatory network, ensuring the appropriate temporal and spatial expression of genes[10]. One type of distal regulatory elements are enhancers, which can increase gene expression in a tissue- and cell-type-specific manner[1]. Enhancers possess sequence-specific binding sites for transcription factors, enabling them to modulate chromatin structure and transcriptional activity[11,12]. The activity of enhancers is marked by specific chromatin signatures, including acetylated histone H3 lysine 27 (H3K27ac) and monomethylated histone H3 lysine 4 (H3K4me1)[1,13]. In order to act on the transcriptional activity, enhancers must be located in open chromatin regions and form physical contact with the target promoters. This kind of interaction is further required for silencers. Analogous to enhancers, these elements recruit and bind co-factors that influence transcriptional activity. In contrast to enhancers, silencers function to repress gene expression upon promoter interaction[1]. Although silencer-specific histone modifications remain largely unexplored, trimethylation of histone H3 lysine 27 (H3K27me3) is frequently observed at repressed gene loci and may indicate silencer activity[1]. The formation of chromatin loops, mediated by structural proteins such as CCCTC-binding factors (CTCF) and cohesin, is required to establish physical contact between enhancers and silencers and the promoter of target genes. This looping architecture organizes the genome into topologically associating domains (TADs), within which regulatory elements and their target genes can interact[1,14]. At the boundaries of these TADs, insulators are frequently found to maintain the segregation. Insulators allow the physical interactions between regulatory elements to be blocked, limiting the influence of distal regulators and preventing them from affecting genes outside their defined loops[1]. Collectively, the interplay of these DNA elements allows fine-tuned temporal and spatial regulation of gene expression across different tissues and cell types.

1.1.2 Non-coding RNAs

In addition to DNA regulatory elements, high-throughput sequencing has discovered that most of the human genome is transcribed under various conditions, producing RNA transcripts that are not translated into proteins. These ncRNAs are essential for fundamental physiological processes and play important roles in regulating gene expression. To date, hundreds of thousands of ncRNAs have been identified and classified into diverse classes, each characterized by different functional properties and mechanisms[3,8,15,16]. Although we are just at the beginning of understanding their molecular functions, ncRNAs have changed the perception of RNAs as simple intermediates in protein synthesis to key regulatory elements of the genome[15].

Since the discovery of ncRNAs, their biological relevance in genome organization and protein production has become increasingly apparent. Sequencing technologies have identified a large number of different classes of ncRNAs with varying functional properties[15,17]. Based on their functional role, ncRNAs are categorized into housekeeping and regulatory ncRNAs (**Figure 1.1**). Housekeeping RNAs, which include small nuclear RNAs (snRNAs), transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and small nucleolar RNAs (snoRNAs), are ubiquitously expressed transcripts essential for protein synthesis[18]. Upon transcription of a pcGene, the messenger RNA (mRNA) is processed and spliced. This splicing process is mediated by snRNAs, which are core components of the splicing machinery[19]. The mature mRNA is then exported into the cytosol, where its sequence is translated into functional proteins. This translation process is mediated by rRNAs that form the structural and catalytic core of ribosomes[20], while tRNAs act as molecular interpreters of the genetic code, pairing the mRNA codons with their corresponding amino acids[21,22]. Additionally, the proper functioning of rRNAs and tRNAs is regulated by snoRNAs, which direct chemical modifications such as methylation and pseudouridylation. These modifications increase the stability and structural integrity of rRNAs and tRNAs, ensuring their efficiency in protein synthesis[23].

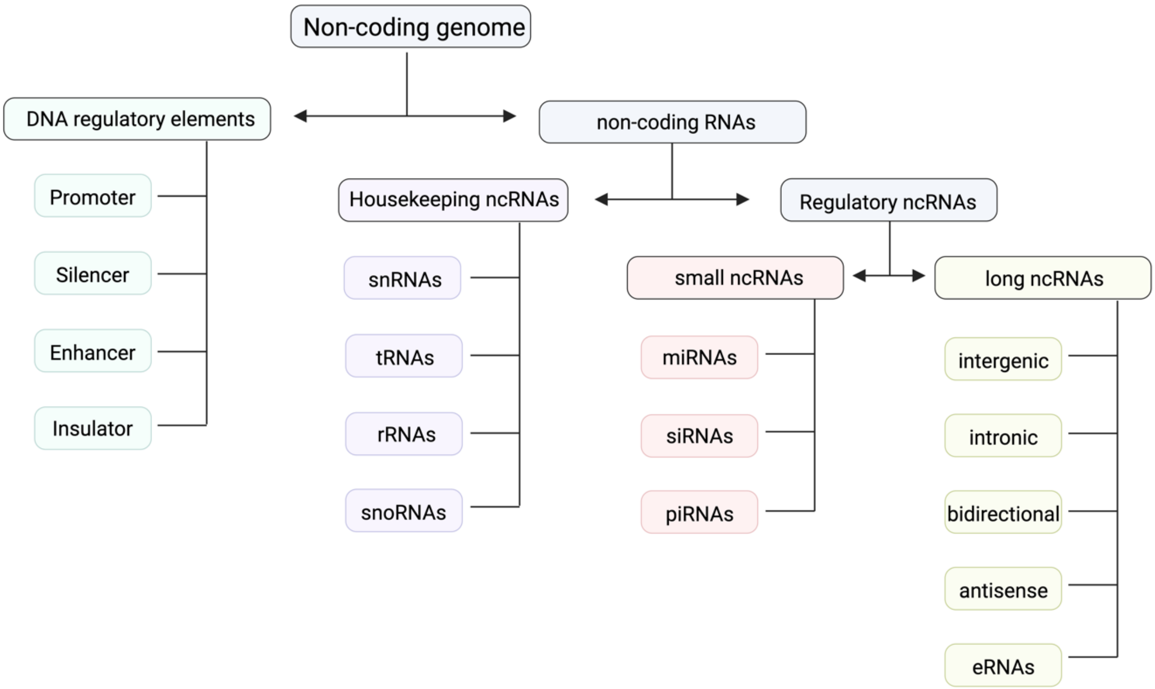


Figure 1.1 Overview of the functional elements of the non-coding genome.
Schematic overview of the functional elements of the non-coding genome, including DNA regulatory elements and the different classes of non-coding RNAs (ncRNAs). Based on their functional role, ncRNAs are categorized into housekeeping and regulatory ncRNAs. Regulatory ncRNAs are further subdivided by length into small and long ncRNAs.

1.1.2.1 Small regulatory ncRNAs

In contrast to housekeeping RNAs, regulatory RNAs control the abundance of proteins and are expressed in a dynamic and cell-type-specific manner. Based on their length, regulatory ncRNAs are subdivided into small and long ncRNAs[18]. Small ncRNAs were among the first ncRNAs identified to function as regulatory transcripts at various layers (**Figure 1.1**)[24]. The definition of small ncRNAs is based on size rather than function and typically includes ncRNAs smaller than 300 base pairs (bp) in length[25]. In 1984, the 93-nucleotide *microRNA F* was discovered in the bacterium *Escherichia coli*, marking the first reported ncRNA with regulatory function. The *microRNA F* exerts its function through base pairing with the mRNA of the *outer membrane protein F,* resulting in ribosome blocking and repression of translation[26,27]. Today, at least three different types comprise the class of small regulatory ncRNAs, including small interfering RNAs (siRNAs), microRNAs (miRNAs), and piwi-interacting RNAs (piRNAs)[28].

Both siRNAs and miRNAs are involved in post-transcriptional silencing by directing Argonaute (Ago) proteins to the nucleic acids of target genes[29]. In invertebrates, siRNAs are abundant and suppress viruses and transposable elements, whereas in mammals, their activity is largely restricted to embryonic stem cells and the germ line[28]. MiRNAs are thought to have evolved from ancestral siRNA pathways and are involved in the regulation of endogenous mRNAs. The biogenesis of miRNAs is commonly based on a canonical pathway that generates ~22 nucleotide short molecules. Upon transcription, the hairpin structure of a primary miRNA is cleaved, resulting in a precursor miRNA that is exported into the cytosol. Here, the RNase III Dicer cleaves the precursor miRNA at the terminal loop to generate a miRNA duplex. This duplex associates with the Ago protein and the passenger strand is discarded, forming the single-stranded RNA-induced silencing complex (RISC). The guide strand directs the complex to the transcripts of the regulatory targets, resulting in either translational repression or degradation[30]. In contrast to the siRNA and miRNA pathways that interact with Ago proteins, piRNAs direct PIWI proteins to methylate DNA or cleave the RNA transcript of target genes. In animals, piRNAs function primarily in the germline to silence transposable elements and maintain genomic stability[31].

1.1.2.2 LncRNAs and their regulatory mechanisms

Besides small regulatory RNAs, a significant proportion of the eukaryotic genome is transcribed into lncRNAs[32-35]. Due to their generally low expression patterns, lncRNAs were initially considered to be transcriptional noise[33]. However, since then, numerous studies have demonstrated their functional roles in developmental and cellular processes[15,17,36]. To date, thousands of lncRNA loci have been identified, with estimates ranging from 16,000 to 100,000 lncRNAs in humans[5,37,38].

LncRNAs are defined as RNA transcripts longer than 500bp that mostly lack coding potential. This definition was intended to exclude housekeeping and small regulatory ncRNAs, which are generally shorter in length[33]. According to their genomic position, lncRNAs are classified as intergenic, located between genes; intronic, located within introns; and antisense, transcribed on the opposite strand of genes. Additionally, lncRNAs that share the transcription start sites with adjacent genes are classified as bidirectional lncRNAs[39].

Compared to pcGenes, lncRNAs share several characteristics, including conserved promoters, exon structures, and splice junctions[33]. Many lncRNAs are transcribed by RNA polymerase II, are polyadenylated, and have 7-methylguanosine caps, suggesting their processing is similar to mRNAs[33,40]. However, lncRNAs also differ from pcGenes. While the promoters and exons of lncRNAs show some conservation, their primary sequences are less conserved across species compared to those of pcGenes[32,33,41]. Furthermore, lncRNAs are primarily localized in the nucleus[33], with expression patterns that are often dynamic and highly cell-type-specific, particularly during later stages of development[40,42]. A hallmark of lncRNAs is that the expression pattern can change rapidly in response to environmental factors, including stress and disease conditions[43-46].

Given their critical role in gene regulation, lncRNAs are involved in various physiological processes, ranging from DNA damage response[47], immune system regulation[48], inflammation[49-51], and metabolism[52,53] to hormone production, signal transduction[54,55], neural functions[56-58], and responses to environmental stresses in plants[33,59,60]. Numerous studies have shown that lncRNAs exert their functional roles at almost all layers of gene regulation, including the modulation of chromatin architecture, orchestrating transcriptional and post-transcriptional processes, as well as facilitating the formation of higher-order structures such as scaffolds and condensates[40]. Through these diverse mechanisms, lncRNAs can influence gene activity by either enhancing or repressing transcription or protein abundance. These effects can occur in *cis*, by affecting loci on the same chromosome, or in *trans*, by targeting genes at distant genomic sites on different chromosomes[33].

1.1.2.2.1 Chromatin regulation by lncRNAs

A regulatory mechanism by which lncRNAs act on gene expression is via the interaction with chromatin. The structure of chromatin is critical for gene activity, with open chromatin facilitating the access of the transcriptional machinery and closed chromatin repressing it[12]. LncRNAs can directly influence chromatin structure through their negative charge, which interacts with histone proteins, resulting in chromatin decondensation and rapid changes in gene expression (**Figure 1.2a**)[40,61]. Furthermore, the interaction with various proteins can modify the state of chromatin in an indirect manner. For example, lncRNAs can recruit chromatin modifiers to gene promoters[40,62] or act as decoys, sequestering these modifiers and preventing them from accessing their target sites (**Figure 1.2a**)[40,63]. Protein complexes like the Polycomb Repressive Complex (PRC) 1 and PRC2 influence gene activity by modulating transcription through histone modifications and chromatin compaction[64]. The lncRNA *ANRIL* is a key example of a regulatory lncRNA that mediates its function through lncRNA-protein interaction with PRC1 and PRC2 to affect the state of chromatin. *ANRIL* facilitates the recruitment of these complexes to the adjacent genes *CDKN2A* and *CDKN2B*, exerting *cis*-

regulatory effects on their expression[40,62]. Another aspect of lncRNA-chromatin interaction is the formation of RNA-DNA hybrids, such as R-loops or RNA-DNA-DNA triplexes[40]. Depending on the context, these hybrid structures can be recognized by transcription factors or chromatin modifiers that activate or inhibit transcription (**Figure 1.2a**)[40,65-67]. The lncRNA *MEG3* is a prime example of a hybrid-forming lncRNA. Guided by GA-rich sequences**,** *MEG3* represses genes of the TGF-β signaling pathway by the formation of triplexes that facilitate the recruitment of PRC2. This interaction leads to chromatin modifications that result in transcriptional silencing[66,68]. However, it is important to note that the prevalence of these structures is still controversial due to the challenges of detecting them *in vivo*[40].

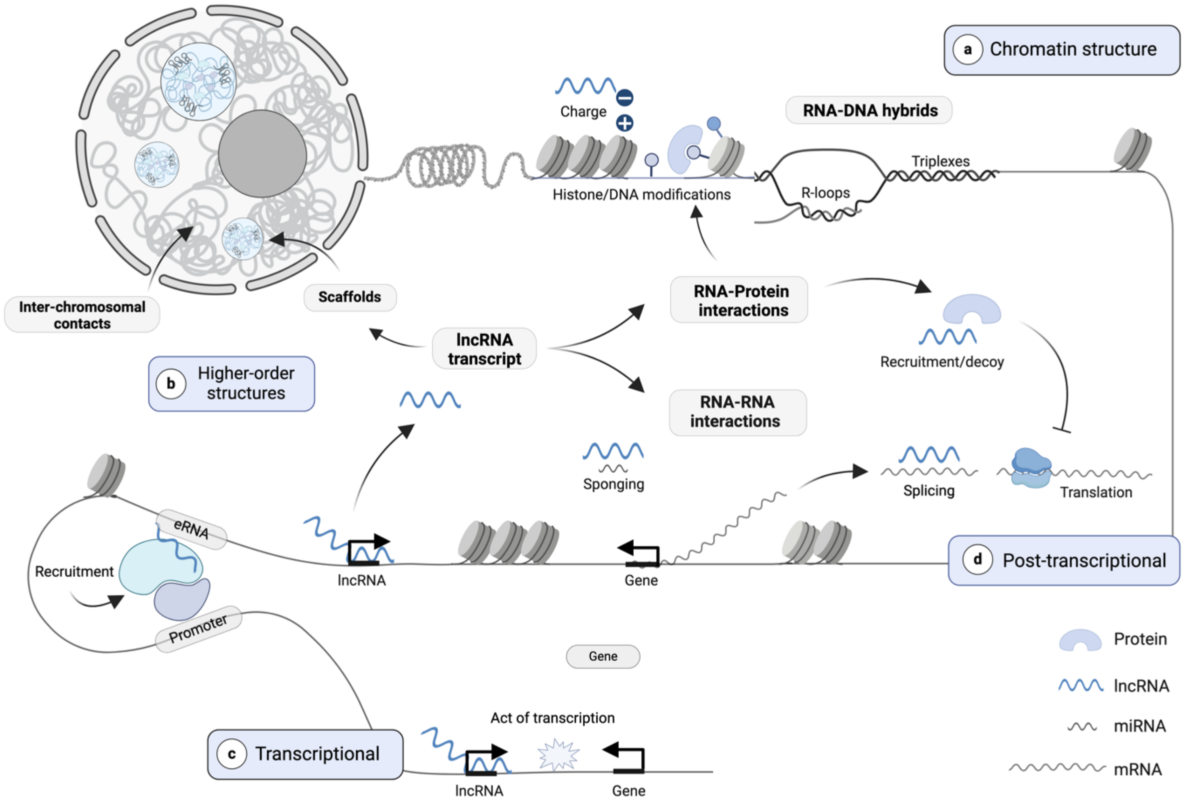


Figure 1.2 Overview of lncRNA-mediated gene regulation.
Schematic overview of the different mechanisms by which long non-coding RNAs (lncRNAs) can regulate gene expression. LncRNAs can modify (**a**) chromatin structure by sequestering or recruiting chromatin-modifying proteins or by the charge of the transcript. Moreover, lncRNAs can bind DNA to form RNA-DNA hybrids such as triplexes and R-loops. Acting as scaffolds, lncRNAs facilitate the assembly of proteins and RNAs leading to the formation of (**b**) higher-order structures, or promoting inter-chromosomal contacts. At the (**c**) transcriptional level, lncRNAs can modify gene expression via the act of transcription, e.g. by transcriptional interference or via the lncRNA transcript itself. Regulatory elements such as enhancer RNAs (eRNAs) can further be transcribed from lncRNA loci and promote gene expression by chromatin looping. Regulatory effects can also be exerted at the (**d**) post-transcriptional level by interacting with RNA-binding proteins to modify signaling pathways or the processing and stability of RNA transcripts. Moreover, lncRNA transcripts can interact with RNA molecules to recruit protein complexes that affect mRNA degradation, splicing, or act as sponges,

competing with target RNAs for miRNA binding. Created in BioRender. Andergassen, D. (2025) https://BioRender.com/b98a895.

1.1.2.2.2 LncRNAs in higher-order structures: Scaffolds and condensates

Furthermore, lncRNAs contribute to the formation of higher-order structures, including scaffolds and nuclear condensates. These structures are membraneless RNA-protein compartments that are essential for various cellular processes[40]. Acting as scaffolds, lncRNAs facilitate the assembly of proteins and RNAs, critical for regulatory activities such as pre-mRNA splicing or transcription[40,69-71]. An example of a lncRNA involved in the formation of higher-order structures is *NEAT1*. *NEAT1* plays both structural and functional roles in the formation of paraspeckles, which allow RNAs and proteins to be sequestered. Thus, lncRNA-mediated gene regulation is not only contributed to individual transcripts but often involves complex regulatory networks of multiple RNAs and proteins, collectively influencing gene expression through their coordinated activity[40]. Additionally, lncRNAs can act as modifiers of nuclear architecture, bringing different chromosomes into proximity and promoting inter-chromosomal contacts. These types of inter-chromosomal contacts allow for gene regulation across different chromosomes in a *trans*-dependent manner (**Figure 1.2b**)[72].

1.1.2.2.3 Transcriptional regulation by lncRNAs

Another mechanism by which lncRNAs regulate gene expression is through directly affecting transcriptional regulation. Thereby, gene expression can be modulated either by the act of transcription or by the lncRNA transcript itself. The mechanisms by which lncRNA activity represses or initiates gene expression are diverse and include interference with the transcriptional machinery, polymerase recruitment, histone modifications[73,74], and changes in chromatin accessibility (**Figure 1.2c**)[75,76]. Thereby, the mechanism is not restricted to a single mode-of-action but can include multiple modalities. The *antisense Igf2r ncRNA* (*Airn*) is a well-studied example of a lncRNA that regulates target gene expression in both a transcript-dependent and transcript-independent manner. In mouse extraembryonic tissues, the *Airn* transcript recruits PRC2[77,78] and G9a[79] to the promoters of multiple distant genes, leading to their silencing. Additionally, the transcription of *Airn* itself leads to the repression of the overlapping *insulin-like growth factor 2 receptor* (*Igf2r*) gene through a mechanism known as transcriptional interference in both extraembryonic and somatic tissues. Thereby, the transcriptional activity of the *Airn* locus sterically blocks the RNA polymerase II at the transcription start site, leading to the silencing of *Igf2r*[40,73,80,81]. Moreover, lncRNAs can be transcribed at active enhancers, resulting in enhancer RNAs (eRNAs)[82,83]. These nonpolyadenylated transcripts play a role in regulating chromatin looping by acting as scaffolds for protein complexes that mediate interactions between promoters and enhancers (**Figure 1.2c**)[40,84,85].

1.1.2.2.4 Post-transcriptional regulation by lncRNAs

Beyond functioning at the transcriptional level, lncRNAs exert regulatory effects at the post-transcriptional stage by interacting with RNA-binding proteins (RBPs) and nucleic acids. These interactions allow lncRNAs to modulate mRNA splicing and signaling pathways, affecting the processing, stability, and degradation of mRNA transcripts (**Figure 1.2d**)[86-88]. Such post-transcriptional regulation is exemplified by the interaction between the lncRNA *NORAD* and Pumilio proteins. Proteins of the Pumilio family bind to specific RNA motifs known as Pumilio Response Elements (PREs) on target mRNAs, promoting their degradation and repressing translation[89]. The lncRNA *NORAD* contains numerous PREs, making it a high-affinity binding partner for Pumilio proteins[90]. In response to DNA damage, *NORAD* is highly expressed and sequesters Pumilio proteins by binding them. This sequestration prevents Pumilio proteins from over-repressing their mRNA targets and maintains genomic stability[40,90].

In addition to these direct interactions with RBPs, lncRNAs can interact with RNA molecules to recruit protein complexes that affect mRNA degradation (**Figure 1.2d**). The STAU1 protein, for example, binds to double-stranded RNA structures of certain mRNAs, promoting their decay. This type of post-transcriptional regulation can be facilitated by lncRNAs that bind to STAU1 mRNA targets with complementary sequences, facilitating STAU1-mediated decay[91]. Moreover, some lncRNAs can affect gene regulation by competing with mRNAs for miRNA binding, commonly known as miRNA sponging (**Figure 1.2d**)[92,93]. These lncRNAs possess complementary sequences to those found in miRNA target sites on mRNAs, preventing them from binding to their intended mRNA targets. Thus, by acting as molecular sponges, lncRNAs can regulate miRNA availability and influence the repression of target genes[40].

In summary, lncRNAs can regulate gene expression through various mechanisms, including transcriptional and post-transcriptional processes[40]. Their interactions with chromatin, proteins, and nucleic acids, as well as their involvement in nuclear condensates, highlight the complex nature of lncRNA-mediated gene regulation.

1.2 X chromosome inactivation: A case study of lncRNA function

In the early 1990s, *H19* was discovered as the first lncRNA in humans and marked the beginning of exploring the functional mechanisms of lncRNAs[94]. Shortly after, in 1991, the lncRNA *Xi-specific transcript* (*XIST*) was identified and found to initiate the process of female XCI[36,95,96].

During mammalian development, one of the two X chromosomes becomes epigenetically silenced to achieve gene dosage compensation between males (XY) and females (XX)[97]. In mice, the upregulation of *Xist* initiates XCI in *cis* on the chromosome from which it is expressed[98]. Early knockout studies demonstrated that *Xist* is essential for the viability and proper development of female mice, as mouse models lacking *Xist* showed lethality early in embryogenesis due to two active X chromosomes in the extra-embryonic tissue[99]. Subsequent studies have shown that this early lethality can be bypassed using conditional knockout models, in which *Xist* is specifically deleted in the epiblast. In these cases, mutant embryos developed to term but failed to survive beyond weaning, attributed to defects in postnatal organ maturation[100,101].

Functional and structural differences arise between the active (Xa) and inactive (Xi) X chromosomes as a result of the inactivation process[98]. In mice, XCI occurs in two successive waves, including an initial non-random phase early in development, followed by random XCI at a later stage. In mouse embryos, the first wave occurs shortly after fertilization between the two- and four-cell stadium, resulting in imprinted silencing of the paternal X chromosome[102]. During oogenesis, the imprinted repression of the maternal *Xist* is established by H3K27me3 of a Polycomb-dependent domain spanning *Xist*. As a result, *Xist* becomes upregulated exclusively from the paternal allele, inducing the non-random inactivation of the paternal X chromosome[98,103]. In extraembryonic lineages, such as the placenta, imprinted XCI is maintained[104], whereas reactivation occurs in cells of the inner cell mass of embryonic day 3.5 blastocysts[102,105]. At embryonic day 5.5, the second wave of XCI occurs, leading to the random inactivation of one of the two X chromosomes[105]. In this phase, *Xist* becomes randomly upregulated from one of the two X chromosomes, initiating the inactivation process and gene silencing. Due to the random nature, somatic tissues exhibit a mosaic pattern of cells with either the maternal or paternal X chromosome active[98]. Although the mechanisms governing the random expression choice of *Xist* are not fully understood, the antisense transcription of *Tsix* has been identified as a repressive regulator of *Xist* during XCI[98,106]. However, further research is needed to elucidate the precise regulatory pathways of *Xist* expression. Subsequently, the Xi remains inactive and is clonally transmitted by mitosis[107]. An exception of this process occurs in primordial germ cells, where Xi reactivation occurs in embryonic day

12.5 (E12.5) embryos to ensure that both X chromosomes are active prior to oogenesis (**Figure 1.3**)[108].

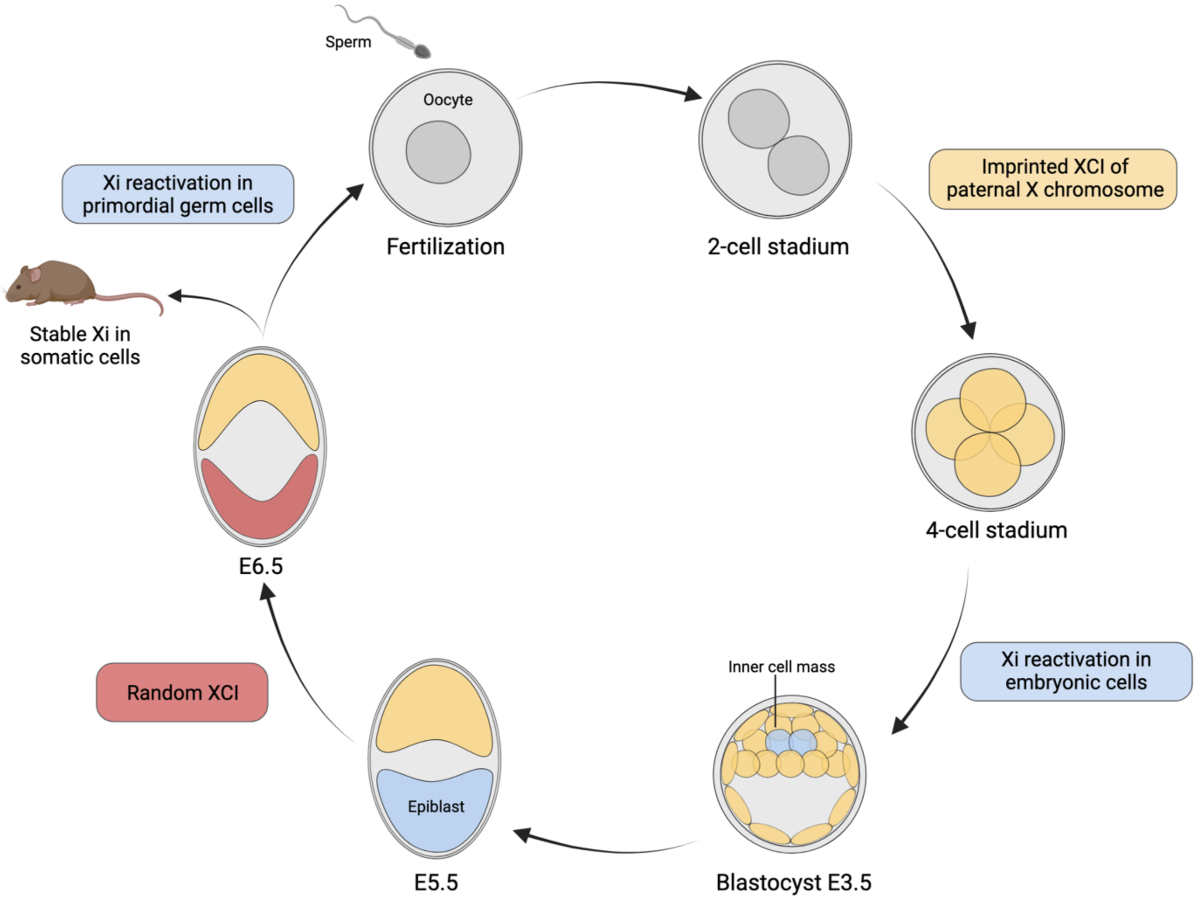

Sperm
Oocyte
Fertilization
2-cell stadium
Imprinted XCI of paternal X chromosome
4-cell stadium
Xi reactivation in embryonic cells
Inner cell mass
Blastocyst E3.5
Epiblast
E5.5
Random XCI
E6.5
Stable Xi in somatic cells
Xi reactivation in primordial germ cells

Figure 1.3 Overview of the cycle of X chromosome inactivation.
In mice, the first wave of X chromosome inactivation (XCI) occurs shortly after fertilization between the two- and four-cell stadium. This process results in imprinted silencing of the paternal X chromosome. Reactivation of the inactive X chromosome (Xi) occurs in E3.5 blastocysts in cells of the inner cell mass, while imprinted XCI is maintained in extraembryonic lineages. The second wave of XCI occurs in cells of the epiblast at embryonic stage E5.5, leading to the random inactivation of one of the two X chromosomes. Subsequently, the Xi remains stable in somatic tissues, except for primordial germ cells, where Xi reactivation occurs in E12.5 embryos to ensure that both X chromosomes are active prior to oogenesis. The figure was adapted and redrawn from Wutz *et al.* 2011[109]. Created in BioRender. Andergassen, D. (2025) https://BioRender.com/b98a895.

1.2.1 Mechanisms of action of the lncRNA *Xist*

Due to its prominent role in XCI, *Xist* became one of the best-studied examples of a *cis*-acting lncRNA with complex modular organization. The *Xist* gene locus transcribes a 17 kilobase (kb) polyadenylated and spliced transcript that contains several repetitive domains (A-F). These domains show partial sequence conservation and are essential for the proper functioning of *Xist*[95]. Once transcribed, the *Xist* RNA remains in the nucleus and coats the future Xi in *cis*. Changes in the 3D conformation of the X chromosome allow *Xist* to reach distant sites, facilitating its spreading across the entire chromosome[110].

The A-repeats of *Xist* recruit the chromatin regulator *SHARP/SPEN*, which activates histone deacetylase 3 (*HDAC3*) present on the X chromosome[111-113]. *HDAC3* then targets H3K27ac, leading to chromatin condensation and gene silencing[114]. The B- and C-repeats of *Xist* bind to the heterogeneous nuclear ribonucleoprotein (hnRNP) K to recruit PRC1 and PRC2[115]. This process strengthens the repressive compartment by inhibiting transcription initiation and sequestering the splicing machinery. After this process, *Xist* becomes dispensable for silencing the X chromosome, as its maintenance is carried forward by epigenetic modifications. These include PRC2-mediated deposition of repressive histone marks, such as H3K27me3 on regulatory regions and DNA methylation by the DNA methyltransferase (DNMT) 1 and DNMT3B[98].

Although the Xi is largely transcriptionally silent, a few genes can overcome the process of XCI and remain active. These so-called escape genes are consequently expressed from both alleles, increasing the gene dosage of females. In mice, approximately 3-7% of the X-linked genes have been reported to escape XCI, while in humans more than 20% are estimated to escape[98,116-118]. This variability in gene expression contributes to an increased phenotypic diversity in females. Gene escape occurs in a constitutive and facultative manner. Constitutive escape genes consistently overcome XCI in most cell lineages and tissues, whereas facultative escapees are tissue or lineage specific and can vary across developmental stages, increasing the cellular diversity of females[78,98,118,119]. While constitutive escape genes often have homologs on the Y chromosome and are required to maintain gene expression dosage, the functional roles and the molecular mechanisms of many facultative escape genes remain to be investigated.

1.2.2 The structure and organization of Xi: Roles of *Dxz4* and *Firre*

As a result of the XCI process, the Xi forms a compact chromatin structure called the Barr body[120]. Chromosome conformation capture methods have shown that the Xi is depleted of TADs compared to the Xa. In contrast, the Xi harbors two conserved megadomains of high intrachromosomal contact, that bisect the inactive chromosome[121-123]. The conserved macrosatellite *Dxz4* is located at the boundaries of these structures and is transcribed into the lncRNA *4933407K13Rik*. On the Xi, *Dxz4* is hypomethylated and contains several Xi-specific CTCF binding sites[124]. Moreover, *Dxz4* contributes to the folding of the Xi by forming a conserved superloop interaction with the *functional intergenic repeating RNA element* (*Firre*) locus (**Figure 1.4**)[122,125].

The *Firre* locus transcribes a well-studied lncRNA and exhibits multiple characteristics specific to the Xi. Similar to the *Dxz4* locus, the *Firre* gene body contains multiple CTCF binding sites that are specific to the Xi and are marked by trimethylated histone H3 lysine 4 (H3K4me3)[78,124]. These binding sites have been shown to anchor the Xi to the nucleolus, supporting its perinucleolar localization[124,126]. The *Firre* locus produces full-length transcripts from the Xa, whereas shorter isoforms have been reported to escape XCI on the Xi[72,78,124]. Furthermore, the *Firre* RNA has been shown to mediate crosstalk between Xa and Xi in somatic cells, with Xa-derived transcripts maintaining H3K27me3 enrichment on Xi (**Figure 1.4**)[124].

Due to these Xi-specific characteristics, *Dxz4* and *Firre* have been hypothesized to play a role in XCI, but multiple studies have indicated that these loci are dispensable for the inactivation process[121,125-129]. It has been shown that the deletion of *Firre* and *Dxz4* in cell lines results in the loss of the superloop interaction and the megastructures present on Xi[128,130]. Interestingly, these changes in the 3D conformational structure did not affect the expression of genes located on the Xi[125]. Another study revealed that the establishment of XCI remains further unaffected, as deleting these loci *in vivo* revealed fertile and viable mutants[127]. However, the mosaic nature of random XCI has made it challenging to assess the impact of these loci on XCI maintenance *in vivo*, as single-cell analyses are required to conclusively determine their precise role.

Although no overt effects have been detected for XCI, the deletion of *Firre* and *Dxz4* has been shown to affect autosomal gene regulation in an organ-specific manner *in vivo*[127,131]. While the functional properties of *Dxz4* remain largely elusive, the lncRNA *Firre* has been extensively studied. Cell culture models have demonstrated a variety of functions of *Firre*, including gene regulation[132], adipogenesis[133], and nuclear architecture[72,128]. *Firre* is abundantly transcribed and contains multiple repeating RNA domains (RRDs) and local repeats[72,134]. It has been shown that the nuclear retention of *Firre* RNA is maintained by these RRDs, which serve as a nuclear retention signal[134]. Further, these repeats allow *Firre* to form *trans*-chromosomal interactions that affect autosomal gene regulation[134]. Upon transcription, *Firre* RNA localizes around its gene body and binds to the nuclear matrix factor hnRNP U, facilitating the tethering of distant chromosomes to co-regulate genes in *trans*[72,134]. Additionally, a recent study investigated the temporal dynamics of *Firre* by monitoring gene expression changes across different time points following *Firre* induction[135]. The authors observed that the RNA of *Firre* acts on the epigenetic and transcriptional landscape within minutes, leading to abundant transcriptional changes on a longer timescale[135].

Loss- and gain-of-function studies in mice have further shown the physiological implications of *Firre*. During hematopoiesis, *Firre* mutants exhibit alterations in blood cell composition[131]. Notably, these effects could be rescued by the transgenic expression of *Firre*, highlighting a *trans*-acting molecular mechanism (**Figure 1.4**)[131]. Additionally, the study showed that overexpression of *Firre* was associated with impaired survival upon exposure to lipopolysaccharides[131]. In humans, duplications of the *FIRRE* locus have been associated with intellectual disability and dysmorphic features[136,137]. Moreover, *FIRRE* has been linked to the survival outcomes of patients with diffuse large B-cell lymphoma and has been shown to promote tumor growth in multiple cancers[138-140]. Although the *Firre* locus is extensively characterized, large-scale phenotyping studies of *Firre* using mouse models are currently lacking.

1.2.3 The imprinted lncRNA *Crossfirre* within the *Firre* locus

An additional lncRNA has been annotated within the *Firre* locus, transcribed in an antisense orientation to *Firre*. This lncRNA, termed *Crossfirre* (*Gm35612*), consists of 3 exons and is embedded in a 50kb long interspersed nuclear element (LINE). LINEs are a group of transposable elements that are hypothesized to facilitate the spreading of XCI across the chromosome[141]. Moreover, an extensive allele-specific analysis identified *Crossfirre* as an imprinted X-linked gene in somatic tissues that is predominantly expressed from the maternal X chromosome[78]. Imprinted expression of *Crossfirre* was detected in adult brains through RNA-sequencing (RNA-seq) analysis and further confirmed by observing maternal enrichment of H3K4me3 at the promoter of *Crossfirre* in mouse embryonic fibroblasts[78]. Considering the imprinted characteristics, the *Crossfirre* locus may be worthwhile investigating for its potential association with imprinted XCI. Prior to this thesis, the *Crossfirre* locus was entirely uncharacterized and the *in vivo* role at the molecular and phenotypic level was unknown. In addition, the relation of *Crossfirre* to both imprinted and random XCI, either independently or in conjunction with *Firre* and *Dxz4*, remained to be explored (**Figure 1.4**).

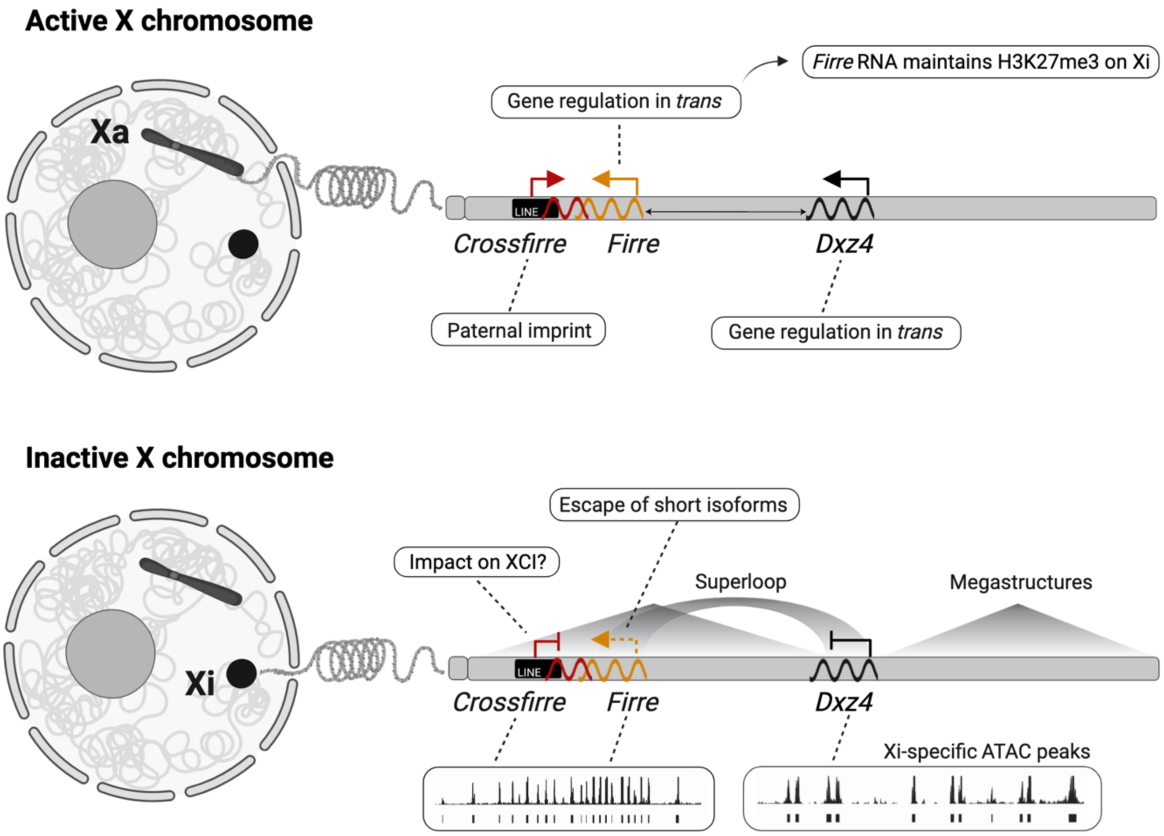


Figure 1.4 Schematic overview of the *Crossfirre*, *Firre,* and *Dzx4* loci.
The active X chromosome (Xa, upper panel) and inactive X chromosome (Xi, lower panel) are shown with the *Crossfirre*, *Firre,* and *Dxz4* loci highlighted. Transcription of these loci is specific to the Xa, with shorter isoforms of *Firre* that may transcribe from the Xi. The superloop interaction between *Firre* and *Dxz4,* and the two megadomains are specific to the Xi. The colors indicate *Crossfirre* (red), *Firre* (orange), and *Dxz4* (black). Created in BioRender. Andergassen, D. (2025) https://BioRender.com/b98a895.

1.3 Allele-specific expression

Mammals are diploid individuals and thus, except for the sex chromosomes in males, the genome consists of two copies of each chromosome, one inherited from the maternal and one from the paternal side. For most genes, both alleles contribute equally to the expression, referred to as biallelic expression. However, a subset of genes, including the lncRNA *Xist* on the female X chromosome, show predominant expression from one allele, a phenomenon known as ASE (**Figure 1.5**)[142].

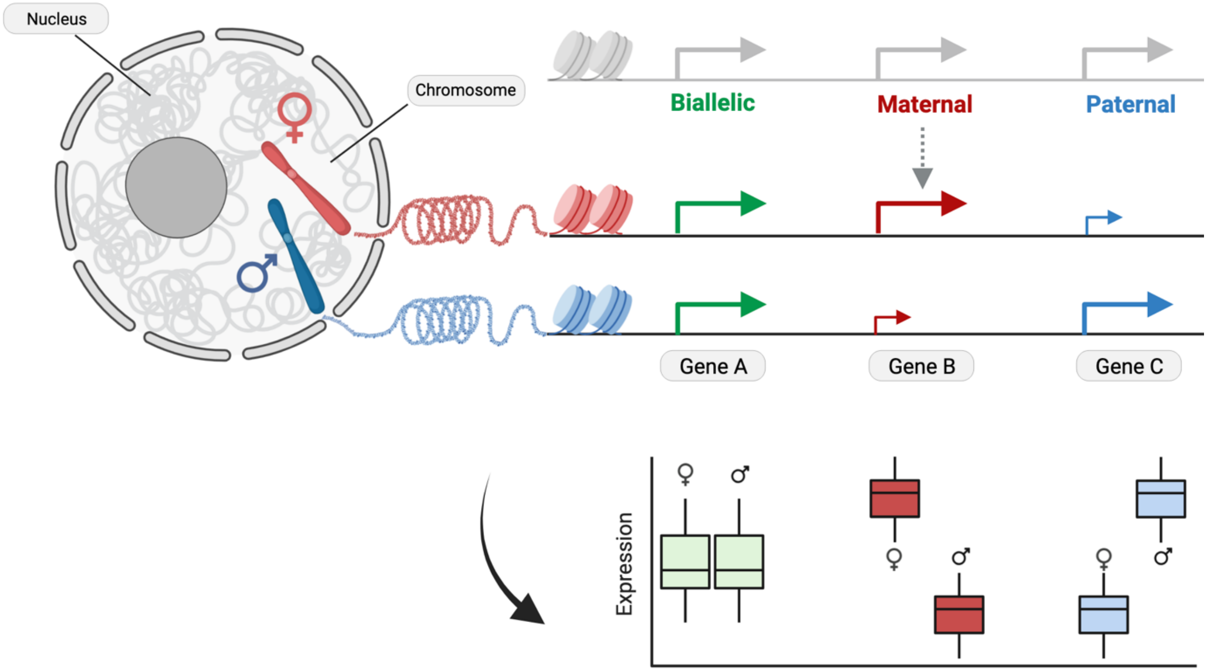


Figure 1.5 The concept of allele-specific expression.
Diploid individuals, such as mammals, possess two alleles of each chromosome, including a maternal and a paternal allele. Usually, gene activity is considered as the sum of expression derived from both chromosomes. In allele-specific expression (ASE) analysis, each allele is considered individually. The majority of genes are expressed biallelically, with both alleles being expressed at equal levels. However, a subset of genes shows ASE where gene expression levels differ between the maternal and paternal alleles. Created in BioRender. Andergassen, D. (2025) https://BioRender.com/b98a895.

ASE occurs throughout the entire genome at both the tissue[78] and single-cell level[143,144]. Several mechanisms can lead to the expression of genes in an allele-specific manner, including random monoallelic expression, such as observed in the case of random XCI, as well as allele-specific differences arising from genetic variation or epigenetic modifications like genomic imprinting[142].

1.3.1 Allele-specific expression arising from genomic imprinting

A well-studied mechanism of ASE is genomic imprinting, which was discovered in the 1980s through pronuclear transplantation in mice[145,146]. Genomic imprinting is a consequence of inheritance and characterized by epigenetic modifications between the alleles, leading to parental-specific gene expression from either the maternal or paternal chromosome[147].

In 1991, *Igf2r* was the first gene in mice to be identified as imprinted, followed by the discovery of *Igf2* and *H19*[148-150]. To date, approximately 100 mouse and 40 human imprinted genes have been discovered, advancing our understanding of the underlying molecular mechanisms[78,151-153]. In mice, the genetic imprint of genes is established in the germline during oocyte and sperm development. This process involves the DNA methylation complexes DNMT1A and DNMT3L to establish de novo methylation marks at gametic differentially methylated regions (gDMRs)[154-157]. After fertilization, genome-wide DNA demethylation occurs in both parental genomes during preimplantation development. However, imprints at gDMRs are protected from this demethylation and are stably maintained in somatic cells throughout mitosis by DNMT1[156,158,159]. An exception occurs in the primordial germ cells, where erasure of the imprints occurs at E12.5 to reset them for gametogenesis[160].

Approximately 80% of the imprinted genes have been identified to be organized in clusters, where a single gDMR regulates the imprinted expression of multiple genes. These gDMRs are defined as imprinting control elements (ICEs) and act in *cis* to repress expression by targeting small clusters of genes[161-163]. Genetic deletion studies of ICEs have identified the responsible gDMRs for multiple imprinted gene clusters, including *Kcnq1*, *Pws/As*, *Gnas*, *Igf2-H19*, *Grb10*, *Dlk1-Meg3* or *Igf2r-Airn*[164-170]. It was demonstrated that the imprinted lncRNAs in these clusters are expressed from the opposite allele as their associated pcGenes. Additionally, it was shown that deleting the ICE on the allele expressing the imprinted lncRNA restored the biallelic expression of the pcGenes. These findings highlight the functional role of lncRNAs in imprinted loci by repressing the pcGenes, as confirmed for *Kcnq1ot1* and *Airn* within the *Kcnq1* and *Igf2r* clusters, respectively[80,171]. For these cases, the ICE is located at the promoters of the lncRNAs, regulating their parental-specific expression and resulting in allele-specific repression of the target genes.

One of the best-studied examples of a regulatory lncRNA that represses several pcGenes in a parental-specific manner is the *Igf2r/Airn* cluster. *Airn* is an imprinted, paternally expressed lncRNA that silences target genes in a *cis*-dependent manner across most tissues[172,173]. Allele-specific analyses of this cluster have been a powerful tool for identifying the regulatory targets of the lncRNA *Airn*. While *Igf2r* is repressed by *Airn* in almost all tissues through the act of transcription, the *Airn* RNA represses six distant genes within a 10 megabase (Mb) window in the placenta (*Pde10a*, *Park2*, *Slc22a3*, *D*act2, *Smoc2*, *Thbs2*), by the recruitment of epigenetic repressors[78,79,148,174,175]. The example of *Airn* illustrates that imprinted clusters controlled by an allele-specific lncRNA provide valuable models for epigenetic discovery, allowing researchers to disentangle the effects of lncRNA expression on one allele compared to the allele lacking lncRNA expression, all within the same nuclear environment[161].

1.3.2 Allele-specific expression arising from genetic variation

Although genomic imprinting is a well-studied phenomenon leading to ASE, the vast majority of ASE is driven by genetic variations between the alleles[176]. Genetic deletions, insertions, and single-nucleotide polymorphisms (SNPs) can affect gene expression in *cis* or *trans*, with heterozygous *cis*-acting SNPs frequently leading to ASE. Thereby, genetic variation can mediate ASE through a variety of mechanisms, including the transcriptional and post-transcriptional level (**Figure 1.6**).

At the transcriptional level, heterozygous SNPs can affect chromatin accessibility, giving rise to ASE in a tissue-specific manner. In addition to imprinting, genetic variants can alter epigenetic marks such as DNA methylation or histone modifications through sequence-dependent allele-specific methylation[177,178]. Moreover, heterozygous polymorphisms can affect the binding affinity of transcription factors to promoters or enhancers, leading to differential gene expression between the alleles (**Figure 1.6**)[179,180].

At the post-transcriptional level, genetic variation can affect the abundance of transcripts and isoforms. Heterozygous SNPs can trigger nonsense-mediated decay, a cellular quality control mechanism that leads to the degradation of mRNAs with premature stop codons[142,181-183] (**Figure 1.6**). Another post-transcriptional mechanism by which genetic variants can affect the transcript abundance of genes is by disrupting RNA binding sites for proteins that are crucial for RNA processing, localization, and translation[184]. Alterations in binding sites, for example those for miRNAs, can further alter transcript degradation rates and contribute to ASE[142,185]. Moreover, variants located near splice sites can lead to alternative splicing, which often involves alternate 3' or 5' exon ends, exon skipping, or intron retention, affecting the number of transcript isoforms within cells (**Figure 1.6**)[142,186]. Importantly, lncRNAs show ASE more frequently than pcGenes and thus may contribute significantly to the presence of ASE in the genes they regulate in *cis*[78,142,187]. Finally, heterozygous variants can further impact the translation of mRNAs by altering the regulatory regions involved in this process, such as secondary mRNA structures, the 5' untranslated region, or the translation start site (**Figure**

1.6)[142,188]. In summary, the influence of genetic variation on the presence of ASE encompasses a wide range of possibilities.

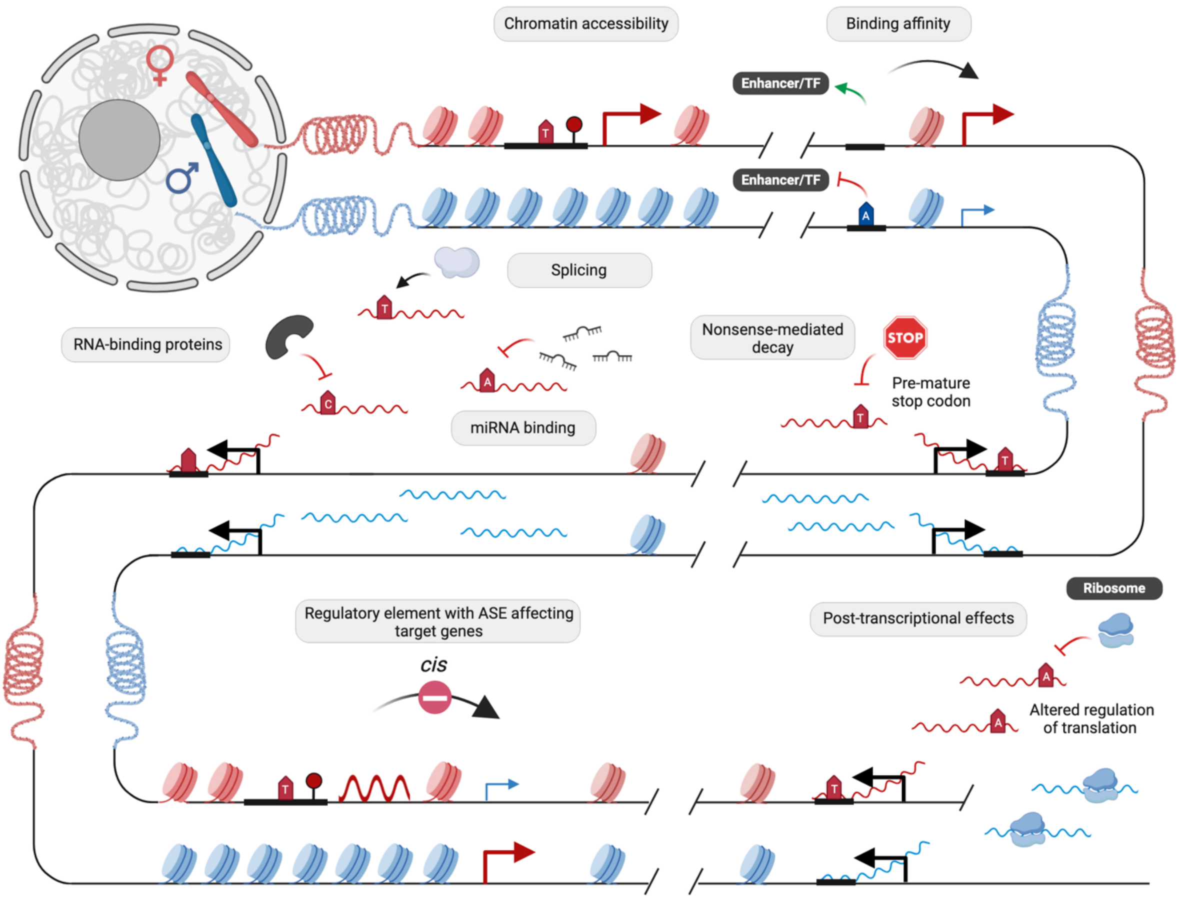


Figure 1.6 Overview of the genetic mechanisms leading to allele-specific expression. Heterozygous single nucleotide polymorphisms (hetSNPs) can affect gene expression in *cis*, leading to allele-specific expression (ASE) through various mechanisms. Chromatin accessibility can be affected by hetSNPs that alter epigenetic marks, such as histone modifications or DNA methylation. Moreover, hetSNPs can affect the binding affinity of transcription factors to promoters or enhancers. At the post-transcriptional level, genetic variation can influence the abundance of transcripts by triggering nonsense-mediated decay or disrupting RNA binding sites. Variants located near splice sites can further lead to alternative splicing. Changes in miRNA binding sites can alter transcript degradation rates and thus contribute to ASE. Additionally, allele-specific ncRNAs can cause ASE in the genes they regulate in *cis*. At the translational level, hetSNPs can alter secondary mRNA structures, the 5' untranslated region or the translation start site. Created in BioRender. Andergassen, D. (2025) https://BioRender.com/b98a895.

1.3.3 Genome-wide allele-specific expression analysis

In order to identify loci with allelic imbalance, ASE analysis involves the quantification of sequencing reads that derive from the paternal and maternal allele. This approach requires that the sequencing reads overlap heterozygous SNPs to distinguish between the alleles.

A robust method for studying ASE is the generation of F1 mouse hybrids by crossing different inbred strains. Given that the SNP information for these strains is known, the individual sequencing reads can be traced back to the allele of origin. Moreover, the number of variants can be maximized by crossing genetically distant strains[189]. In contrast, human studies require prior genotyping and phasing to identify the corresponding alleles. While genetic variants can be called directly from RNA-seq data, this approach fails to detect gene features with monoallelic expression. In such cases, these sites are misclassified as homozygous because only one allele is expressed[142]. Despite the need for prior SNP calling, the high genetic diversity of humans has led to a widespread prevalence of ASE gene loci across the population. Analyses of the GTEx project, which includes RNA-seq data from nearly 1,000 individuals across 54 different tissues[190], have demonstrated that a substantial proportion of genes show ASE in at least one sample[142,191]. However, it is essential to note that the occurrence of ASE in individual samples is not necessarily biological meaningful but rather a result of the genetic variation in outbred populations[142,191].

Due to the presence of heterozygous variants in ASE mapping, the alignment of sequencing data can be biased towards the allele that is more similar to the reference genome. This mapping bias can result in the identification of false positives and must be considered[192]. Several strategies can minimize this effect, including the use of SNP-tolerant mappers[193-196] or the alignment to masked references[197], personalized genomes, or haplotypes[142,198,199]. Subsequently, different computational approaches can be used to identify ASE from bulk or scRNA-sequencing (scRNA-seq) data to resolve allelic imbalances on tissue or cell-type level[142].

1.3.3.1 Statistical methods to identify allele-specific expression

Different statistical models are used to assess ASE, which can be classified into two categories: those designed for ASE mapping in individual samples and those designed to identify putative regulatory variants across populations[142]. Further, sequencing reads can be mapped to individual heterozygous SNPs or haplotypes encompassing multiple variants, for example within a gene locus[142]. Haplotypes are sets of polymorphisms that are likely to be inherited together and provide a more comprehensive picture than individual variants. Including the haplotype information has been shown to increase the power of ASE mapping[200].

A straightforward approach to statistically assess for ASE within individuals is binomial testing. A binomial test compares the number of sequencing reads corresponding to the maternal and paternal allele against the null hypothesis that both alleles are expressed equally, meaning with the same probability[142]. A bioinformatics tool that uses the binomial test to identify ASE from high-throughput sequencing data is Allelome.PRO. Allelome.PRO was designed to provide the entire picture of ASE loci in F1 mouse hybrids, including biallelic, imprinted and strain-biased genes. Furthermore, Allelome.PRO calculates an allelic ratio which is the proportion of sequencing reads from one allele relative to the total reads at a specific locus[201]. The ease of use and interpretation of the results have contributed to the success of Allelome.PRO as an established tool for ASE mapping.

Another computational approach to test for allelic expression are Bayesian models. Bayesian models provide a probabilistic framework to generate robust estimates of ASE[202-204]. By integrating prior knowledge and updating probabilities, Bayesian approaches can offer deep insights into complex patterns of ASE. The parameters of ASE variation across an individual's genes can be learned using these models[203,204]. Moreover, intra-individual ASE data can be combined with total gene expression variation across individuals to detect regulatory variants[142]. By extending the haplotype information derived from population phasing to include non-coding regions, these models further allow the identification of putative regulatory variants and provide insights into the mechanisms driving ASE[142,200].

1.4 Unraveling the function of ncRNAs

Despite the rapid advances in the field of ncRNAs, our understanding remains fragmented and incomplete, with functional insights often lacking[36]. To date, less than 1% of the identified loci have been experimentally characterized[17,205].

For lncRNAs, genetically modified mouse models are considered gold-standard experiments to unravel their functional roles, a process that is time-consuming and laboratory-extensive[101]. The diverse mechanisms inherent to the regulatory nature of lncRNAs require comprehensive experimental strategies to unravel the precise functions and mechanisms. Especially for lncRNAs it often remains challenging to distinguish whether regulatory effects arise from the transcript, the act of transcription, or the underlying DNA sequence and thus, a variety of experimental approaches are needed to disentangle their mechanisms. A critical starting point for characterizing novel lncRNAs is the whole-gene ablation *in vivo* to identify potential functional consequences. This approach can be complemented by more refined strategies, such as polyadenylation-terminator insertion or promoter deletion, to induce transcriptional termination. These techniques allow researchers to differentiate between the effects of DNA elements inherent to the lncRNA locus and those driven by the transcriptional process or the transcript[101]. To further disentangle effects due to the transcription, promoter activity, or the lncRNA transcript, the gene body of lncRNAs can be replaced with a reporter gene. In addition, transgene rescue experiments can be used to distinguish whether a lncRNA acts in *cis* or *trans*. In some cases, lncRNAs may further encode small functional peptides that can be identified by introducing frameshift or start codon mutations[101].

Following a knockout experiment, characterization of the functional roles requires comprehensive molecular phenotyping using multi-omics approaches such as epigenetic profiling and transcriptomic analysis. In particular, knockout and knockdown experiments enable the identification of dysregulated genes, allowing researchers to pinpoint the molecular functions and regulatory targets of the ncRNA in question. Comparative analysis of dysregulated genes can reveal the pathways affected by the ncRNA, providing insights into its role in biological processes[206]. By understanding the pathways and molecular interactions involved, researchers can infer the broader physiological or developmental implications of a ncRNA and predict potential effects on cellular function and disease mechanisms.

In addition, the functional consequences of dysfunctional regulatory RNAs can be revealed by phenotypic analysis. For ncRNAs these effects are often subtle and context-dependent[33]. NcRNAs frequently exert pleiotropic effects, meaning that their influence can vary significantly across developmental stages, tissues, or in response to environmental cues, leading to distinct phenotypic outcomes under various conditions. Detecting these nuanced effects requires comprehensive sampling across diverse tissues, cell types, and developmental stages, as well as large-scale phenotyping efforts encompassing a wide range of tests[207]. This approach enables researchers to capture subtle variations in gene expression, cellular function, and organismal health that may arise due to the ncRNA, providing a clearer understanding of their contributions to complex phenotypes and potential disease associations[207].

An illustrative example of a lncRNA with no essential phenotype upon deletion in mice is the highly abundant lncRNA *Malat1*. Several loss-of-function studies in cell culture models highlight the importance of *Malat1* for nuclear speckle formation[33,208,209]. However, multiple researchers did not detect any overt phenotype after genetic removal of the gene[208-210]. Nevertheless, *Malat1* has significant implications for the progression of multiple cancers and diseases[33,211,212]. The context-specific functionality of *Malat1* underscores the need for extensive molecular and large-scale phenotyping to identify the functional roles of ncRNAs.

To facilitate the selection of candidate loci for experimental investigation, there is a growing need for computational methods that predict the functional role and regulatory targets of ncRNAs. Bioinformatic tools that allow their prioritization based on predicted interactions and associations to tissues or conditions can reduce the experimental effort and increase the likelihood of investigating functionally relevant ncRNAs. However, the computational identification of ncRNA-targets and their mechanisms is challenging. Due to the low sequence conservation among species, functional predictions based on paralogs or orthologs with similar sequences is complex[17]. Unlike pcGenes, whose sequences are rich in functional information, ncRNAs mostly lack sequence-function relationships[17]. Additionally, the low expression levels often lead to the underrepresentation of ncRNA transcripts in sequencing data. Bulk RNA-seq approaches favor highly abundant RNAs, resulting in the undersampling of many ncRNAs[39]. The dynamic and cell-type-specific nature of ncRNA expression further complicates their detection, as many ncRNAs are expressed only in rare subpopulations of cells, at specific time points during development, or in response to environmental factors[33]. So far, computational methods, such as genotype-expression correlation studies have been used to predict the targets of regulatory loci. These studies test for genotypes associated with the expression level of genes across samples. The resulting statistically significant associations are defined as eQTLs, which represent genetic variations linked to the expression of a gene[213]. However, genotype-expression correlation studies require large sample sizes to obtain sufficient statistical power. Due to the dynamic expression patterns and temporal variations of ncRNAs, these methods have consequently failed to identify a large number of regulatory targets and mechanisms of ncRNAs.

To date, GWAS that rely on collections of DNA samples from individuals with different phenotypes, such as healthy and diseased, have uncovered hundreds of thousands of disease-associated variants by statistically testing genetic variations for their association with phenotypes[214]. Approximately 90% of the identified GWAS variants are located within the non-coding genome[215,216]. Interestingly, given their critical role in regulating gene expression, ncRNA dysregulation has been associated to a vast range of human traits and diseases, such as cardiovascular and infectious diseases, cancer, and neurological disorders[15]. Notably, a total of 371,647 risk variants have been mapped to lncRNA loci, accounting for 45% of all identified human GWAS variants[217]. However, the lack of a functional understanding of the vast majority of ncRNAs, including their regulatory targets and mechanisms of action, poses a significant challenge in elucidating how variants within the non-coding genome contribute to disease. Thus, one of the major challenges today is to unravel the targets and mechanisms of action of ncRNAs[15,17].

2 Scientific aim

Despite the advances in ncRNA research, our understanding of their functional roles and regulatory targets remains fragmented and incomplete. To date, less than 1% of the identified ncRNA loci have been experimentally characterized[17,205]. Thus, the ability to predict how ncRNAs translate into diseases is limited. This dissertation aimed to contribute to the functional understanding of ncRNAs by exploring experimental and computational approaches.

Project 1: Investigating the *in vivo* contribution of the *Crossfirre* locus alone and in combination with *Firre* and *Dxz4*

Prior to the thesis, the characterization of the lncRNA *Crossfirre*, including its involvement in XCI biology, remained entirely unexplored. Additionally, the contributions of *Firre* and *Dxz4* to random XCI in adult tissues have not been fully addressed. Finally, the impact of these X-linked loci on gene regulation and the phenotypic consequences in loss-of-function models, individually and in combination, have not been investigated. To address these knowledge gaps, the first project aimed to investigate the *in vivo* role of *Crossfirre*, *Firre,* and *Dxz4* by performing comprehensive multi-omics analyses and large-scale phenotypic characterization using one of the largest genetically modified X-linked mouse cohorts. This cohort included mouse models carrying: (i) a deletion of *Crossfirre, Firre, and Dxz4*, (ii) double deletions of *Crossfirre-Firre* and *Firre-Dxz4*, and (iii) a triple knockout (TKO) including the removal of *Crossfirre-Firre* and *Dxz4* (**Figure 2.1**).

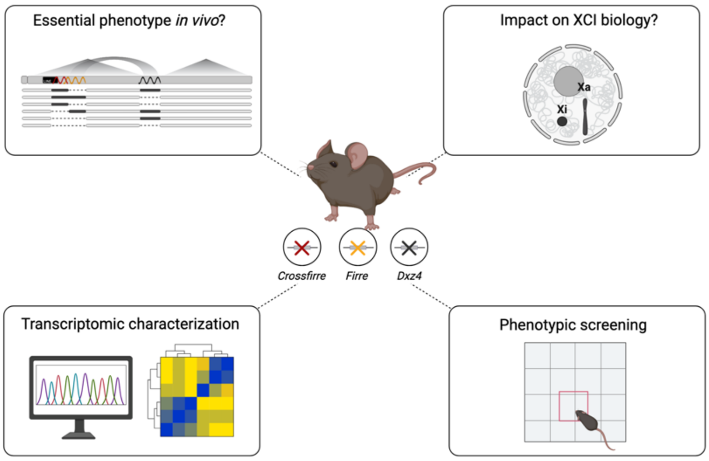


Figure 2.1 Schematic overview to investigate the *in vivo* role of *Crossfirre*, *Firre*, and *Dxz4*. The study investigated whether loss-of-function models lacking the *Crossfirre*, *Firre*, and *Dxz4* loci, individually and in combination, exhibit essential phenotypes *in vivo*. In addition, the role of these loci in X chromosome inactivation (XCI) biology was investigated, complemented by comprehensive transcriptomic and phenotypic analyses. Created in BioRender. Andergassen, D. (2025).

Project 2: Decoding the targets and mechanisms of the non-coding genome through allele-specific genomics

Due to the cost- and time-extensive nature of experimentally characterizing ncRNAs in the laboratory, the second project focused on predicting their target genes and mechanisms *in silico*. This computational approach aims to facilitate the selection of future candidate ncRNAs for experimental validation. ASE analyses, which compare allelic expression levels within the same cellular environment, provide a highly controlled and sensitive system to overcome gene dosage compensatory mechanisms and mitigate the dynamic expression patterns of ncRNAs. It is hypothesized that the allelic bias of a *cis*-acting regulatory ncRNA would be reflected in the allelic imbalance of the proximate targets. Consequently, ASE analyses provide a powerful tool to predict the regulatory targets and mechanisms of *cis*-acting ncRNAs. This project aimed to identify the regulatory *cis*-acting ncRNAs in mice and humans by developing a bioinformatics framework that predicts their targets and mechanisms based on the allelic expression patterns (**Figure 2.2**).

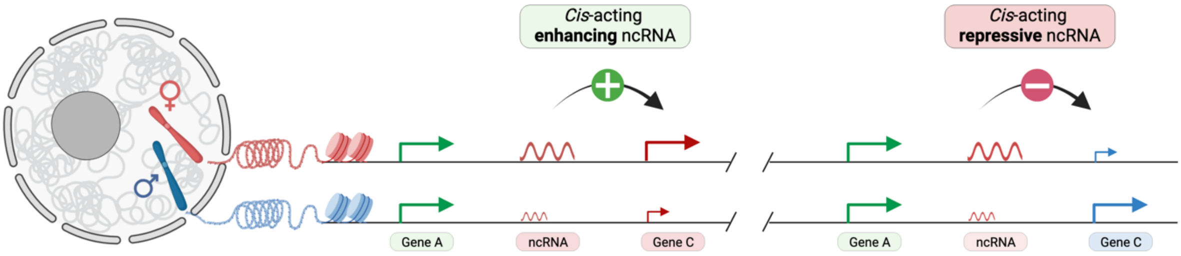

Cis-acting **enhancing** ncRNA
Cis-acting **repressive** ncRNA
♀
♂
Gene A
ncRNA
Gene C
Gene A
ncRNA
Gene C

Figure 2.2 Schematic overview of the allele-specific approach to predict ncRNA-targets. Schematic overview of the allele-specific concept to predict the regulatory targets of *cis*-acting ncRNAs. It is hypothesized that the allelic bias of a *cis*-acting regulatory ncRNA is reflected in the allelic imbalance of the proximate target. Depending on whether the allelic bias between ncRNA and pcGene is towards the same or opposite alleles, it is further assumed that the mechanism can be inferred as either enhancing or repressive. Created in BioRender. Andergassen, D. (2025) https://BioRender.com/b98a895.

3 Materials

3.1 Wet-lab materials

3.1.1 Chemicals, reagents, and consumables

Substances or consumables	Source
Chloroform (Trichloromethane, CHCl3)	Roth (Karlsruhe, Germany)
Dulbecco's Phosphate Buffered Saline	Thermo Fisher Scientific Inc. (Waltham, USA)
Ethanol ≥ 99,5%	Roth (Karlsruhe, Germany)
Isopropanol ≥ 99,8%	Roth (Karlsruhe, Germany)
RNase Zap	Sigma Aldrich (Taufkirchen, Germany)
RNase-free water	Thermo Fisher Scientific Inc. (Waltham, USA)
TRIzol Reagent	Thermo Fisher Scientific Inc. (Waltham, USA)

3.1.2 Primer

Primer name	Sequence	Source
fwd_Crossfirre_Crossfirre-firre	AGAACAGCCCTGGAGGAAAT	Sigma Aldrich
fwd_Dxz4	ACAGTGCATCAAAAGCACACG	Sigma Aldrich
fwd_Dxz4_WT	AGTTGGGAGCGAAGCAGAAA	Sigma Aldrich
rev_Crossfirre	GTAGGCAAGCCTGAGGAAAA	Sigma Aldrich
rev_Crossfirre_Crossfirre-firre_WT	TCTCTTGTAAGAGTTCCCATGTGT	Sigma Aldrich
rev_Crossfirre-firre	CCTGGGTCCTCTATAAAAGCAACAG	Sigma Aldrich
rev_Dxz4	CCTGGTGGCACAGAACTCTA	Sigma Aldrich

3.1.3 gRNAs

Name (target)	Protospacer+PAM	Source
gRNA_Crossfirre_up	GATCTTTACCCCACAGTATA**AGG**	Integrated DNA Technologies
gRNA_Crossfirre_down	GGGATGGCCACACCTCACAA**TGG**	Integrated DNA Technologies
gRNA_Crossfirre-firre_up	AATGGGTCCAGGTATTGGCG**GGG**	Integrated DNA Technologies

gRNA_Crossfirre-firre_down	CTAAAAGGATTAGGGTCTCT**TGG**	Integrated DNA Technologies
gRNA_Dxz4_up	CATGCTGCTTTTATGTGCTT**CGG**	Integrated DNA Technologies
gRNA_Dxz4_down	TACTGAAGGAATCGTATGAC**CGG**	Integrated DNA Technologies

3.1.4 Antibodies

Name	**Source**
Anti-mouse CD16/CD32 (Fc Block)	BD Biosciences
TER-119-PE antibodies	Thermo Fisher Scientific Inc. (Waltham, USA)
Zombie Green™ viability dye	BioLegend (San Diego, USA)

3.1.5 Mouse strains

Mouse strain	**Source**
CAST/EiJ	Jackson Laboratory, JAX: Strain #000928
C57BL/6J	Jackson Laboratory, JAX: Strain #000664
B6D2F1/J	Jackson Laboratory, JAX: Strain #100006

3.1.6 Kits

Kit name	**Source**
Chromium Next GEM Single Cell 3' Reagent Kits v3.1 (Dual Index)	10x Genomics (Pleasanton, USA)
Illumina® Stranded mRNA Prep Ligation Kit	Illumina (SanDiego, USA)
RNeasy mini columns	Qiagen (Düsseldorf, Germany)
TruSeq stranded Illumina®	Illumina (SanDiego, USA)
Unstranded TruSeq libraries	Illumina (SanDiego, USA)

3.1.7 Instruments

Devices	**Source**
Agilent 2100 Bioanalyzer	Agilent (Santa Clara, USA)

Agilent 4200 TapeStation	Agilent (Santa Clara, USA)
GentleMACS™ Dissociator	Miltenyi Biotec (Bergisch Gladbach, Germany)
HiSeq 2500	Illumina (SanDiego, USA)
NovaSeq 6000	Illumina (SanDiego, USA)
Qubit 2.0 Fluorometer	Thermo Fisher Scientific Inc. (Waltham, USA)
Vortex-Genie 2	Scientific industries (Bohemia, USA)

3.2 Bioinformatic requirements

3.2.1 Software

Software	**Version**	**Source**
Adobe Acrobat	2024.004.20272	Adobe Systems Incorporated (San Jose, USA)
Adobe Illustrator	25.4.1	Adobe Systems Incorporated (San Jose, USA)
Allelome.LINK	1.0	This thesis
Allelome.PRO	1.0	Andergassen *et al.*, 2015[201]
Allelome.PRO v2.0	2.0	This thesis
awk	20200816	Aho *et al.*, 1987[218]
bedtools	2.30.0	Quinlan *et al.*, 2010[219]
bowtie2	2.3.5.1	Langmead *et al.*, 2012[220]
cellranger	6.1.2	Zheng *et al.*, 2017[221]
curl	7.76.1	Hostetter *et al.*, 1997[222]
deeptools	3.3.0	Ramirez *et al.*, 2016[223]
fastqc	0.11.6	Andrews *et al.*, 2010[224]
gatk	3.8	McKenna *et al.*, 2010[225]
htseq	0.11.3	Anders *et al.*, 2015[226]
macs2	2.1.4	Zhang *et al.*, 2008[227]
Perl	5.32.1	Wall *et al.*, 1994[228]
Python	2.7	Python Software Foundation[229]
R	3.6.3	R Core Team 2023[230]
rseqc	2.6.4	Wang *et al.*, 2012[231]
samtools	1.12	Danecek *et al.*, 2021[232]
sed	4.8	N/A

sinto	0.8.1	https://timoast.github.io/sinto/index.html
SNPsplit	0.3.2	Krueger *et al.*, 2016[233]
sra-tools	2.11.0	SRA Toolkit Development Team
star	2.6.0c	Dobin *et al.*, 2013[194]
ucsc-bedtobigbed	377	Kent *et al.*, 2010[234]
ucsc-fetchchromsizes	377	Kent *et al.*, 2010[234]
ucsc-wigtobigwig	377	Kent *et al.*, 2010[234]
vcftools	0.1.16	Danecek *et al.*, 2011[235]

3.2.2 R packages

Package	**Version**	**Source**
AnnotationDbi	1.64.1	Pagès *et al.*, 2023[236]
AnnotationFilter	1.26.0	Morgan *et al.*, 2023[237]
AnnotationHub	3.10.1	Morgan *et al.*, 2024[238]
ape	5.8	Paradis *et al.*, 2019[239]
apeglm	1.24.0	Zhu *et al.*, 2019[240]
base	4.3.1	R Core Team[230]
beeswarm	0.4.0	Eklund *et al.*, 2021[241]
biomaRt	2.58.2	Durinck *et al.*, 2005[242]
BSgenome	1.70.2	Pagès *et al.*, 2024[243]
CePa	0.8.0	Gu *et al.*, 2022[244]
ChIPpeakAnno	3.36.1	Zhu *et al.*, 2013[245]
circlize	0.4.16	Gu *et al.*, 2014[246]
clusterProfiler	4.10.1	Yu *et al.*, 2012[247]
ComplexHeatmap	2.18.0	Gu *et al.*, 2016[248]
cowplot	1.1.3	Wilke *et al.*, 2024[249]
data.table	1.16.0	Barrett *et al.*, 2024[250]
DESeq2	1.42.1	Love *et al.*, 2014[251]
devtools	2.4.5	Wickham *et al.*, 2022[252]
dplyr	1.1.4	Wickham *et al.*, 2023[253]
EnhancedVolcano	1.20.0	Blighe *et al.*, 2023[254]
enrichplot	1.22.0	Yu *et al.*, 2023[255]

ensembldb	2.26.0	Rainer *et al.*, 2019[256]
eulerr	7.0.2	Larsson et al., 2024[257]
fdrtool	1.2.18	Klaus *et al.*, 2024[258]
gdata	3.0.0	Warnes *et al.*, 2023[259]
GenomeInfoDb	1.38.8	Arora *et al.*, 2024[260]
GenomeInfoDbData	1.2.11	Bioconductor Core Team, 2023[261]
GenomicAlignments	1.38.2	Lawrence *et al.*, 2013[262]
GenomicFeatures	1.54.4	Lawrence *et al.*, 2013[262]
GenomicRanges	1.54.1	Lawrence *et al.*, 2013[262]
ggbeeswarm	0.7.2	Clarke *et al.*, 2023[263]
ggbreak	0.1.2	Xu *et al.*, 2021[264]
ggExtra	0.10.1	Attali *et al.*, 2023[265]
ggforce	0.4.2	Pedersen *et al.*, 2024[266]
ggfun	0.1.6	Yu *et al.*, 2024[267]
ggnetwork	0.5.13	Briatte *et al.*, 2024[268]
ggplot2	3.5.1	Wickham *et al.*, 2016[269]
ggpubr	0.6.0	Kassambara *et al.*, 2023[270]
ggraph	2.2.1	Pedersen *et al.*, 2024[271]
ggrastr	1.0.2	Petukhov *et al.*, 2023[272]
ggrepel	0.9.6	Slowikowski *et al.*, 2024[273]
ggridges	0.5.6	Wilke *et al.*, 2024[274]
ggsci	3.2.0	Xiao *et al.*, 2024[275]
ggsignif	0.6.4	Ahlmann-Eltze *et al.*, 2021[276]
ggtree	3.10.1	Yu *et al.*, 2017[277]
GOSemSim	2.28.1	Yu *et al.*, 2010[278]
gplots	3.1.3.1	Warnes *et al.*, 2024[279]
grid	4.3.1	R Core Team, 2023[230]
gridBase	0.4-7	Murrell *et al.*, 2014[280]
gridExtra	2.3	Auguie *et al.*, 2017[281]
gridGraphics	0.5-1	Murrell *et al.*, 2020[282]
gtools	3.9.5	Warnes *et al.*, 2023[283]
igraph	2.0.3	Csardi *et al.*, 2006[284]

IRanges	2.36.0	Lawrence *et al.*, 2013[262]
karyoploteR	1.28.0	Gel *et al.*, 2017[285]
karyotapR	1.0.1	Mays *et al.*, 2023[286]
leiden	0.4.3.1	Kelly *et al.*, 2023[287]
leidenbase	0.1.31	Ewing *et al.*, 2024[288]
limma	3.58.1	Wu *et al.*, 2012[289]
matrixStats	1.4.1	Bengtsson *et al.*, 2024[290]
MuDataSeurat	0.0.1.0	Bredikhin *et al.*, 2023[291]
org.Hs.eg.db	3.18.0	Carlson *et al.*, 2023[292]
org.Mm.eg.db	3.18.0	Carlson *et al.*, 2023[292]
pheatmap	1.0.12	Kolde *et al.*, 2019[293]
plyr	1.8.9	Wickham *et al.*, 2011[294]
png	0.1-8	Urbanek *et al.*, 2022[295]
qvalue	2.34.0	Storey *et al.*, 2023[296]
RColorBrewer	1.1-3	Neuwirth *et al.*, 2022[297]
readr	2.1.5	Wickham *et al.*, 2024[298]
readxl	1.4.3	Wickham *et al.*, 2023[299]
reprex	2.1.1	Bryan *et al.*, 2024[300]
Rsamtools	2.18.0	Morgan *et al.*, 2023[301]
scales	1.3.0	Wickham *et al.*, 2023[302]
Seurat	4.1.1	Hao *et al.*, 2021[303]
simplifyEnrichment	1.12.0	Gu *et al.*, 2023[304]
stringr	1.5.1	Wickham *et al.*, 2023[305]
SummarizedExperiment	1.32.0	Morgan *et al.*, 2023[306]
sctransform	0.4.1	Hafemeister *et al.,* 2019[307]
tibble	3.2.1	Müller *et al.*, 2023[308]
tidygraph	1.3.1	Pedersen *et al.*, 2024[309]
tidyr	1.3.1	Wickham *et al.*, 2024[310]
tidyverse	2.0.0	Wickham *et al.*, 2019[311]
UpSetR	1.4.0	Gehlenborg *et al.*, 2019[312]
VennDiagram	1.7.3	Chen *et al.*, 2022[313]
writexl	1.5.0	Ooms *et al.*, 2024[314]

3.2.3 Public data, databases, and annotations

Name	Source
ENCODE blacklist genes	Amemiya *et al.*, 2019[315]
GENCODE M25 GRCm38.p6 20191146	Frankish *et al.*, 2019[316]
GENCODEv26 annotation	Frankish *et al.*, 2019[316]
Gene expression omnibus	Edgar *et al.*, 2002[317]
GTEx v8 release	GTEx Consortium atlas, 2020[190]
GTEx_v8_finemapping_DAPG.txt	Wen *et al.*, 2017[318]
mm10 genome (version 2020-A)	Zheng *et al.*, 2017[221]
Molecular Signatures Database	Subramanian *et al.*, 2005[206]
NHGRI-EBI GWAS Catalog v1.0	Sollis *et al.*, 2023 [215]
phASER_WASP_GTEx_v8_matrix.gw_phased.txt	Castel *et al.*, 2020[191]
RefSeq gene annotation GRCm38/mm10 (2018)	O'Leary *et al.*, 2016[319]
Sanger database	Keane *et al.*, 2011[189]
SNP file CAST x BL6 (15,438,314 variants)	Andergassen *et al.*, 2019[127]
SNP file CAST x FVB (16,988,479 variants)	Andergassen *et al.*, 2019[175]

3.2.4 Public sequencing data

Name	Sample identifier	Source
Pl_FixC_++_1	GSM3636720	GSE127554[127]
Pl_FixC_++_2	GSM3636721	GSE127554[127]
Pl_FixC_++_3	GSM3636722	GSE127554[127]
Pl_DxC_++_1	GSM3636709	GSE127554[127]
Pl_DxC_++_2	GSM3636710	GSE127554[127]
Pl_FDxC_++_1	GSM3636697	GSE127554[127]
Pl_FDxC_++_2	GSM3636698	GSE127554[127]
Pl_FDxC_++_3	GSM3636699	GSE127554[127]
Pl_FixC_-+_1	GSM3636714	GSE127554[127]
Pl_FixC_-+_2	GSM3636715	GSE127554[127]
Pl_FixC_-+_3	GSM3636716	GSE127554[127]
Pl_DxC_-+_1	GSM3636703	GSE127554[127]

Pl_DxC_-+_2	GSM3636704	GSE127554[127]
Pl_DxC_-+_3	GSM3636705	GSE127554[127]
Pl_FDxC_-+_1	GSM3636691	GSE127554[127]
Pl_FDxC_-+_2	GSM3636692	GSE127554[127]
Pl_FDxC_-+_3	GSM3636693	GSE127554[127]
Pl_CxFi_++_1	GSM3636741	GSE127554[127]
Pl_CxFi_++_2	GSM3636742	GSE127554[127]
Pl_CxFi_++_3	GSM3636743	GSE127554[127]
Pl_CxD_++_1	GSM3636729	GSE127554[127]
Pl_CxD_++_2	GSM3636730	GSE127554[127]
Pl_CxD_++_3	GSM3636731	GSE127554[127]
Pl_CxFD_++_1	GSM3636735	GSE127554[127]
Pl_CxFD_++_2	GSM3636736	GSE127554[127]
Pl_CxFD_++_3	GSM3636737	GSE127554[127]
Pl_CxFi_+-_1	GSM3636738	GSE127554[127]
Pl_CxFi_+-_2	GSM3636739	GSE127554[127]
Pl_CxFi_+-_3	GSM3636740	GSE127554[127]
Pl_CxD_+-_1	GSM3636726	GSE127554[127]
Pl_CxD_+-_2	GSM3636727	GSE127554[127]
Pl_CxD_+-_3	GSM3636728	GSE127554[127]
Pl_CxFD_+-_1	GSM3636732	GSE127554[127]
Pl_CxFD_+-_2	GSM3636733	GSE127554[127]
Pl_CxFD_+-_3	GSM3636734	GSE127554[127]
Br_FDxFD_--_1	GSM3636580	GSE127554[127]
Br_FDxFD_--_2	GSM3636581	GSE127554[127]
Br_FDxFD_--_3	GSM3636582	GSE127554[127]
Br_FDxFD_--_4	GSM3636583	GSE127554[127]
Br_WT_++_1	GSM3636584	GSE127554[127]
Br_WT_++_2	GSM3636585	GSE127554[127]
Br_WT_++_3	GSM3636586	GSE127554[127]
Br_WT_++_4	GSM3636587	GSE127554[127]
He_FDxFD_--_1	GSM3636588	GSE127554[127]

He_FDxFD_--_2	GSM3636589	GSE127554[127]
He_FDxFD_--_3	GSM3636590	GSE127554[127]
He_FDxFD_--_4	GSM3636591	GSE127554[127]
He_WT_++_1	GSM3636592	GSE127554[127]
He_WT_++_2	GSM3636593	GSE127554[127]
He_WT_++_3	GSM3636594	GSE127554[127]
He_WT_++_4	GSM3636595	GSE127554[127]
Ki_FDxFD_--_1	GSM3636596	GSE127554[127]
Ki_FDxFD_--_2	GSM3636597	GSE127554[127]
Ki_FDxFD_--_3	GSM3636598	GSE127554[127]
Ki_FDxFD_--_4	GSM3636599	GSE127554[127]
Ki_WT_++_1	GSM3636600	GSE127554[127]
Ki_WT_++_2	GSM3636601	GSE127554[127]
Ki_WT_++_3	GSM3636602	GSE127554[127]
Ki_WT_++_4	GSM3636603	GSE127554[127]
Li_FDxFD_--_1	GSM3636604	GSE127554[127]
Li_FDxFD_--_2	GSM3636605	GSE127554[127]
Li_FDxFD_--_3	GSM3636606	GSE127554[127]
Li_FDxFD_--_4	GSM3636607	GSE127554[127]
Li_WT_++_1	GSM3636608	GSE127554[127]
Li_WT_++_2	GSM3636609	GSE127554[127]
Li_WT_++_3	GSM3636610	GSE127554[127]
Li_WT_++_4	GSM3636611	GSE127554[127]
Lu_FDxFD_--_1	GSM3636617	GSE127554[127]
Lu_FDxFD_--_2	GSM3636618	GSE127554[127]
Lu_FDxFD_--_3	GSM3636619	GSE127554[127]
Lu_FDxFD_--_4	GSM3636620	GSE127554[127]
Lu_WT_++_1	GSM3636621	GSE127554[127]
Lu_WT_++_2	GSM3636622	GSE127554[127]
Lu_WT_++_3	GSM3636623	GSE127554[127]
Lu_WT_++_4	GSM3636624	GSE127554[127]
Sp_FDxFD_--_1	GSM3636625	GSE127554[127]

Sp_FDxFD_--_2	GSM3636626	GSE127554[127]
Sp_FDxFD_--_3	GSM3636627	GSE127554[127]
Sp_FDxFD_--_4	GSM3636628	GSE127554[127]
Sp_WT_++_1	GSM3636629	GSE127554[127]
Sp_WT_++_2	GSM3636630	GSE127554[127]
Sp_WT_++_3	GSM3636631	GSE127554[127]
Sp_WT_++_4	GSM3636632	GSE127554[127]
Sp_DxD_--_1	GSM3636633	GSE127554[127]
Sp_DxD_--_2	GSM3636634	GSE127554[127]
Sp_FixFi_--_1	GSM3636635	GSE127554[127]
Sp_FixFi_--_2	GSM3636636	GSE127554[127]
Sp_FixFi_--_3	GSM3636637	GSE127554[127]
Pl_E12_5_CF_1	GSM1970843	GSE75957[78]
Pl_E12_5_CF_2	GSM1970844	GSE75957[78]
Pl_E12_5_FC_1	GSM1970845	GSE75957[78]
Pl_E12_5_FC_2	GSM1970846	GSE75957[78]
Pl_E12_5_CxRSDel_++_3	SRR8753471	GSE128513[175]
Pl_E12_5_CxRSDel_++_2	SRR8753472	GSE128513[175]
Pl_E12_5_CxRSDel_++_1	SRR8753473	GSE128513[175]
Pl_E12_5_CxRSDel_+-_3	SRR8753474	GSE128513[175]
Pl_E12_5_CxRSDel_+-_2	SRR8753475	GSE128513[175]
Pl_E12_5_CxRSDel_+-_1	SRR8753476	GSE128513[175]
NPC_XX2	SRR3933589	GSE84646[320]
NPC_XX4	SRR3933595	GSE84646[320]
Female_Spleen_Rep2	SRR8119821	PRJNA497970[321]
Female_Spleen_Rep1	SRR8119822	PRJNA497970[321]
Male_Spleen_Rep2	SRR8119826	PRJNA497970[321]
Male_Spleen_Rep1	SRR8119827	PRJNA497970[321]
Male_Kidney_Rep2	SRR8119832	PRJNA497970[321]
Male_Kidney_Rep1	SRR8119833	PRJNA497970[321]
Male_Heart_Rep2	SRR8119834	PRJNA497970[321]
Male_Heart_Rep1	SRR8119835	PRJNA497970[321]

Male_Cerebrum_Rep2	SRR8119836	PRJNA497970[321]
Male_Cerebrum_Rep1	SRR8119837	PRJNA497970[321]
Male_Liver_Rep2	SRR8119838	PRJNA497970[321]
Male_Liver_Rep1	SRR8119839	PRJNA497970[321]
Female_Cerebrum_Rep1	SRR8119850	PRJNA497970[321]
Female_Cerebrum_Rep2	SRR8119851	PRJNA497970[321]
Female_Liver_Rep1	SRR8119852	PRJNA497970[321]
Female_Liver_Rep2	SRR8119853	PRJNA497970[321]
Female_Lung_Rep1	SRR8119854	PRJNA497970[321]
Female_Lung_Rep2	SRR8119855	PRJNA497970[321]
Female_Heart_Rep1	SRR8119856	PRJNA497970[321]
Female_Heart_Rep2	SRR8119857	PRJNA497970[321]
Female_Kidney_Rep1	SRR8119858	PRJNA497970[321]
Female_Kidney_Rep2	SRR8119859	PRJNA497970[321]
Male_Lung_Rep1	SRR8119864	PRJNA497970[321]
Male_Lung_Rep2	SRR8119865	PRJNA497970[321]
aBr_CF_1	SRR3085966	GSE75957[78]
aBr_CF_2	SRR3085967	GSE75957[78]
aBr_FC_1	SRR3085968	GSE75957[78]
aBr_FC_2	SRR3085969	GSE75957[78]
MEF_K4m3_CF_1	SRR2038034	GSE69168[201]
MEF_K4m3_CF_2	SRR2038035	GSE69168[201]
MEF_K4m3_FC_1	SRR2038036	GSE69168[201]
MEF_K4m3_FC_2	SRR2038037	GSE69168[201]

4 Methods

Parts of the methods described in the sections of chapters **4**, including **4.1** (**4.1.1**, **4.1.2**), **4.2** (**4.2.1**, **4.2.2**), **4.3**, **4.5.** (**4.5.1**, **4.5.2**, **4.5.3**), and **4.6** have been previously published in a similar form by the author of this thesis[322]. Additionally, methods of the chapter **4**, including **4.1** (**4.1.2**), **4.2** (**4.2.1**), **4.4** (**4.4.1**, **4.4.2**), **4.5** (**4.5.2**, **4.5.4**) and **4.6** have been described similarly by the author of this thesis in a submitted manuscript (see **9.2 Submitted manuscripts, 1.**).

As a bioinformatician, my task in the projects was on the computational biology, including the development of analysis pipelines and the execution of downstream analyses following next-generation sequencing (NGS). Therefore, the wet lab procedures comprising the generation of knockout mouse models and NGS sample preparation have been carried out in collaboration. However, brief descriptions of all wet lab steps are included to provide a comprehensive overview of the entire experimental framework.

4.1 Animal studies

Animals were housed in pathogen-free environments at Harvard University's Biological Research Infrastructure and the Institute of Pharmacology and Toxicology at the Technical University of Munich. All animal experiments conducted at the Institute of Pharmacology and Toxicology followed the EU guideline 2010/63 and the German Animal Welfare Act (Tierschutzgesetz and Tierschutzversuchstierverordnung). Approval was granted by the District Administrative Office of the City of Munich, Veterinary Office of the City of Munich, in accordance with Section 11, Paragraph 1, Sentence 1, No. 1 of the German Animal Welfare Act.

4.1.1 Generation of *Crossfirre*, *Firre,* and *Dxz4* knockout mouse models

To investigate the *in vivo* effects of the *Crossfirre* locus alone and in combination with *Firre* and *Dxz4* in Project 1, three knockout mouse models were generated in collaboration: (i) a single *Crossfirre* deletion encompassing the 50kb LINE cluster attached to the gene locus (Δ*Crossfirre*), (ii) a double deletion of the *Crossfirre* and *Firre* loci (Δ*Crossfirre-Firre*), and (iii) a triple deletion of *Crossfirre*, *Firre*, and *Dxz4*.

The Δ*Crossfirre* and Δ*Crossfirre-Firre* knockout mouse models were generated as previously described for Δ*Firre* and Δ*Dxz4*[127,131]. CAST/EiJ (CAST), B6D2F1/J (F1 BL6 and DBA/2J), and C57BL/6J (BL6) mouse strains were obtained from the Jackson Laboratory. Zygotes of pronuclear stage 3 were isolated from superovulated B6D2F1/J females mated with BL6 males[323]. Cas9 mRNA (200 ng/µl) and two guide RNAs flanking each locus (50 ng/µl) were co-injected into the zygotes, which were cultured to the blastocyst stage and implanted into pseudopregnant CD-1 females (Charles River)[127,323]. Progenies were screened for the deletions using polymerase chain reaction (PCR) and Sanger sequencing. The sequences of the PCR primers and guide RNAs (gRNAs) are listed in the **Materials sections 3.1.2** and **3.1.3**.

In addition, TKO mouse models were generated by crossing ∆*Crossfirre-Firre* males with ∆*Dxz4* females[127]. As a result, female offspring inherited the *Crossfirre-Firre* deletion on the paternal X chromosome and the *Dxz4* knockout on the maternal X chromosome. To obtain mouse models with all three deletions on the same X chromosome, females were further mated with BL6 males and offspring were screened for meiotic recombination between *Crossfirre-Firre* and *Dxz4*. To minimize strain bias and CRISPR-Cas9 off-target effects, founder mice (75% BL6 background) were backcrossed twice with BL6 mice, resulting in an expected BL6 strain background of 93%. Similarly, wildtype (WT) controls were generated by backcrossing the founder mice to match the strain background of the knockout mouse models[127]. For the phenotypic analysis at the German Mouse Clinic (GMC), TKO mouse models underwent two additional backcrosses, resulting in an expected BL6 background of 98%.

All knockout mice were analyzed with the previously published ∆*Firre* and ∆*Dxz4* single-deletion and the ∆*Firre-Dxz4* double-deletion mouse models[127,131]. Combined, this set of mutants provides a comprehensive framework for examining the *in vivo* contributions of the X-linked LINE cluster, the megastructures and open chromatin specific to Xi, and the Xa-specific expression of *Crossfirre*, *Firre*, and *Dxz4*.

4.1.2 Collection of tissue samples

For Project 1, one of the objectives was to investigate the effect of the *in vivo* deletions of *Crossfirre*, *Firre*, and *Dxz4* on imprinted XCI. Therefore, reciprocal crosses between WT CAST and BL6 mutants (∆*Crossfirre*, ∆*Crossfirre-Firre*, TKO) were performed to generate mice with deletions on either the Xi (CAST x BL6) or Xa (BL6 x CAST). Placentas were harvested at E12.5 from three biological replicates per genotype, resulting in the following number of samples: ∆*Crossfirre*: Xa *n* = 3, Xi *n* = 3; ∆*Crossfirre-Firre*: Xa *n* = 3, Xi *n* = 3; TKO: Xa *n* = 3, Xi *n* = 3. In conjunction with these samples, E12.5 placentas from the previously published ∆*Dxz4*, ∆*Firre*, and ∆*Firre-Dxz4* mouse models, as well as sample-matched WT data, were reanalyzed from Andergassen *et al.*[127] (Δ*Firre*: Xa *n* = 3, Xi *n* = 3; Δ*Dxz4*: Xa *n* = 3, Xi *n* = 3; Δ*Firre-Dxz4*: Xa *n* = 3, Xi *n* = 3; WT: BL6 x CAST *n* = 8; CAST x BL6 *n* = 9).

For further investigation of the TKO effect on random XCI, homozygous TKO females were mated to CAST males. Thus, heterozygous TKO (-/+ TKO x CAST) and WT (+/+ TKO x CAST) mice were obtained as littermates. Spleens of female F1 offspring were harvested at six weeks of age and used for scRNA-seq (TKO *n* = 1, WT *n* = 1) and bulk RNA-seq (TKO *n* = 3, WT *n* = 3).

Furthermore, samples were collected from adult homozygous TKO mice (-/- TKO, *n* = 3) to generate a transcriptomic bodymap. Three TKO mice (-/- TKO, BL6) were sacrificed at six weeks of age, and organs, including the brain, liver, lung, kidney, heart, and spleen, were isolated (*n* = 3 per tissue). Additionally, spleens were collected from six-week-old Δ*Crossfirre* (*n* = 3) and Δ*Crossfirre-Firre* (*n* = 2) mouse models. Sample-matched transcriptomic data from the previously published Δ*Firre*, Δ*Firre-Dxz4*, and WT tissue samples were included in the analyses[127].

To identify the protein-coding target genes of ncRNAs in Project 2, F1 hybrid mice (BL6 x CAST) were generated by crossing BL6 females with CAST males. At nine weeks of age, the liver, heart, kidney, spleen, brain, and lung were harvested from female mice (*n* = 3 per tissue), resulting in 18 tissue samples. All samples were snap-frozen in liquid nitrogen and stored at -80°C.

4.2 RNA extraction and library preparation

4.2.1 Sample preparation for bulk RNA-seq

For Project 1, RNA was extracted from Δ*Crossfirre*, Δ*Crossfirre-Firre*, and TKO tissue samples (*n* = 41) using TRIzol lysates and RNeasy mini columns (Qiagen). Total mRNA was used to generate strand-specific PolyA+ mRNA libraries for placentas (TruSeq stranded Illumina, *n* = 18) and unstranded TruSeq libraries for six-week-old adult organs (*n* = 23). A Qubit 2.0 Fluorometer was used to assess library concentrations and an Agilent 2100 Bioanalyzer to determine library fragment size and purity. Sequencing was performed at the Harvard University Bauer Sequencing Core on a HiSeq 2500 (75bp paired-end).

For heterozygous TKO and WT spleen samples (*n* = 6), strand-specific libraries were generated using the Illumina Stranded mRNA Prep Ligation Kit. The Agilent's TapeStation System was employed to assess library concentrations, and sequencing was performed at Helmholtz Munich using a NovaSeq 6000 (50bp paired-end). For Project 2, sequencing libraries were generated from tissue samples (brain, liver, lung, heart, spleen, kidney) of nine-week-old F1 hybrid mice (BL6 x CAST, *n* = 18). Individual samples (50-100 mg) were homogenized in 1 ml TRIzol using the GentleMACS Dissociator (program RNA_02_0). The isolation of RNA was performed as described in the manufacturer's instructions (Invitrogen, TRIzol Reagent, Cat. #15596018) with 1 ml of homogenized tissue solution. Subsequently, 100 ng of RNA and the Illumina Stranded mRNA Prep Ligation Kit were used to generate poly-A captured sequencing libraries. The fragment length and concentration of the RNA-seq libraries were evaluated using the Agilent TapeStation System, and sequencing was conducted at Helmholtz Munich on a NovaSeq 6000 platform (50bp paired-end).

4.2.2 Sample preparation for scRNA-seq

For Project 1, scRNA-seq was performed on spleen samples obtained from heterozygous TKO mice (-/+ TKO x CAST, *n* = 1) and WT littermates (+/+ TKO x CAST, *n* = 1). Spleens were harvested from six-week-old female mice and dissociated between glass slides to generate a single-cell suspension. The suspension was strained through 70 µm and 30 µm filters and incubated with an Fc-blocker for 15 minutes to prevent non-specific antibody binding. Subsequently, cells were stained with Zombie Green (Viability, BioLegend) to assess cell viability and TER-119-PE antibodies (Erythrocytes, ThermoFisher) to label erythrocytes. Each sample was incubated with Cell Multiplexing Oligos (10x) to add unique barcodes and pool samples into a single 10x reaction. Following this step, cells were subjected to fluorescence-activated cell sorting (FACS) to remove non-viable cells (Zombie Green-positive) and erythrocytes (TER-119-positive). The viable, non-erythrocytic cells (Zombie Green-negativ, TER-119-negative) were quantified and used to generate a single-cell library with the Chromium Next GEM Single Cell 3’ Reagent Kits v3.1 (Dual Index) featuring Feature Barcode technology for Cell Multiplexing (10x). Sequencing was performed at Helmholtz Munich using the NovaSeq 6000 platform.

4.3 Phenotypic analysis with the German Mouse Clinic

A cohort of 30 WT (male *n* = 15, female *n* = 15) and 26 TKO (male *n* = 13, female *n* = 13) mice underwent a primary phenotypic screening at the GMC[207,324]. The GMC provides large-scale phenotyping services for mouse mutants, conducting over 550 standardized phenotyping tests across a broad range of categories. These categories include cardiovascular health, clinical chemistry, pathology, behavior, metabolism, immunology/allergy, dysmorphology, biomarkers, eyes, neurology, and nociception.

Mice were housed in individually ventilated cages and were considered pathogen-free in agreement with the Federation of European Laboratory Animal Science Associations (FELASA) recommendations. Adherent to the GMC housing conditions and the directive 2010/63/EU German national law, all animals had access to standard mouse chow and water. The authority of the district government of Upper Bavaria approved all animal experiments.

4.4 Development of bioinformatics pipelines

To facilitate the bioinformatic analysis of NGS data for ASE identification and ncRNA-target gene predictions, the Allelome.PRO v2.0 and Allelome.LINK pipelines were developed as part of the projects (**Figure 4.1**). A detailed manual for Allelome.PRO v2.0 and Allelome.LINK is provided in the **Appendix section 10.1**. Both pipelines are available at the GitHub page of the Andergassen Lab (https://github.com/AndergassenLab/Allelome.LINK).

4.4.1 Updating the Allelome.PRO pipeline to Allelome.PRO v2.0

The previously published Allelome.PRO pipeline is a bioinformatics tool that processes high-throughput sequencing data to identify allele-specific genomic patterns[201]. The pipeline was initially developed to classify gene loci of F1 hybrid mice into biallelic, imprinted, strain-biased, or non-informative. As a result, the tool requires sequencing data from reciprocal crosses, making it unsuitable for single samples. To extend the utility of Allelome.PRO to individual samples, single cells, and humans, the pipeline has been updated to Allelome.PRO v2.0.

To allow the analysis of single samples, the following input requirements were removed from the primary pipeline: main_title, ratio, y_axis, fdr_param, strains, for_c1, for_c2, rev_c1, rev_c2. Additionally, the classification scheme has been revised to categorize loci as allele-specific or biallelic, removing the previous distinction between strain-biased and imprinted. Accordingly, the output categories Imprinted: Maternal (MAT), Imprinted: Paternal (PAT), Strain bias: Strain 1, Strain bias: Strain 2, Not informative, and No SNP have been removed from the pipeline.

Furthermore, Allelome.PRO has been simplified by omitting the false discovery rate (FDR)-based mock comparison and the user-defined ratio filter, as well as the output files: <name>_IG.txt, <name>_SG.txt, <name>_locus_full.txt, <name>_SNP_full.txt, <name>.pdf, and info.txt. A new option to filter loci based on total read coverage has been added to the pipeline to enhance computational efficiency. Moreover, a comprehensive log file was introduced to facilitate the tracking and debugging process. The scripts pileup_filter.pl, read_count.pl, bed_creator_SNP.sh, and bed_creator.sh have been written in R and integrated into the score.R script to reduce complexity. Furthermore, Allelome.PRO v2.0 has been enhanced with a user-friendly interface by implementing direct command-line parsing. A detailed overview of the updated pipeline is provided in **Figure 4.1a** and **Appendix section 10.1**.

4.4.2 Developing the Allelome.LINK extension tool

The Allelome.LINK pipeline was developed as an extension tool of Allelome.PRO v2.0 to facilitate the target and mechanism prediction of *cis*-acting genomic loci based on the allele-specific pattern. The pipeline was written using the R programming language and is designed to accept the output of Allelome.PRO v2.0 (locus_table.txt) as input. Subsequently, the tool links allele-specific loci within a user-defined range based on ASE (**Figure 4.1b**).

First, the input data is filtered for genomic positions with sufficient read coverage and allelic bias using user-defined cutoff values. Informative loci are intersected and linked if they occur within predefined window sizes. The regulatory mechanism between interaction sites is inferred as either enhancing or repressive, depending on the correlation or anti-correlation of the allelic bias towards the same or opposite alleles.

A linkage score (LS) is calculated to rank individual linkages using the following equation:

$$LS = \log_{10}\left(\min_{|AS_1|,|AS_2|} +1\right) \times (1 - |\Delta\mathrm{AR}|) \qquad (1)$$

Here, AS refers to the allelic score calculated by Allelome.PRO v2.0. This score is derived from a binomial test using the number of maternal and paternal reads, with an assumed probability of 0.5[201]. The resulting *p*-value indicates the likelihood of an allelic bias at a given locus. To increase the robustness of the linkage score, Allelome.LINK utilizes the minimum AS from both loci, adjusted by 1 to avoid the decadic logarithm of zero. The adjusted value is then multiplied by 1−|ΔAR|, where ΔAR is the difference in the allelic ratios calculated as:

$$\Delta\mathrm{AR} = |AR_1 - 0.5| - |AR_2 - 0.5| \qquad (2)$$

The AR of both loci are centered by subtracting 0.5 before calculating their absolute difference. This method ensures that equal weight is assigned to allele-specific biases toward maternal and paternal alleles. As a result, similar allelic ratios will yield a small ΔAR, exerting a more pronounced impact on the linkage score. This approach is based on the assumption that co-regulated loci exhibit similar changes in their allelic ratios.

As an output, the Allelome.LINK pipeline generates a folder with linkage tables in text format, including all candidate predictions, as well as BED and BEDPE files for intuitive genome browser visualization. Furthermore, a log file is generated to track the analysis run. The Allelelome.PRO v2.0 and Allelome.LINK pipelines can be exerted by a simple one-line command. Detailed instructions for using the pipelines are provided in the **Appendix section 10.1**.

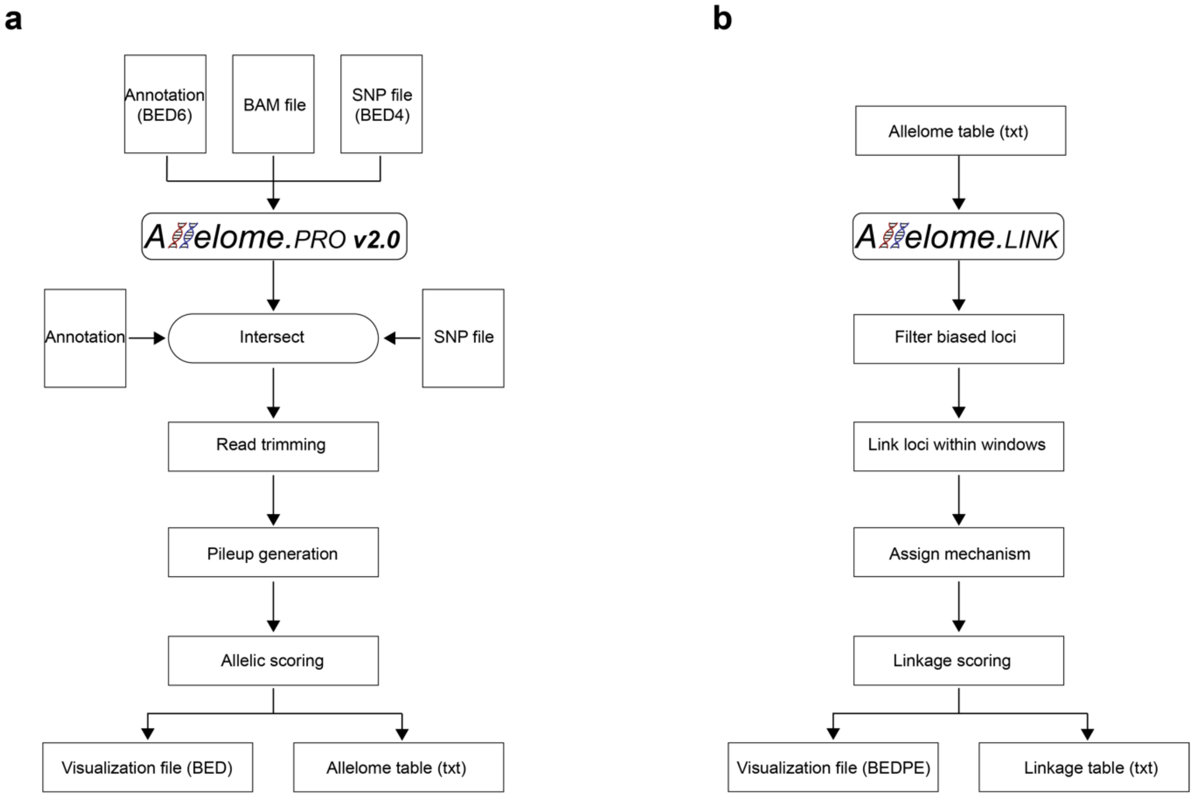


Figure 4.1 Summary of the allele-specific analysis pipelines.
a, Overview of the Allelome.PRO v2.0 pipeline. As input files, Allelome.PRO v2.0 requires a sample BAM file, a SNP file (BED4) and an annotation file (BED6). The pipeline starts by intersecting the SNP and annotation files, followed by read trimming to ensure that each read covers only one SNP. It then generates a pileup file, which records the variants of the reads and is used to calculate the allelic score for each locus. In addition, the pipeline produces a genome browser visualization file (BED) and a classification table containing the allelic scores for informative loci (locus_table.txt).
b, Overview of the Allelome.LINK pipeline. The Allelome.LINK pipeline starts by using the locus_table.txt output of an Allelome.PRO v2.0 run as input. The data is filtered to include only loci with sufficient read coverage and allelic bias according to user-defined thresholds. Allele-specific regions that co-occur within specified window sizes are linked with each other. Each interaction is classified as enhancing or repressive based on allelic correlation or anti-correlation. The output is a genome browser visualization file (BEDPE) and a linkage table with the predicted linkage information.

4.5 Bioinformatic analysis

All bioinformatics scripts associated with Project 1 and the bioinformatics pipelines for Project 2 are publicly available at the GitHub page of the Andergassen Lab (https://github.com/AndergassenLab/).

4.5.1 Analysis of ATAC-seq data

Two public ATAC-sequencing (ATAC-seq) datasets were downloaded from the Gene Expression Omnibus (GEO) database and used for Project 1. The first dataset was derived from the Sequence Read Archive (SRA) project PRJNA497970 and comprised ATAC-seq data from six different organs of eight-week-old BL6 mice of both sexes (lung, cerebrum, spleen, liver, heart, and kidney)[321]. Two replicates were obtained per tissue and sex (males *n* = 12, females *n* = 12). The second dataset includes ATAC-seq of two samples of clonal F1 neural progenitor cells (129S1/SvImJ x CAST F1) downloaded from the study GSE84646[320]. The corresponding accession numbers are listed in **Materials section 3.2.4**. A complete overview of the steps comprising the ATAC-seq workflow is shown in **Figure 4.2**.

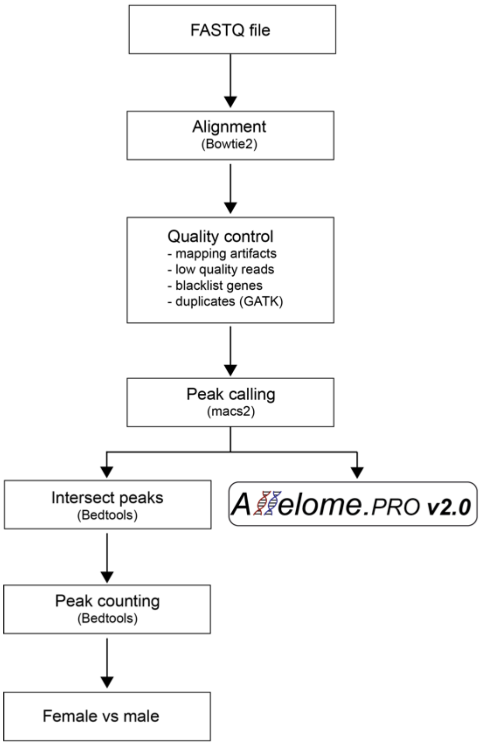


Figure 4.2 Overview of the ATAC-seq workflow.
The ATAC-seq workflow starts with the alignment of raw FASTQ files using the Bowtie2 aligner. Aligned data is subjected to quality control, which includes the removal of mapping artifacts, low-

quality reads, blacklist genes, and duplicates. Processed data is then used to call ATAC peaks using macs2 and the callpeak function. The identified peaks are used as annotation for the Allelome.PRO v2.0 pipeline to perform allele-specific analyses. Additionally, peaks are intersected by tissue and sex to quantify their distribution across the genome. A binomial test compares the epigenetic profiles between females and males.

4.5.1.1 Data alignment and quality control

Public ATAC-seq data was aligned using the Bowtie2 aligner[220] with the default parameters for paired-end data. An index reference was created using the bowtie2-build command and the GENCODE M25 GRCm38.p6 20191146 reference[316]. Post-alignment, quality control procedures were implemented to ensure data integrity. This process included the removal of mapping artifacts with bp lengths ≥ 2000 or ≤ 38, mitochondrial and low-quality reads (MAPQ < 20). In addition, ENCODE blacklist genes (blacklist.v2.bed)[315] and duplicates identified by GATK MarkDuplicates (version 4.1.0.0)[225] were excluded for downstream analysis.

4.5.1.2 Peak calling and epigenetic profiling

After quality control, the processed data was subjected to broad peak calling using macs2 callpeak[227]. Peaks identified in the organs of eight-week-old BL6 mice[321] were intersected by tissue and sex. The number of peaks within 100kb sliding windows (50kb overlap) was quantified across the whole genome using bedtools and the intersectBed command[219]. A binomial test was employed to calculate $\log_{10}$ *p*-values, using the median number of peaks per window across all tissues with an expected probability of 0.5 to compare the epigenetic profile between females and males. Peaks that were more abundant in females were assigned positive values, while peaks that were more prevalent in males were assigned negative values.

4.5.1.3 Allele-specific analysis of neural progenitor cells

Allele-specific analysis was performed on samples derived from clonal F1 neural progenitor cells[320]. A SNP file containing 20,563,466 variants was generated for the 129S1/SvImJ and CAST strains using SNP information obtained from the Sanger database (mgp.v5.merged.snps_all.dbSNP142.vcf)[189]. As an annotation file, a 50kb sliding window with a 25kb overlap was generated using the mm10 genome as a reference. Both files were used as input for the Allelome.PRO v2.0 analysis, which was run to obtain the allele-specific information for each sliding window, utilizing a read cutoff of ≥ 1 read per SNP and a total read cutoff of ≥ 50 reads per gene.

4.5.2 Analysis of bulk RNA-seq data

Bulk RNA-seq data was sequenced at the Helmholtz Munich Genomics Core Facility and the Harvard University Bauer Sequencing Core. The sequencing facilities performed de-multiplexing and adapter trimming of the raw data, and the resulting raw FASTQ files were provided per sample. Furthermore, publicly available RNA-seq data was downloaded from the GEO database from the following studies: Andergassen *et al.*, 2017[78], 2019[175], 2019[127]. All SRA projects and the corresponding accession numbers are listed in the **Materials section 3.2.4**. A detailed overview of the workflow for processing bulk RNA-seq data is provided in **Figure 4.3**.

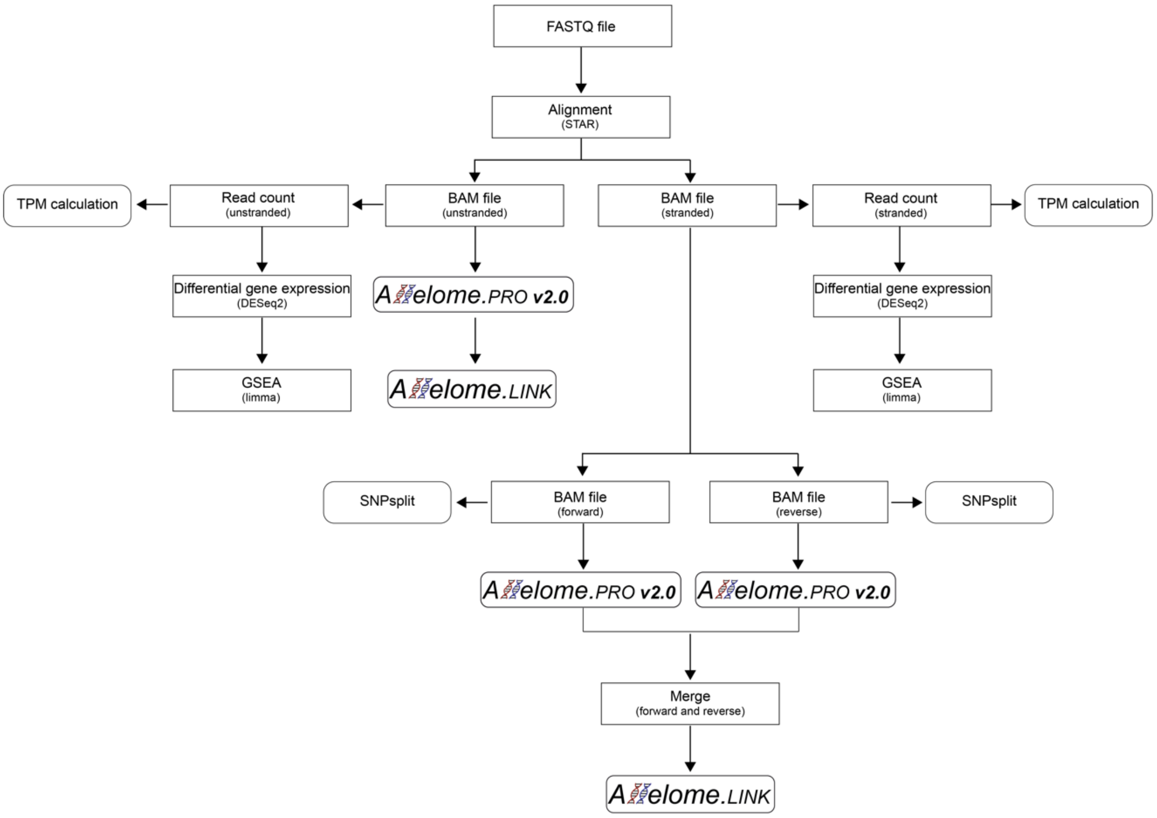


Figure 4.3 Overview of the RNA-seq workflow.
Demultiplexed and adapter-trimmed FASTQ files are obtained from the sequencing facility for each sample. Alignment is performed using STAR[194] for both paired-end and single-end reads. Depending on the sequencing library, reads are quantified as either stranded or unstranded using HTseq[226]. Raw read counts are used to identify differentially expressed genes and to calculate transcripts per million (TPM) for data normalization. The aligned BAM file of unstranded data can be used directly for allele-specific analyses with Allelome.PRO v2.0 and Allelome.LINK. To perform allele-specific analyses in a strand-specific manner, BAM files are split into forward and reverse strands and analyzed individually using Allelome.PRO v2.0. The results per strand are merged and used as input for Allelome.LINK. BAM files can further be used to split the sequencing reads to the allele of origin using SNPsplit[233].

4.5.2.1 Alignment and read quantification

Raw FASTQ files obtained from the sequencing facility or downloaded from the GEO database underwent initial quality control using the FastQC tool (version 0.69). High-quality data was aligned to the GENCODE_M25GRCm38.p6_201911 primary assembly[316] using STAR (version 2.6.0c)[194]. Alignments containing non-canonical junctions, multimappers, or sequencing reads with intron sizes greater than 100,000bp were excluded for downstream analysis. The alignment was performed in paired-end or single-end mode, depending on the sequencing run.

Quantification of aligned reads was conducted with HTseq-count, applying the *–stranded reverse* or *–stranded no* flag based on the strand-specificity of the data (HTSeq version 0.11.3)[226]. Therefore, the gencode.vM25.primary_assembly.annotation.gtf was used as an annotation file[316]. Given that GENCODE vM25 did not annotate *Crossfirre* (*Gm35612*), the locus was manually added from the RefSeq gene annotation[319] to enable the quantification of reads for samples associated with Project 1. Due to the absence of the strand-specific information, the last exon of *Crossfirre* was removed to mitigate bias from the overlapping *Firre* locus.

Raw read counts were normalized by calculating the transcripts per million (TPM) values, accounting for gene length and sequencing depth. Strand-specific BAM files were separated by strand using custom R and Perl scripts. The strand orientation was determined using the infer_experiment.py script, and BigWig files were generated using bam2wig.py[231].

4.5.2.2 Assignment of sequencing reads to the alleles

Sequencing data was further mapped to the allele of origin to facilitate visualization of ASE with a genome browser. Gene coverage was maximized by merging individual FASTQ files for each tissue across replicates. N-masked genomes were generated for FVB/CAST and BL6/CAST SNPs using data from the Sanger database (mgp.v5.merged.snps_all.dbSNP142.vcf)[189] and the SNPsplit_genome_preparation command (SNPsplit v0.3.2)[233]. Read mapping was performed using STAR (version 2.6.0c) with the following parameters: *--outFilterIntronMotifs RemoveNoncanonical, --alignIntronMax 100000, --outFilterMultimapNmax 1, --outSAMattributes NH HI NM MD, --alignEndsType EndToEnd*. The same annotation file was used as described for FVB/CAST and BL6/CAST. Following alignment, BAM files were split according to strand orientation and assigned to the alleles using SNPsplit (version 0.3.2) with default parameters[233].

4.5.2.3 Differential gene expression and gene set enrichment analysis

Differential gene expression analysis was performed in R using the DESeq2 library (version 1.32.0)[251]. Dysregulated genes with an adjusted $\log_2$ fold change ($\log_2$FC) ≥ 1 or ≤ -1 (using lfcShrink with apeglm)[240] and a FDR ≤ 0.01 were considered significant. The R package limma[289] was used to identify enriched gene sets based on the DESeq2 test statistic. Gene sets containing between 10 and 500 genes were obtained from the Molecular Signatures Database (MSigDB, c5.go.v7.4.symbols)[206] and significance was determined using an FDR-adjusted *p*-value ≤ 0.1. For the top 100 gene sets with the lowest FDR-adjusted *p*-values, a similarity matrix was calculated using the R package simplifyEnrichment[304]. Network plots were created using the igraph library[284] and clusters were obtained by the walktrap method[304].

4.5.2.4 Allele-specific expression analysis for bulk RNA-seq data

RNA-seq data from mouse samples were analyzed for ASE using the updated Allelome.PRO v2.0 pipeline. Each sample was processed individually per replicate and strand when feasible. The RefSeq gene annotation (GRCm38/mm10)[319] was used for all analyses. Depending on the F1 hybrid cross, various SNP files were generated using the helper script createSNPbedfile.sh[201] and SNP information from the Sanger database (mgp.v5.merged.snps_all.dbSNP142.vcf)[189]. A cutoff of ≥ 1 read per SNP was applied to all samples, while the total read cutoff per gene was adjusted for each dataset according to sequencing depth and research question.

For placental samples of Project 1 obtained from reciprocal crosses between CAST and BL6, including both WT and various knockout mouse models (Δ*Crossfirre*, Δ*Firre*, Δ*Dxz4*, Δ*Crossfirre-Firre*, Δ*Firre-Dxz4*, and TKO), as well as for the heterozygous TKO spleen samples, a previously described SNP file containing 15,438,314 variants between the CAST and BL6 strains was used[127]. This SNP file includes only BL6 alleles shared with the BALB/cJ, DBA/2J, and 129 mouse strains to reduce potential confounding effects due to strain background. A total read cutoff of ≥ 30 reads per gene was applied to ensure robust allele-specific predictions. The median was used to summarize the allelic ratios between replicates.

Publicly available RNA-seq data of the placenta was further used for Project 2 to investigate the ASE pattern of genes expressed from the X chromosome[78]. Biological replicates were pooled for the forward and reverse cross, and Allelome.PRO v2.0 was run using a SNP file for CAST x FVB including 20,581,027 polymorphisms. To compare the results between the forward and reverse cross, the allelic ratios of the reverse cross were adjusted by subtracting them from 1. A total read coverage of ≥ 20 reads per gene was used for the analysis.

For the BL6 x CAST F1 bodymap generated for Project 2, strand-specific data were analyzed using Allelome.PRO v2.0 with a SNP file including 20,635,313 variants between BL6/CAST and a total read cutoff of ≥ 20 reads per gene. The strand-specific results were combined per sample and the biological replicates per tissue were combined as follows: For each tissue, the lowest allelic score across all replicates was selected to be as robust as possible. Median values were calculated to summarize total reads and allelic ratios. The downstream analysis included only autosomal genes.

For the *Airn* knockout models used in Project 2[78,174], unstranded sequencing data were pooled across replicates and analyzed with Allelome.PRO v2.0 using a total read coverage of ≥ 10 reads and a previously published SNP file containing 16,988,479 variants between the CAST and FVB strains[175]. The allelic ratios of replicates were summarized using the median.

4.5.2.5 Linking ncRNAs to protein-coding targets

Linking ncRNAs to their protein-coding targets was performed using the Allelome.LINK pipeline. The results of the Allelome.PRO v2.0 analyses were used as input for Allelome.LINK (locus_table.txt).

For tissue samples from nine-week-old F1 hybrid mice (BL6 x CAST), including heart, spleen, lung, liver, kidney, and brain, ASE was defined using an allelic ratio cutoff of ≤ 0.3 or ≥ 0.7. The window size for linking allele-specific loci was set to ±100kb, as the significant enrichment of allele-specific ncRNAs nearby allele-specific pcGenes was observed within this distance. To filter for non-coding to protein-coding linkages, the coding information of each gene was obtained from the RefSeq annotation[319].

For the placental *Airn* knockout data, the same allelic ratio cutoff was used with an expanded window size of ±4000kb to encompass the entire *Airn*/*Igf2r* cluster. The *Airn* target prediction was conducted per replicate and for pooled samples to determine if pooling improves the accuracy of the Allelome.LINK results.

Moreover, a chromosome-wide linkage analysis was applied to publicly available RNA-seq data of the placenta to identify the target genes of the lncRNA *Xist*[78]. The allelic ratio cutoff was raised to ≥ 0.75 or ≤ 0.25 to ensure stringency for long-range predictions across a whole chromosome.

4.5.3 Analysis of scRNA-seq data

4.5.3.1 Pre-processing scRNA-seq data

FASTQ files were acquired from the Helmholtz Munich Genomics Core Facility. Raw sequencing data was processed utilizing the Cell Ranger multi-pipeline (Cell Ranger 6.1.2 toolkit, 10x Genomics) to demultiplex, align, and quantify individual sequencing reads[221]. The mm10 genome version 2020-A, a pre-built Cell Ranger reference from 10x Genomics, was used as a reference. Further downstream analysis of the single-cell data was conducted in R using the Seurat package[303].

Sample files were merged, and quality control was applied to exclude low-quality cells. Empty droplets and droplets with multiple cells present were removed by filtering cells with < 500 or > 5000 detected genes and < 2000 or > 20000 molecules. Moreover, cells with > 10% mitochondrial reads were excluded, as elevated mitochondrial read counts indicate damaged cells. Genes present in ≤ 10 cells were further removed for downstream analysis.

Raw counts were normalized using SCTransform[307], incorporating the regression for mitochondrial reads. Data integration was carried out for WT and heterozygous TKO samples using the IntegrateData() function of the Seurat package. Dimensionality reduction was achieved using the top 40 principal components computed by RunPCA() and used for Uniform Manifold Approximation and Projection (UMAP) visualization. Clusters were identified using FindClusters() with a resolution parameter of 0.4. Cell types were manually assigned based on marker gene expression.

4.5.3.2 Allele-specific expression analysis for scRNA-seq data

Allele-specific analysis for scRNA-seq data was performed at two different resolutions. First, Allelome.PRO v2.0 was used to determine the allelic ratio of each chromosome for individual cells to identify whether the Xa was the BL6 or CAST allele. Subsequently, cells were sorted according to the Xi status and aggregated as pseudobulk to increase gene coverage. Thus, this approach allows to overcome the issue of random XCI and allows for an allele-specific analysis at the gene-level for cells with the same Xa.

BAM files were generated for each cell using the sinto package and the filterbarcodes option (https://timoast.github.io/sinto/index.html) to conduct the allele-specific analysis at single-cell resolution. The individual single-cell BAM files were used as input for Allelome.PRO v2.0, along with the previously described SNP file containing 15,438,314 variants[127]. A chromosome-wide annotation was generated using the mm10 genome as a reference. A total read cutoff of ≥ 10 reads per chromosome and ≥ 1 read per SNP was applied. Subsequently, cells with the same Xa status were pooled into four different pseudobulk samples by aggregating the read counts per gene (WT BL6 Xa, TKO BL6 Xa: X chromosomal allelic ratio ≤ 0.3 or WT CAST Xa, TKO CAST Xa: X chromosomal allelic ratio ≥ 0.7). Cells with an X chromosomal allelic ratio between 0.3 and 0.7 were likely duplicates and were excluded from the analysis (WT = 2.99%, TKO = 2.49%). For each pseudobulk sample, individual BAM files were generated with sinto and used as input for Allelome.PRO v2.0. The RefSeq gene annotation was used to obtain gene-level ASE classification[319]. Allelome.PRO v2.0 was run with a read cutoff of ≥ 1 read per SNP and ≥ 30 total reads for each of the four samples (WT BL6 Xa, TKO BL6 Xa, WT CAST Xa, TKO CAST Xa). The entire scRNA-seq workflow is shown in **Figure 4.4**.

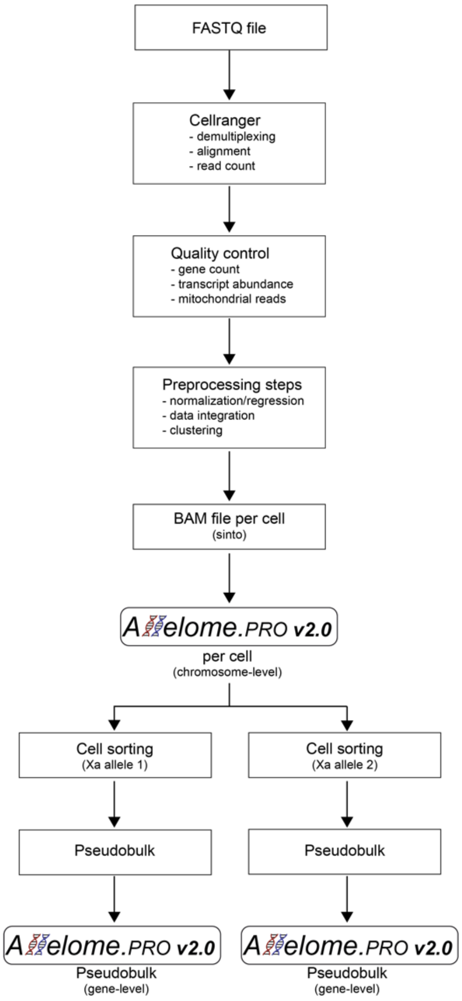


Figure 4.4 Overview of the scRNA-seq workflow for XCI-based cell sorting and gene-level ASE analysis.

FASTQ files from the sequencing facility are pre-processed using the Cell Ranger multi-pipeline (Cell Ranger 6.1.2 toolkit, 10x Genomics[221]) to demultiplex, align and quantify individual sequencing reads. Quality control is performed using R and the Seurat package[303], taking into account gene counts, transcript abundance, and the proportion of mitochondrial reads. This is followed by standard pre-processing steps, including normalization, regression for mitochondrial reads, data integration, clustering, and cell-type annotation. To determine the allelic ratio of each chromosome for individual cells, the BAM file is split into individual BAM files per cell and Allelome.PRO v2.0 is run with a chromosome-wide annotation. Cells are then computationally sorted according to the XCI status and aggregated as pseudobulk to increase gene coverage for a subsequent gene-level allele-specific expression (ASE) analysis.

4.5.4 Analysis of human data from the GTEx project

4.5.4.1 Data pre-processing and allele-specific linking

The Allelome.LINK strategy was further applied to human data from the GTEx project[190]. The publicly available haplotype dataset from the GTEx v8 release provided 153 million allele-specific measurements from 838 individuals, comprising 15,253 samples and 54 tissues (phASER_WASP_GTEx_v8_matrix.gw_phased.txt)[191]. Metadata, including the sample information, was retrieved from the GTEx portal (GTEx_Analysis_v8_Annotations_SampleAttributesDS.txt) and used to separate the expression data by sample using custom R scripts.

Small ncRNAs (combined exon length ≤ 200bp) and overlapping genes were removed from the dataset to minimize allelic bias due to overlapping gene loci. This filtering resulted in 8,106 informative genes (total read cutoff ≥ 20), including 6,281 pcGenes (*n* = 27,155,698 expression values) and 1,825 ncRNA loci (*n* = 924,440 expression values). Subsequently, Allelome.LINK was used to assign ncRNAs to their target genes applying an ASE cutoff of ≤ 0.3 or ≥ 0.7 and a window size of ±100kb. The results were filtered for non-coding to protein-coding linkages. All analysis steps were based on the GENCODEv26 annotation[316] to ensure consistency with the gene annotation of the GTEx project.

4.5.4.2 Validation of ncRNA-to-target linkages using eQTL data

The predicted ncRNA-to-target linkages were validated using sample-matched eQTL data from the GTEx v8 release (GTEx_v8_finemapping_DAPG.txt)[318]. This dataset provided 21,648,584 fine-mapped eQTLs across 49 tissues. eQTL data was unavailable for the Bladder, Cervix - Ectocervix, Cervix - Endocervix, Fallopian Tube, and Kidney - Medulla. Genomic locations and target genes were available for 21,412,255 eQTLs. The ncRNA-to-target linkages were overlapped with the eQTL data by genomic position. A linkage was confirmed if an overlapping eQTL was predicted to influence the expression of the same target gene as predicted for the ncRNA by Allelome.LINK.

4.5.4.3 Integrating GWAS data to assign non-coding risk variants to protein-coding targets

To assign non-coding risk variants to their protein-coding targets, public GWAS data was downloaded from the NHGRI-EBI GWAS Catalog v1.0, including 132,201 unique SNPs (gwas_catalog_v1.0-associations_e110_r2023-10-11.tsv)[215]. The dataset was filtered to exclude non-mappable SNPs and epistatic interactions, resulting in 119,287 variants for downstream analysis. Subsequently, the variant positions were overlapped with the informative ncRNAs, which yielded 1,059 informative SNPs. Non-coding risk variants that overlapped with linked ncRNAs were assigned to the same protein-coding target as predicted by Allelome.LINK.

4.6 Statistical analysis

Data analysis was performed with R version 3.6.3. Depending on the research question, the appropriate statistical test was performed as stated in the figure legends. The Shapiro-Wilk test was used to assess the normality of the data distribution. Based on the data distribution, Pearson or Spearman correlation coefficients were calculated to evaluate significant correlations. Differences between two groups were assessed using either the Wilcoxon rank-sum test for non-normally distributed data or the t-test for normally distributed data. Fisher's exact tests were used to compare categorical proportions between two variables. Binomial tests were applied to determine whether the presence of an expected proportion is consistent with equal probability. Results were considered significant with varying *p*-value thresholds, as indicated for the respective tests based on stringency. If needed, FDR correction was applied to adjust for multiple testing.

Genotype effects were further assessed by the GMC depending on the research question and the assumed parameter distribution. These included Wilcoxon rank-sum tests, t-tests, ANOVA and post-hoc tests, linear models, or Fisher's exact tests. A *p*-value < 0.05 was considered statistically significant for the observed phenotypes with no adjustment for multiple testing.

5 Results

The results described in the sections of chapter **5.1** (**5.1.1**, **5.1.2**, **5.1.3**, **5.1.4**, **5.1.5**, **5.1.6**, **5.1.7**) have been previously published in a similar form by the author of this thesis[322]. Additionally, the results presented in chapter **5.2** (**5.2.1, 5.2.2, 5.2.3, 5.2.4, 5.2.5, 5.2.6, 5.2.7**) have been outlined similarly by the author in a submitted manuscript (see **9.2 Submitted manuscripts, 1.**).

5.1 Project 1: Investigating the *in vivo* contribution of the *Crossfirre* locus alone and in combination with *Firre* and *Dxz4*

The lncRNA *Crossfirre* was recently detected as an imprinted gene on the X chromosome, showing predominantly maternal expression in somatic tissues[78]. Prior to this work, the *in vivo* contribution of this locus, both alone and in combination with *Firre* and *Dxz4*, remained unknown. Thus, the first project aimed to investigate the *in vivo* role of *Crossfirre*, *Firre,* and *Dxz4,* including their relationship to imprinted and random XCI, as well as autosomal gene regulation.

5.1.1 *Crossfirre*, *Firre*, and *Dxz4* are the most female-specific loci in chromatin accessibility

The initial aim of the study was to characterize the epigenetic profile and the expression pattern of the *Crossfirre* locus. The *Crossfirre* locus comprises three exons and is embedded in a 50kb LINE element. The lncRNA is transcribed from the forward strand, antisense to the lncRNA *Firre*, and was previously identified as the only maternally expressed X-linked gene in somatic tissues (**Figure 5.1a**)[78]. To confirm the imprinted status of *Crossfirre*, publicly available brain data from F1 hybrids of reciprocal crosses (FVB x CAST, CAST x FVB) were reanalyzed for the allelic expression status of *Crossfirre*[78]. In line with previous work, predominant expression was observed from the maternal allele independent of the XCI status, confirming imprinting of the *Crossfirre* locus (**Figure 5.1b**). Further validation of the imprinting of *Crossfirre* was achieved by reanalyzing H3K4me3 ChIP-sequencing data from F1 mouse embryonic fibroblasts[201], which confirmed maternal H3K4me3 enrichment at the *Crossfirre* promoter (**Figure 5.1c**).

The expression rates of *Crossfirre*, *Firre*, and *Dxz4* were further examined in the six major organs (brain, heart, lung, liver, kidney, spleen) of adult mice, revealing low to moderate expression. The highest expression of *Crossfirre* was detected in the brain, while the lowest expression was observed in the liver (**Figure 5.1d**). A correlation analysis of the TPM levels between *Crossfirre*, *Firre*, and *Dxz4* was performed to explore potential co-regulation. However, no significant correlation was noted between the lncRNAs (*Crossfirre*-*Firre*: R = 0.36, *p*-value = 0.4893; *Crossfirre*-*Dxz4*: R = 0.33, *p*-value = 0.5273; *Firre*-Dxz4: R = 0.57, *p*-value = 0.2419, **Figure 5.1e**).

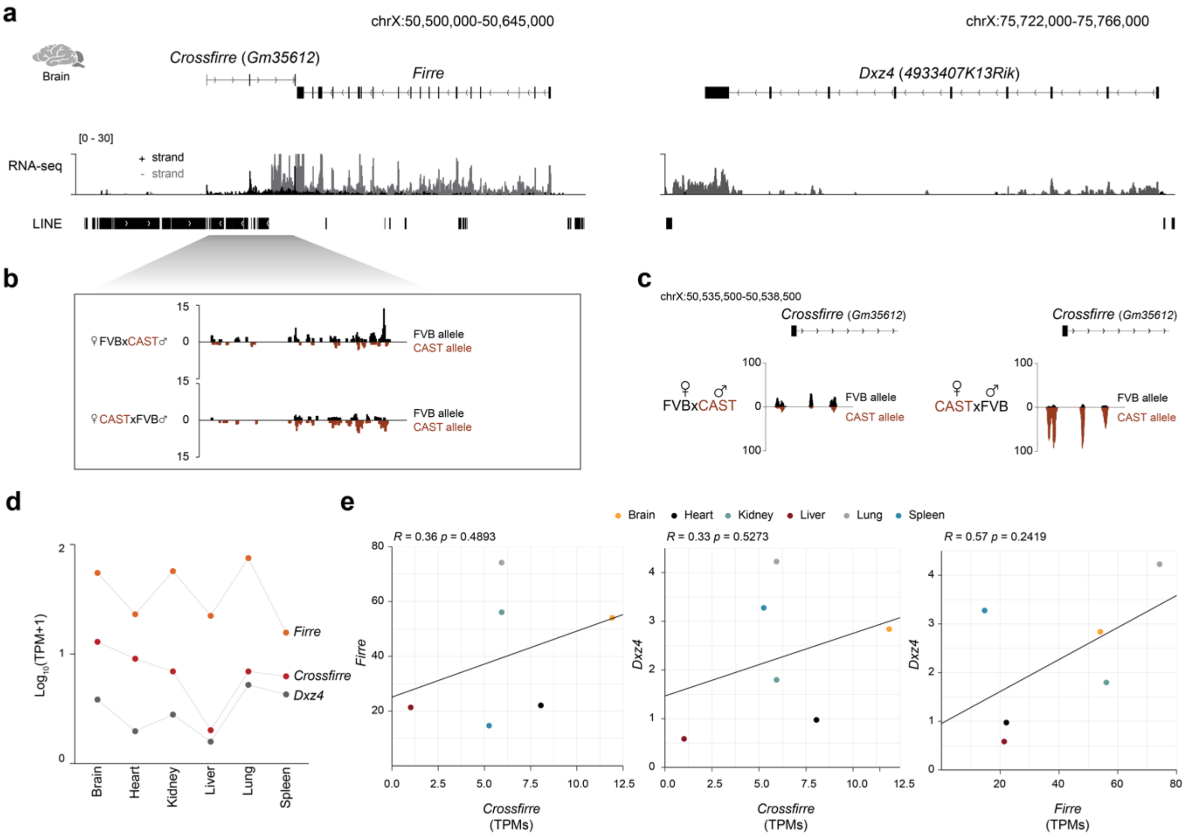


Figure 5.1 Expression dynamics of *Crossfirre*, *Firre*, and *Dxz4* across mouse organs.
a, Genome browser visualization of mouse brain RNA-seq data for the *Crossfirre* (*Gm35612*), *Firre,* and *Dxz4* loci. Sequencing reads are separated by strand orientation. The data was obtained from Andergassen *et al.*, 2017[78].
b, Allele-specific splitting of RNA-seq data from adult F1 hybrid brains. The genome browser tracks display sequencing reads from the forward strand at the *Crossfirre* locus of reciprocal F1 hybrids (FVB x CAST, CAST x FVB). Reads are labeled according to their allele of origin (FVB allele: black, CAST allele: brown).
c, H3K4me3 data of F1 mouse embryonic fibroblasts from Andergassen *et al.*, 2015[201]. Shown is the allele-specific splitting for the aligned sequencing reads towards the FVB and CAST allele.
d, Expression of the *Crossfirre*, *Firre,* and *Dxz4* loci across various adult mouse tissues. To avoid the decadic logarithm of zero, the mean transcripts per million (TPM) were adjusted with a pseudo number of 1 prior to $\log_{10}$-transformation. Expression levels for *Crossfirre* (red), *Firre* (orange), and *Dxz4* (gray) are shown for the brain, heart, kidney, liver, lung, and spleen.
e, TPM correlation analysis for *Crossfirre*, *Firre*, and *Dxz4*. Pearson correlation coefficients were calculated for the mean TPM values between *Crossfirre* and *Firre*, *Crossfirre* and *Dxz4*, and *Firre* and *Dxz4* across multiple organs. The figure was modified from Hasenbein *et al.*, 2024[322].

Previous studies have shown that the *Firre* and *Dxz4* loci contain several Xi-specific transcription start sites with CTCF binding[78,121,124,128]. Thus, the epigenetic profile of these loci was further investigated using publicly available ATAC-seq data[321]. Peak calling confirmed the presence of multiple sites of open chromatin at these loci in the brain (**Figure 5.2a**). Notably, female mice exhibited a more pronounced chromatin accessibility profile at the *Crossfirre*-*Firre* and *Dxz4* loci compared to males (ATAC peaks: *Crossfirre-Firre n* = 21, *Dxz4 n* = 11; male ATAC peaks: *Crossfirre-Firre n* = 9, *Dxz4 n* = 2, **Figure 5.2a**). To further investigate this pattern, the epigenetic ATAC-seq profile of all loci was analyzed across multiple female and male organs. Interestingly, this analysis identified that *Crossfirre*, *Firre*, and *Dxz4* are the most female-specific chromatin accessibility loci genome-wide (**Figure 5.2b-c**). Additionally, an allele-specific analysis of ATAC-seq data from neural progenitor cells[320] confirmed that the female-specific open chromatin originates from the Xi (**Figure 5.2d**).

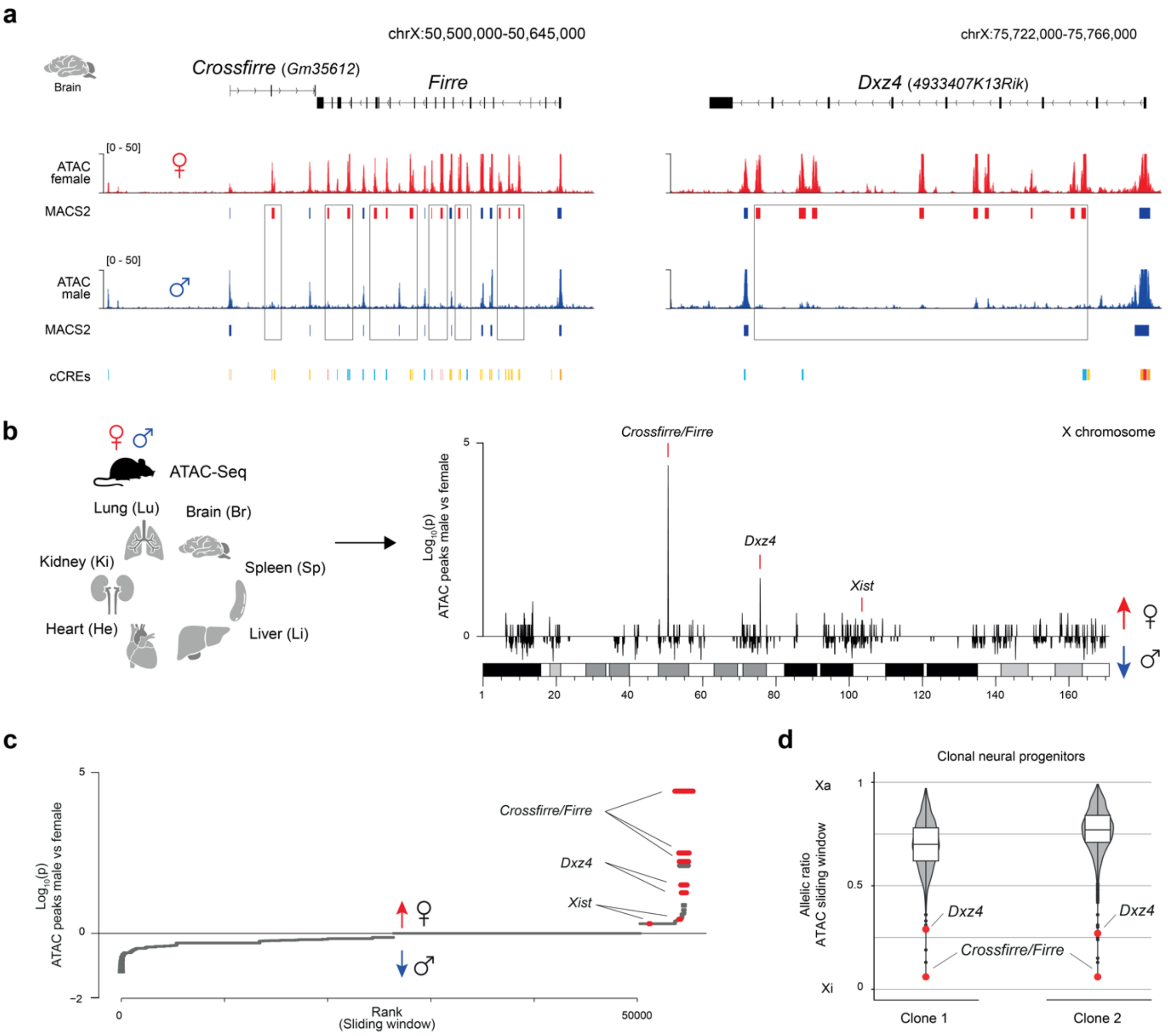


Figure 5.2 ***Crossfirre*****,** ***Firre,*** **and** ***Dxz4*** **are the top female-specific loci.**

a, Genome browser visualization of ATAC-seq data[321] of the *Crossfirre-Firre* and *Dxz4* loci. ATAC-seq data from female (red) and male (blue) brain samples are displayed. Peaks were called using macs2[227] and compared between the sexes. Female-specific ATAC peaks are highlighted with boxes. cCRES: ENCODE Candidate *Cis*-Regulator Elements, CTCF: blue, Promoter: red, DNase/H3K4me3: pink, Proximal enhancer: orange, Distal enhancer: yellow.

b, Epigenetic ATAC-seq profile across six adult mouse tissues[321]. Peaks were called per tissue and sex and quantified within 100kb windows across the X chromosome. A binomial test was used to calculate $\log_{10}$ *p*-values, using the median number of peaks per window across all tissues. Peaks more frequent in females were assigned positive values, while those more frequent in males were assigned negative values.

c, Analysis as in (**b**), but across the entire genome.

d, Allele-specific analysis for neural progenitor cells[320]. The allelic ratios are shown for 50kb sliding windows across the X chromosome. Boxplots represent the interquartile range with the median, and whiskers indicate 1.5x the interquartile range. The figure was modified from Hasenbein *et al.*, 2024[322].

5.1.2 The deletion of *Crossfirre* does not affect viability and development

To further investigate the *in vivo* role of *Crossfirre*, individually and in combination with *Firre* and *Dxz4*, several knockout mouse models were generated, including a *Crossfirre* deletion (Δ*Crossfirre*), a *Crossfirre-Firre* double deletion (Δ*Crossfirre-Firre*), and a *Crossfirre-Firre-Dxz4* triple deletion (TKO). The samples were analyzed with the previously published single-deletions of Δ*Firre* and Δ*Dxz4* and the Δ*Firre-Dxz4* double-deletion mouse models (**Figure 5.3**)[127,131].

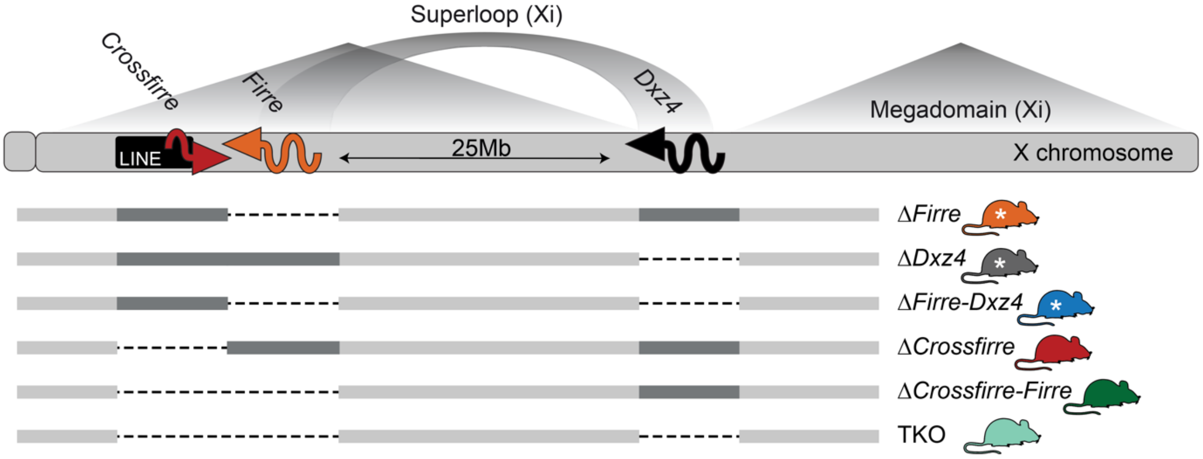
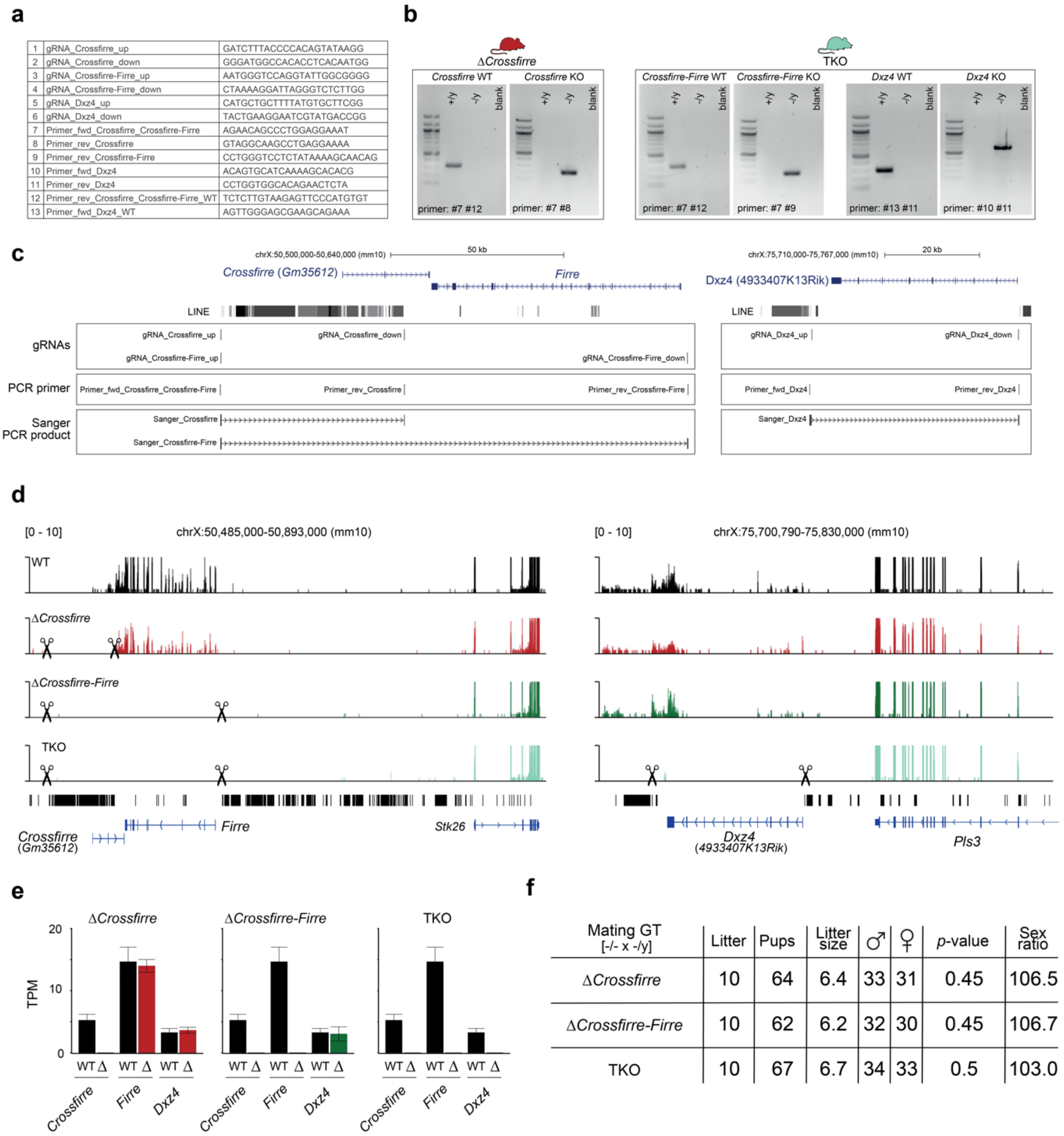


Figure 5.3 Overview of mutant mouse models.
Schematic overview of the mutant mouse models used in the study. The X chromosome is shown, with the *Crossfirre*, *Firre*, and *Dxz4* loci highlighted. Transcription of these loci is specific to the active X chromosome. In contrast, the superloop interaction between *Firre* and *Dxz4*, and the two megadomains are observed exclusively on the inactive X chromosome (Xi). Below the X chromosome, the different mutant mouse models are shown, with deletions indicated by dotted lines. Stars denote mouse models described in previous studies that were reanalyzed for the study[127,131]. The figure was obtained from Hasenbein *et al.*, 2024[322].

The presence of each knockout was confirmed by genotyping and subsequent Sanger sequencing of the PCR product (**Figure 5.4a-c**). RNA-seq of spleens from homozygous mutant mice was further performed to validate the loss of expression (**Figure 5.4d**). Notably, the loss of *Crossfirre* did not affect *Firre* expression, and the combined deletion of *Crossfirre* and *Firre* did not alter the expression levels of *Dxz4* (**Figure 5.4e**). Furthermore, no developmental abnormalities were observed in any of the three mutant mouse models, as homozygous knockout strains were viable and fertile, and offspring displayed average sex ratios and expected litter sizes (**Figure 5.4f**). Taken together, these results suggest that the *Crossfirre* locus, individually and combined with *Firre* and *Dxz4*, is dispensable for development.



1	gRNA_Crossfirre_up	GATCTTTACCCCACAGTATAAGG
2	gRNA_Crossfirre_down	GGGATGGCCACACCTCACAATGG
3	gRNA_Crossfirre-Firre_up	AATGGGTCCAGGTATTGGCGGGG
4	gRNA_Crossfirre-Firre_down	CTAAAAGGATTAGGGTCTCTTGG
5	gRNA_Dxz4_up	CATGCTGCTTTTATGTGCTTCGG
6	gRNA_Dxz4_down	TACTGAAGGAATCGTATGACCGG
7	Primer_fwd_Crossfirre_Crossfirre-Firre	AGAACAGCCCTGGAGGAAAT
8	Primer_rev_Crossfirre	GTAGGCAAGCCTGAGGAAAA
9	Primer_rev_Crossfirre-Firre	CCTGGGTCCTCTATAAAAGCAACAG
10	Primer_fwd_Dxz4	ACAGTGCATCAAAAGCACACG
11	Primer_rev_Dxz4	CCTGGTGGCACAGAACTCTA
12	Primer_rev_Crossfirre_Crossfirre-Firre_WT	TCTCTTGTAAGAGTTCCCATGTGT
13	Primer_fwd_Dxz4_WT	AGTTGGGAGCGAAGCAGAAA

Mating GT [-/- x -/y]	Litter	Pups	Litter size	♂	♀	*p*-value	Sex ratio
Δ*Crossfirre*	10	64	6.4	33	31	0.45	106.5
Δ*Crossfirre-Firre*	10	62	6.2	32	30	0.45	106.7
TKO	10	67	6.7	34	33	0.5	103.0

Figure 5.4 *Crossfirre*, *Firre,* and *Dxz4* mutant strains are viable and develop normally.

a, Sequences of guide RNAs and primers used to generate and genotype Δ*Crossfirre*, Δ*Crossfirre-Firre*, and Δ*Dxz4* mouse models.

b,Genotyping approach for identifying knockout and wildtype (WT) alleles for Δ*Crossfirre*, Δ*Crossfirre-Firre*, and Δ*Crossfirre-Firre-Dxz4* (TKO).

c, Genome browser visualization of Sanger sequencing of the PCR products from knockout bands of the *Crossfirre-Firre* and *Dxz4* loci.

d, Genome browser visualization of RNA-seq tracks covering the *Crossfirre-Firre* and *Dxz4* loci. RNA-seq data was obtained from adult spleens of WT (black), Δ*Crossfirre* (red), Δ*Crossfirre-Firre* (green), and Δ*Crossfirre-Firre-Dxz4* (TKO, turquoise) mouse models. Scissor symbols highlight the corresponding deletion. Long interspersed nuclear elements and a gene annotation are shown below the sequencing tracks.

e, Transcript per million (TPM) values for *Crossfirre*, *Firre,* and *Dxz4*. Shown are the mean TPMs for all three lncRNAs in the adult spleen of Δ*Crossfirre*, Δ*Crossfirre-Firre*, and TKO. Error bars display the standard deviation.

f, Overview of the sex distribution for the breeding of homozygous Δ*Crossfirre*, Δ*Crossfirre-Firre*, and TKO mouse models. A one-sided binomial test was used to calculate *p*-values. The figure was modified from Hasenbein *et al.*, 2024[322].

5.1.3 The deletion of *Crossfirre*, *Firre*, and *Dxz4* does not alter imprinted XCI

Given the imprinting of *Crossfirre*, it was hypothesized that this locus could serve as a marker for imprinted XCI and is therefore involved in the XCI process. To test this hypothesis, the study investigated whether the knockout of *Crossfirre*, individually or combined with *Firre* and *Dxz4*, affects imprinted XCI in the placenta. Because the paternal X chromosome is epigenetically silenced in imprinted XCI, deletions inherited from the paternal allele affect the Xi, whereas deletions inherited from the maternal allele are present on Xa. Placentas were isolated from E12.5 WT and heterozygous offspring of reciprocal crosses (CAST x -/+ BL6, -/+ BL6 x CAST, n = 53). This approach allowed for the evaluation of both the effects of the loss of the lncRNA expression (deletion on Xa) and the disruption of the epigenetic patterns and megastructures (deletion on Xi).

The relative expression of *Crossfirre*, *Firre*, and *Dxz4* was examined in TKO models. Mutants with the deletion on Xi showed expression levels similar to the WT conditions. In contrast, mutants with the deletion on Xa lost the expression of *Crossfirre*, *Firre*, and *Dxz4*, confirming the Xa-specific expression of these lncRNAs (**Figure 5.5a**). Differential gene expression analysis was performed for all generated mutant strains and the previously published Δ*Firre* and Δ*Dxz4* single and double deletion models[127]. Interestingly, this analysis revealed very few dysregulated genes, regardless of whether the deletion was placed on Xa or Xi (|shrunk $\log_2$FC| ≥ 1, FDR ≤ 0.01, **Figure 5.5b**, mean ΔXi = 10, mean ΔXa = 37). When the deletion was present on Xa, *Firre* was the only gene shared between the double knockouts Δ*Crossfirre-Firre* and Δ*Firre-Dxz4*, and the TKO. For the Δ*Crossfirre-Firre* and TKO Xa mutants, two additional pseudogenes (*Gm13340*, *Gm13436*) were shared, while one pseudogene (*Rpsa-ps10*) was shared when the deletion was present on Xi (**Figure 5.5c**).

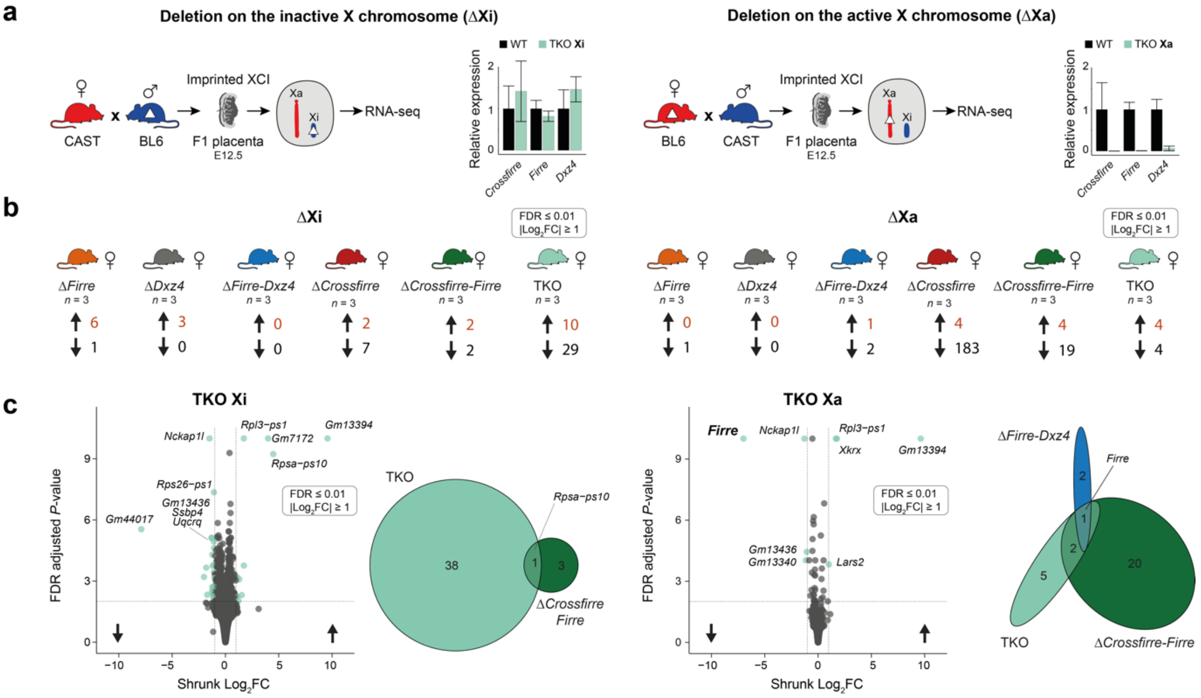


Figure 5.5 Effect of the *Crossfirre*, *Firre,* and *Dxz4* deletion on placental gene expression.
a, Overview of the breeding scheme to place the deletion on the inactive (Xi) or active (Xa) X chromosome. Due to imprinted X chromosome inactivation in the placenta, the paternal X chromosome is epigenetically silenced. Thus, deletions inherited from the paternal allele affect Xi, while deletions from the maternal allele affect Xa. The relative mean expression of *Crossfirre*, *Firre,* and *Dxz4* is shown for triple knockout (TKO) mutants and wildtype (WT). Error bars indicate the standard deviation.
b, Results of a differential gene expression analysis for mutant strains with the deletions on Xi or Xa. E12.5 placentas from three biological replicates per genotype were analyzed. Numbers indicate the up- and downregulated genes with a significance threshold of FDR ≤ 0.01 and a |shrunk $\log_2$FC| ≥ 1.
c, Volcano plot of the differential gene expression analysis for the TKO model with the deletions on Xi and Xa. The Venn diagram shows the dysregulated genes shared between the Δ*Firre-Dxz4* (blue), Δ*Crossfirre-Firre* (green), and TKO (turquoise) genotypes. The figure was obtained from Hasenbein *et al.*, 2024[322].

To determine whether the absence of these loci affects the expected maternal expression of X-linked genes, an allele-specific analysis was performed using RNA-seq data from E12.5 placentas. Notably, more CAST-specific escape genes were detected in WT BL6 x CAST F1 hybrid samples (**Figure 5.6a**). This observation is a well-known phenomenon that has previously been reported as strain-specific escape[78,325,326]. To illustrate the reproducibility of this pattern, the allelic ratios were plotted for all WT samples from both forward and reverse crosses (n = 17), showing a consistently higher frequency of CAST-specific escapees[127] (**Figure 5.6b**).

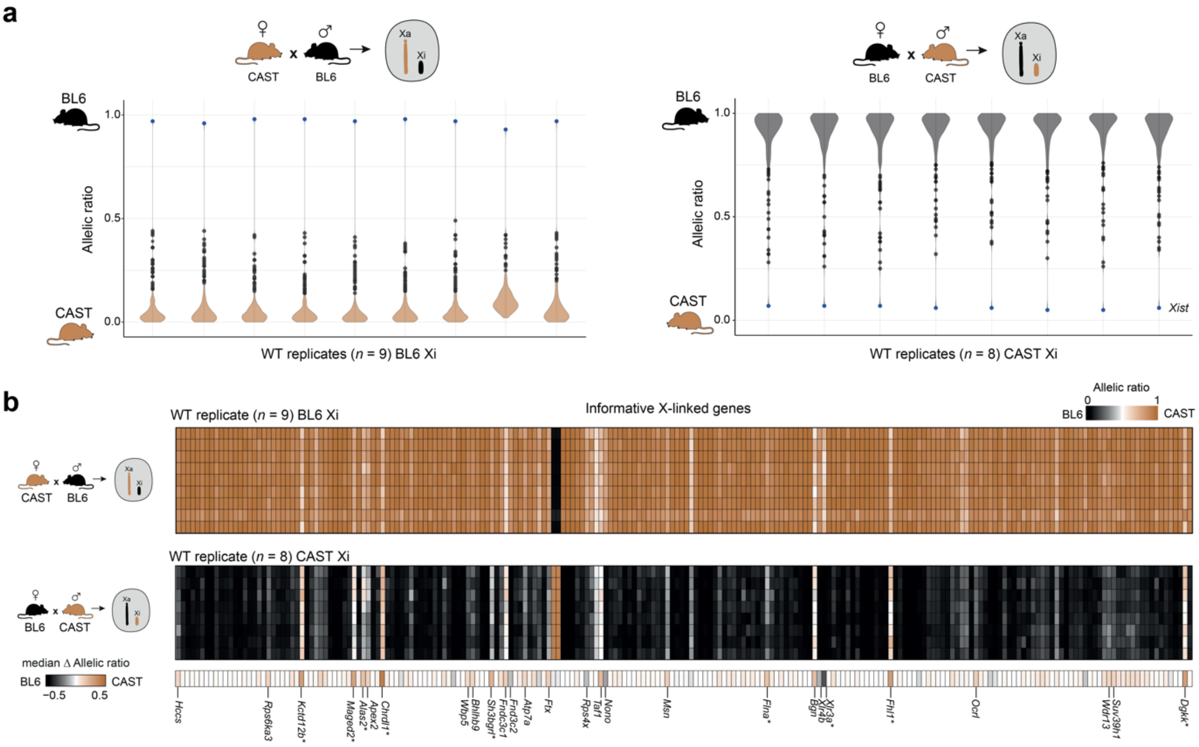


Figure 5.6 Strain-specific escape results in more escape genes from CAST Xi.
a, Violin plots displaying the median allelic ratios of informative X-linked genes for wildtype samples from forward (BL6 inactive X chromosome (Xi), left, *n* = 9) and reverse crosses (CAST Xi, right, *n* = 8). Colors indicate the strain background of the active X chromosome (Xa, CAST: brown, BL6: black), with blue dots marking the allelic ratios of *Xist*.
b, Heatmap representing the median allelic ratios of X-linked genes for wildtype samples of the forward (BL6 Xi, upper panel, *n* = 9) and reverse crosses (CAST Xi, lower panel, *n* = 8). Colors represent the strain background of the Xa (CAST: brown, BL6: black). Genes with an allelic ratio change between BL6 Xi and CAST Xi of ≥ 0.1 are labeled. An asterisk indicates previously identified strain-specific escape genes[127]. The figure was obtained from Hasenbein *et al.*, 2024[322].

The WT allelic ratios were further compared to all mutant strains and showed no changes in the median allelic ratios of X-linked genes, independent of whether the deletions were on Xi or Xa (**Figure 5.7a-b**). The *Crossfirre* deletion further encompassed the removal of a 50kb LINE cluster, DNA elements hypothesized to facilitate the spreading of XCI to escaping-prone regions[141,327]. Therefore, local regions in proximity to the *Crossfirre*, *Firre*, and *Dxz4* loci were further examined for subtle changes in *cis*. However, the allelic ratios of nearby genes were not affected by the absence of these loci (**Figure 5.7b-c**). In conclusion, the absence of the imprinted *Crossfirre* locus on Xi or Xa, individually or combined with *Firre* and *Dxz4*, does not impair imprinted XCI in the placenta.

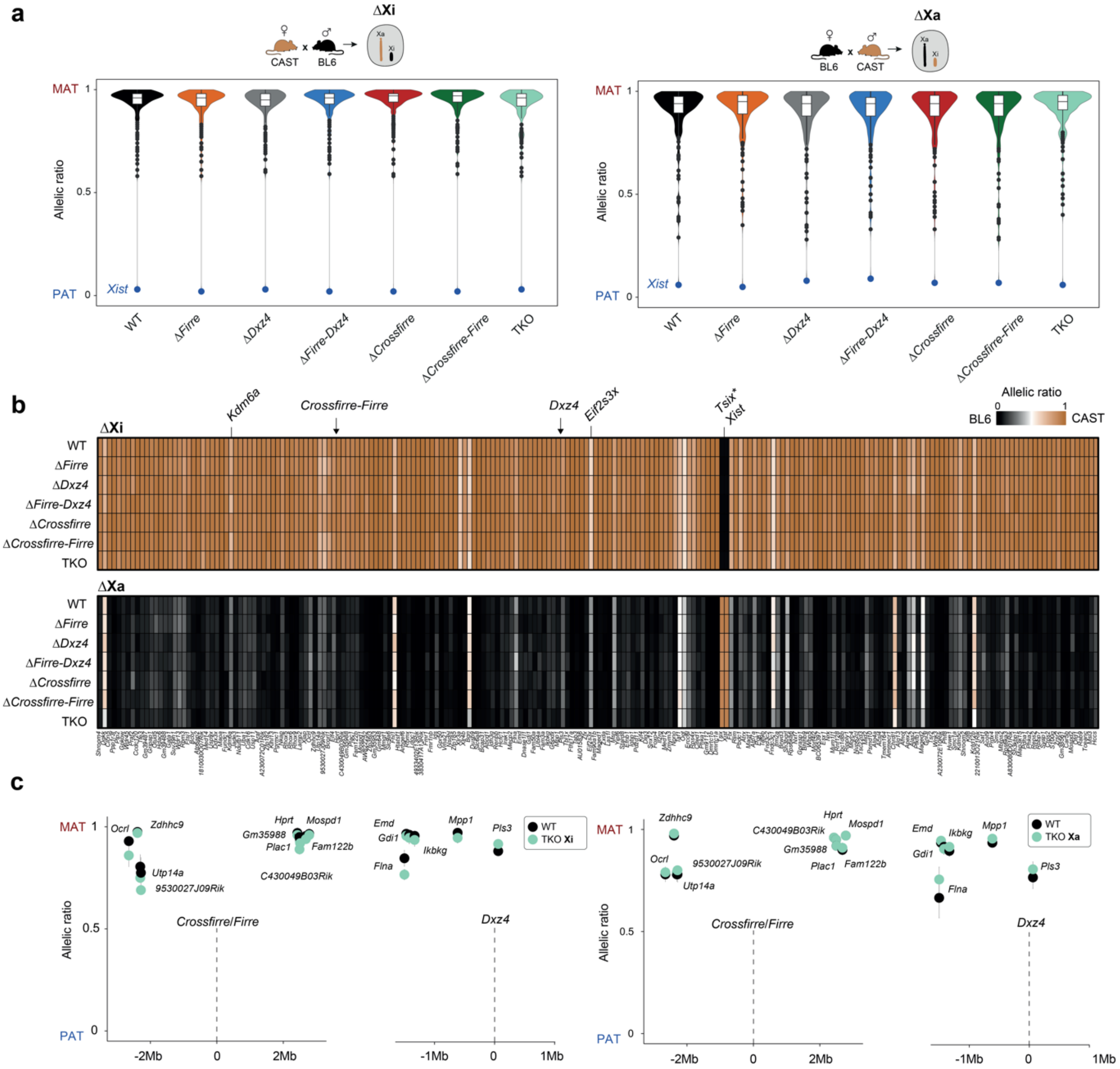


Figure 5.7 Imprinted XCI is not affected by deleting *Crossfirre* alone or in combination with *Firre* and *Dxz4*.

a, Violin plots displaying median allelic ratios of X-linked genes for wildtype (WT) and knockout models with the deletion on Xi (left, *n* = 7) or Xa (right, *n* = 7). Colors denote the different mutant genotypes, and blue dots indicate the allelic ratios of *Xist*. Boxes represent the interquartile range, with whiskers extending to 1.5x the interquartile range.

b, Heatmap illustrating the median allelic ratios of informative X-linked genes for WT samples and mutant strains from forward (BL6 Xi, upper panel) and reverse (CAST Xi, lower panel) crosses. Colors indicate the strain background of Xa (CAST: brown, BL6: black). Arrows emphasize the approximate position of *Crossfirre*, *Firre,* and *Dxz4*. *Note: Artifact of the non-strand-specific analysis. *Tsix* expression is biased by *Xist*, which is transcribed in the antisense direction.

c, Overview of the median allelic ratios and standard deviations for genes in close proximity to *Crossfirre/Firre* (±2 Mb) and *Dxz4* (±1 Mb). Results for Δ*Crossfirre-Firre-Dxz4* (TKO) mutants, with the deletions on Xi (left) or Xa (right) are compared to strain background matched WT samples. The figure was modified from Hasenbein *et al.*, 2024[322].

5.1.4 Random XCI maintenance remains unaffected in TKO mutants

Random XCI results in each X chromosome being either active or inactive in a mosaic pattern across cells in adult organs, complicating the study of how the knockouts affect random XCI. To address this challenge, allele-specific analyses have been performed on single cells, which allow the random nature of XCI to be overcome (**Figure 5.8**).

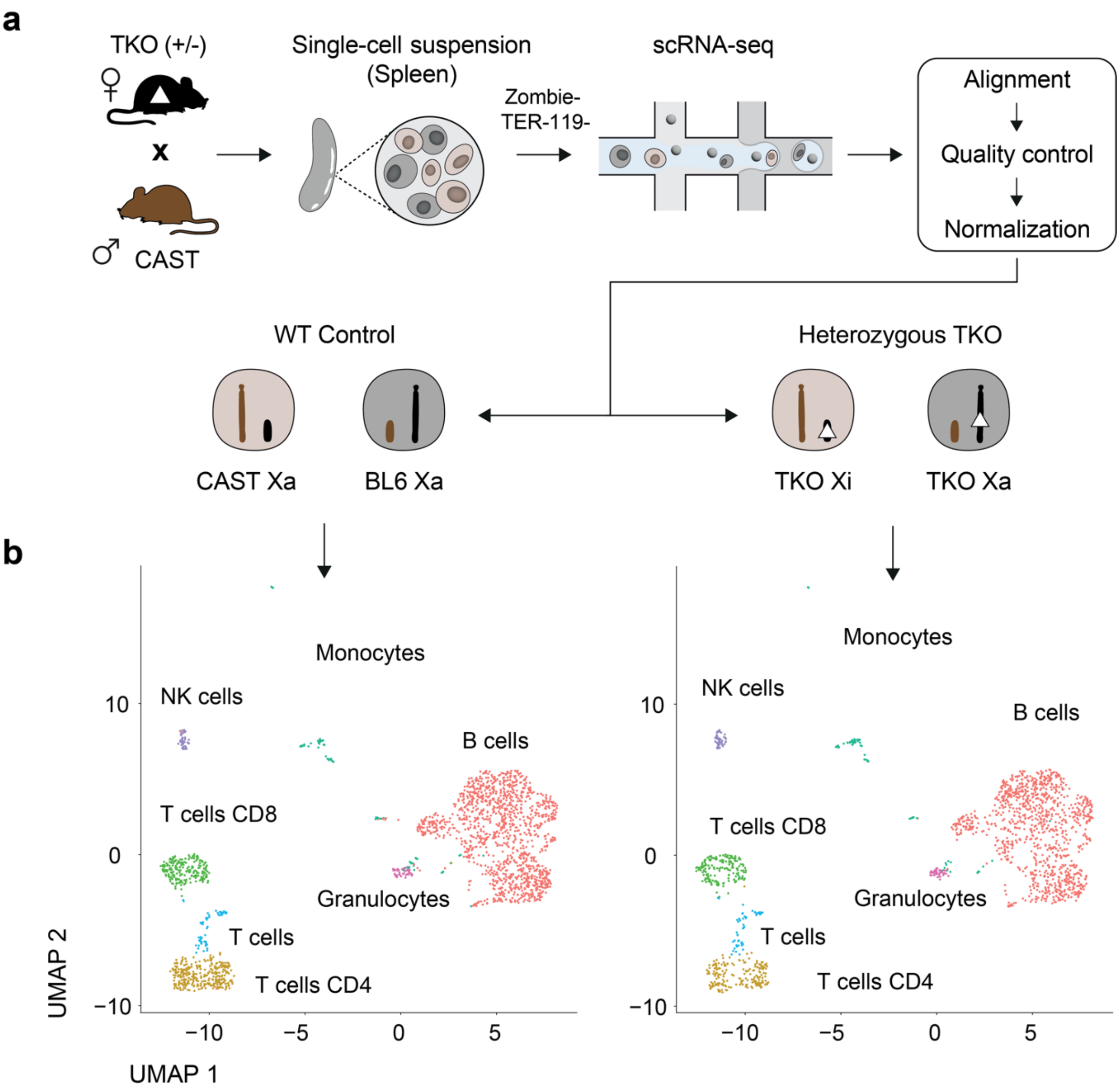


Figure 5.8 Overview of the scRNA-seq workflow of adult F1 spleens.
a, Overview of the scRNA-seq workflow to obtain wildtype (WT) and heterozygous Δ*Crossfirre-Firre-Dxz4* (TKO) cells with the deletion on either Xi or Xa. CAST males were crossed with -/+ TKO females (BL6), and scRNA-seq was performed on the spleens of a WT and a heterozygous F1 hybrid. Preprocessing steps of the sequencing data included alignment, quality control, and data normalization. Based on the genotype (WT, -/+TKO), two datasets were obtained with a mixed cell population containing cells with the CAST and BL6 allele Xa. Consequently, the TKO deletion was present on Xi or Xa.

b, UMAP visualization of the unsupervised clustering of the cells in (**a**), including WT and -/+ TKO samples. The figure was obtained from Hasenbein *et al.*, 2024[322].

Heterozygous TKO females (-/+ BL6) were crossed with WT males (+/+ CAST) to obtain female F1 littermates of a -/+ TKO and a WT sample. The spleen from these two animals were harvested at six weeks of age and used for scRNA-seq, resulting in the allele-specific single-cell transcriptome of 1642 heterozygous and 2043 WT cells after quality control (**Figure 5.8a**). Clustering revealed the expected cell types and cell-type proportions for both samples (**Figure 5.8b**).

Subsequently, an allele-specific analysis was performed at the chromosome-level by aggregating reads from the same chromosome for individual cells. Biallelic expression was observed for autosomes of most WT and heterozygous TKO cells (**Figure 5.9a**). For the X chromosome, a subset of cells showed biallelic expression (WT n = 61, TKO n = 40). These cells are likely to be duplicates and were excluded from further analysis. The majority of WT cells showed an X chromosomal allelic ratio ≥ 0.7 (n = 1342, 65.7%), indicating that the CAST allele was Xa. In contrast, 640 (31.3%) WT cells had an allelic ratio ≤ 0.3, suggesting that the BL6 allele was Xa (**Figure 5.9b**). Consequently, WT cells exhibited the expected XCI skewing ratio in F1 hybrids between CAST and BL6 (70% CAST Xa, 30% BL6 Xa[328]). Interestingly, cells from the heterozygous TKO population showed a more pronounced skewing ratio, with 82.8% CAST Xa (TKO on Xi, n = 1359) and 14.8% BL6 Xa (TKO on Xa, n = 243, **Figure 5.9b**). To validate this shift in the XCI skewing pattern of mutant cells, bulk RNA-seq was performed for -/+ TKO (n = 3) and WT samples (n = 3, **Figure 5.9c**). However, the allelic ratios of X-linked genes showed a similar skewing pattern for mutant and WT samples. Moreover, no significant differences were detected in the allelic ratios of the lncRNA *Xist* (**Figure 5.9d-e**).

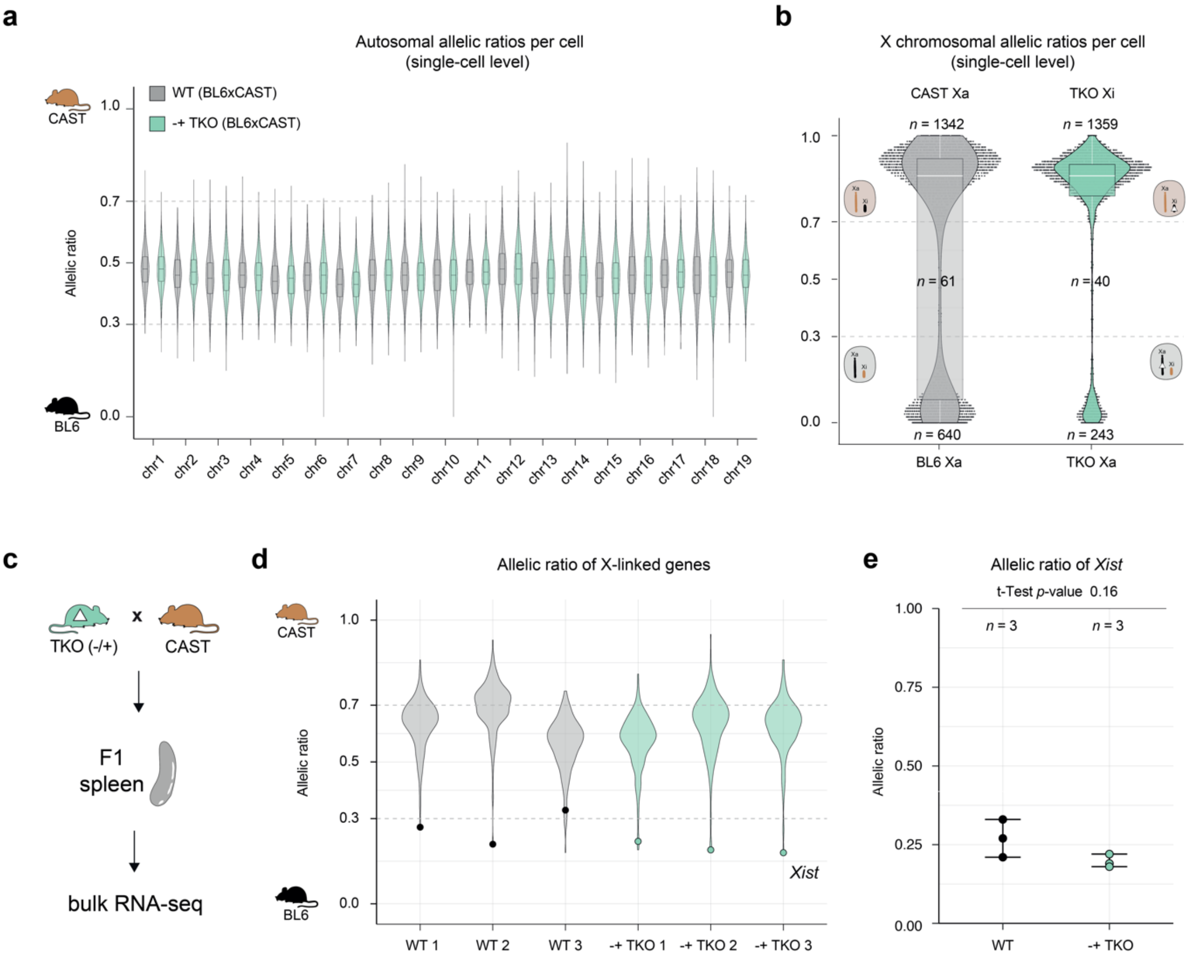


Figure 5.9 Allele-specific analysis of spleens using scRNA-seq and bulk RNA-seq data.
a, Violin plot for autosomes showing the allelic ratios of wildtype (WT, gray) and -/+ ∆*Crossfirre-Firre-Dxz4* (TKO, turquoise) samples at the chromosome-level. An allelic ratio of 1 corresponds to 100% CAST expression, and 0 corresponds to 100% expression from the BL6 allele. Boxplots represent the interquartile range from the median, with whiskers indicating 1.5x the interquartile range.
b, Violin plot showing the allelic ratios of the X chromosome for single cells from WT and -/+ TKO samples. An allelic ratio of ≥ 0.7 indicates that the CAST allele is Xa, whereas a ratio of ≤ 0.3 indicates that the BL6 allele is Xa. Individual dots represent single cells.
c, Schematic overview of the bulk RNA-seq workflow to obtain WT and -/+ TKO samples. CAST males and -/+ TKO females (BL6) were mated, and spleens were harvested from female F1 offspring (*n* = 6). Subsequently, bulk RNA-seq was performed.
d, Violin plot displaying the allelic ratios of X-linked genes per genotype and replicate (*n* = 3 per genotype). The dots highlight the allelic ratio of *Xist*.
e, Comparison of the allelic ratios of *Xist* for WT (*n* = 3) and heterozygous TKO (*n* = 3) bulk RNA-seq samples. Differences in the allelic ratios were assessed using a t-test. The figure was obtained from Hasenbein *et al.*, 2024[322].

To identify potential gene-level effects in mutant mice on random XCI, cells were computationally sorted according to the strain background of Xa and pooled as pseudobulk to maximize read coverage. This analysis allowed for a background matched comparison between WT and TKO cells with the deletion present on either Xa or Xi. Known escape genes such as *Kdm6a*, *Eif2s3x*, *Ftx*, and *Kdm5c* were detected as biallelically expressed, while *Xist* was expressed exclusively from the Xi, validating the approach (**Figure 5.10a**). However, similar to the results for imprinted XCI, no changes in the allelic ratios of X-linked genes were observed between WT and mutant samples with the TKO on Xa or Xi (**Figure 5.10a**). Thus, the deletion of the *Crossfirre*, *Firre*, and *Dxz4* loci does not impact random XCI biology *in vivo*.

Although the XCI biology was not affected by the TKO, it is noteworthy that the cell-type composition was altered in heterozygous TKO samples compared to the WT (**Figure 5.10b**). In spleen samples with the TKO on Xi, fewer CD4 T cells were present (Fisher odds = 0.71, *p*-value = 0.003). Conversely, for cells with the TKO on the Xa, the proportion of B cells was reduced in the spleen sample of mutant mice (Fisher odds = 0.71, *p*-value = 0.023, **Figure 5.10c**). Taken together, these findings demonstrate differences in cell-type composition depending on whether the deletion is present on Xa or Xi.

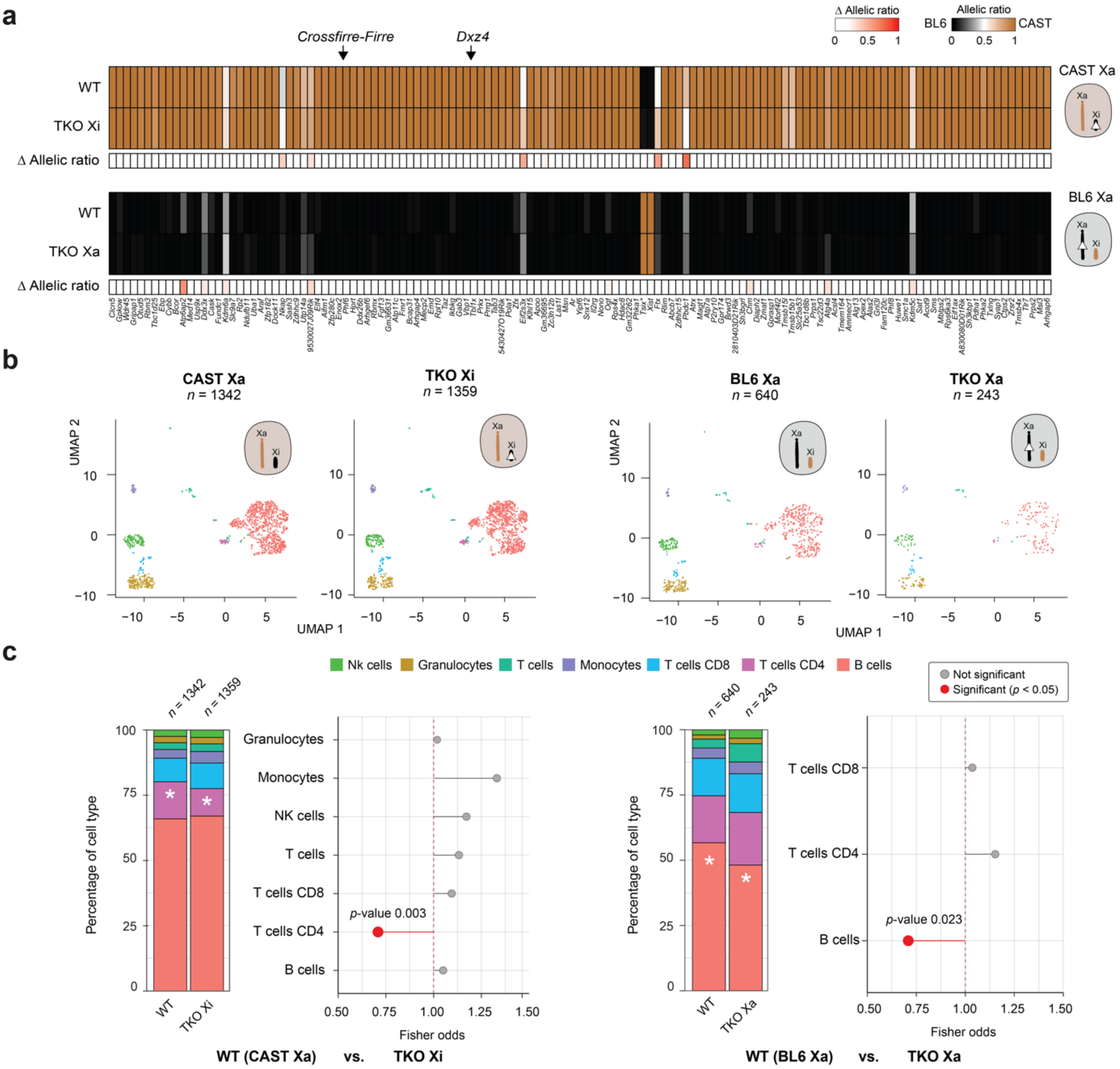


Figure 5.10 Pseudobulk and cell-type composition analysis using scRNA-seq data.
a, Heatmap of pseudobulk samples showing the median allelic ratio of informative X-linked genes. Sequencing reads were aggregated for wildtype (WT) and -/+ Δ*Crossfirre-Firre-Dxz4* (TKO) samples with the TKO located on either Xi (upper panel) or Xa (lower panel). The color scale represents the allelic ratio and ranges from 1 (brown, CAST Xa) to 0 (black, BL6 Xa). The heatmap also shows the absolute delta change in allelic ratios between WT and heterozygous TKO samples. Arrows mark the approximate positions of *Crossfirre*, *Firre,* and *Dxz4*. *Note: The observed expression bias of *Tsix* is an artifact of the non-strand-specific analysis influenced by *Xist,* which is transcribed in the antisense direction.
b, Uniform Manifold Approximation and Projection (UMAP) visualization of the unsupervised clustering of the WT and heterozygous TKO cells, divided by cells with the CAST allele Xa or the BL6 allele Xa. For the heterozygous TKO, this resulted in the deletion being present on Xi or Xa, respectively. Different colors represent distinct cell types.
c, Cell-type composition shifts depending on whether the TKO deletion is located on Xi (left panel) or Xa (right panel). Bar plots show the percentage of each cell type for WT and -/+ TKO samples, with colors corresponding to cell types as in (**b**). Asterisks indicate statistically significant changes in cell-type proportions as determined by Fisher's exact test (*p*-value ≤ 0.05). For cell types with ≥ 20

cells, the odds ratio of the Fisher test is shown next to the bar graph. Statistically significant changes are highlighted in red. The figure was modified from Hasenbein *et al.*, 2024[322].

5.1.5 Mice lacking *Crossfirre, Firre*, and *Dxz4* show upregulation of multiple autosomal pathways

Previously, *Firre* and *Dxz4* have been shown to affect autosomal gene expression in adult organs[127,131]. To further investigate the additional effect of *Crossfirre*, a transcriptomic bodymap was generated for homozygous TKO mice, including the spleen, kidney, lung, heart, liver, and brain. In addition, organ and age-matched samples from the previously published ∆*Firre-Dxz4* mutants were reanalyzed. Interestingly, the additional knockout of *Crossfirre* resulted in an 11.4-fold increase in the number of dysregulated genes (TKO *n* = 1190, ∆*Firre-Dxz4 n* = 104, FDR ≤ 0.01, |shrunk $\log_2$FC| ≥ 1, **Figure 5.11a**). In the TKO, the spleen exhibited the highest number of differentially expressed genes (*n* = 417), with the majority located on autosomes (97.1%, **Figure 5.11a**). In total, 148 differentially expressed genes were common in ≥ 2 tissues, with 93.2% sharing the direction of dysregulation. Genes shared in ≥ 5 organs were upregulated in all samples (**Figure 5.11b**). Subsequently, gene set enrichment analysis (GSEA) revealed that the common dysregulation led to a predominant upregulation of mitochondrial and ribosomal gene-sets across five tissues except the brain (FDR ≤ 0.1, **Figure 5.11c**).

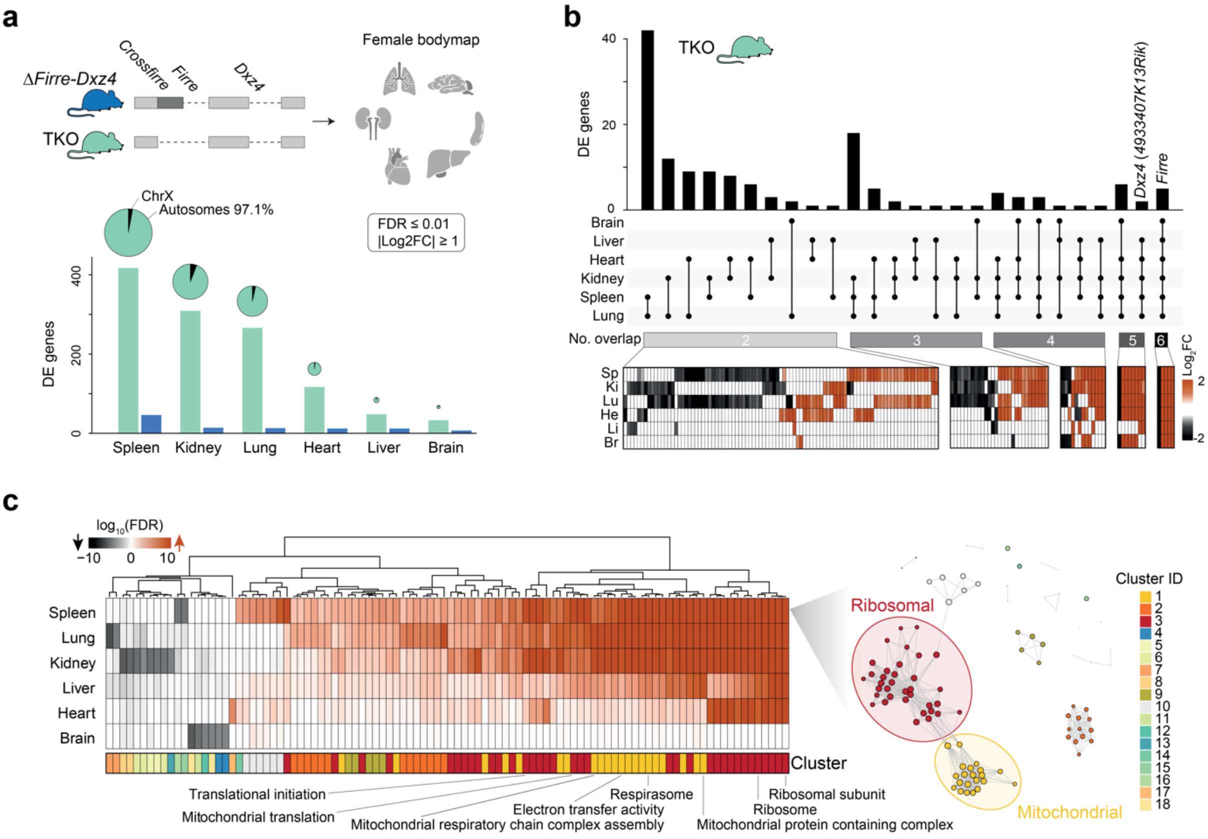


Figure 5.11 Homozygous triple deletion of *Crossfirre*, *Firre*, and *Dxz4* results in widespread autosomal dysregulation observed across multiple organs.

a, Transcriptomic analysis of homozygous Δ*Firre-Dxz4* and Δ*Crossfirre-Firre-Dxz4* (TKO) mouse models. A transcriptomic bodymap was generated for adult female -/- TKO mice covering six different organs and analyzed together with the previously published sample- and age-matched Δ*Firre-Dxz4* mouse models[127] (wildtype n = 4; Δ*Firre-Dxz4* n = 4; TKO n = 3). The number of significantly differentially expressed genes (DEGs, DEseq2: FDR ≤ 0.01, $|\log_2 FC| \geq 1$) is shown as a bar graph for each tissue and genotype (TKO: turquoise, Δ*Firre-Dxz4*: blue). Pie charts illustrate the distribution of TKO DEGs on autosomes and the X chromosome, respectively, with sizes proportional to the number of DEGs.

b, Number of shared DEGs in the TKO across organs. The accompanying heatmap presents the $\log_2$ fold changes for genes shared between the tissues. The color code indicates up- (orange) and down-regulated (black) genes.

c, Heatmap of the top 100 most significantly enriched gene sets based on $\log_{10}$(FDR) values from the TKO gene set enrichment analysis (left; FDR ≤ 0.1). The color code represents up- and down-regulated gene sets. Different colors indicate cluster IDs, which were calculated based on gene similarity (n = 18). The network plot shows the different gene set clusters from the spleen, highlighting the ribosomal (cluster ID: 3 n = 35) and mitochondrial (cluster ID: 1 n = 21) clusters. The figure was obtained from Hasenbein *et al.*, 2024[322].

5.1.6 Double deletion of *Crossfirre* and *Firre* drives the autosomal gene dysregulation

To identify the driving loci of the observed molecular phenotype in TKO organs, spleens from all homozygous mouse mutant models were further investigated, including Δ*Crossfirre*, Δ*Firre*, Δ*Dxz4*, Δ*Crossfirre-Firre*, Δ*Firre-Dxz4* and TKO. Again, differential gene expression analysis was performed using sample-matched WT controls (**Figure 5.12a**). Most of the dysregulated genes were detected in mutant models with the combined *Crossfirre-Firre* deletion (TKO: *n* = 417, Δ*Crossfirre-Firre*: *n* = 103, FDR ≤ 0.01, |shrunk $\log_2$FC| ≥ 1), with the majority of genes being dysregulated in the same direction (*n* = 73, 70.87%, **Figure 5.12b**). Notably, the knockout of either Δ*Crossfirre* or Δ*Firre* individually resulted in few dysregulated genes (Δ*Crossfirre*: *n* = 9, *Firre*: *n* = 7), suggesting a combined effect of both loci. Subsequent GSEA revealed that the *Crossfirre-Firre* knockout reproduced the molecular phenotype of the TKO in the spleen, including the upregulation of mitochondrial and ribosomal pathways (**Figure 5.12c**). Interestingly, the molecular phenotype could not be reproduced in any of the single deletion models, indicating that *Crossfirre* and *Firre* combined affect the autosomal gene dysregulation observed in TKO samples.

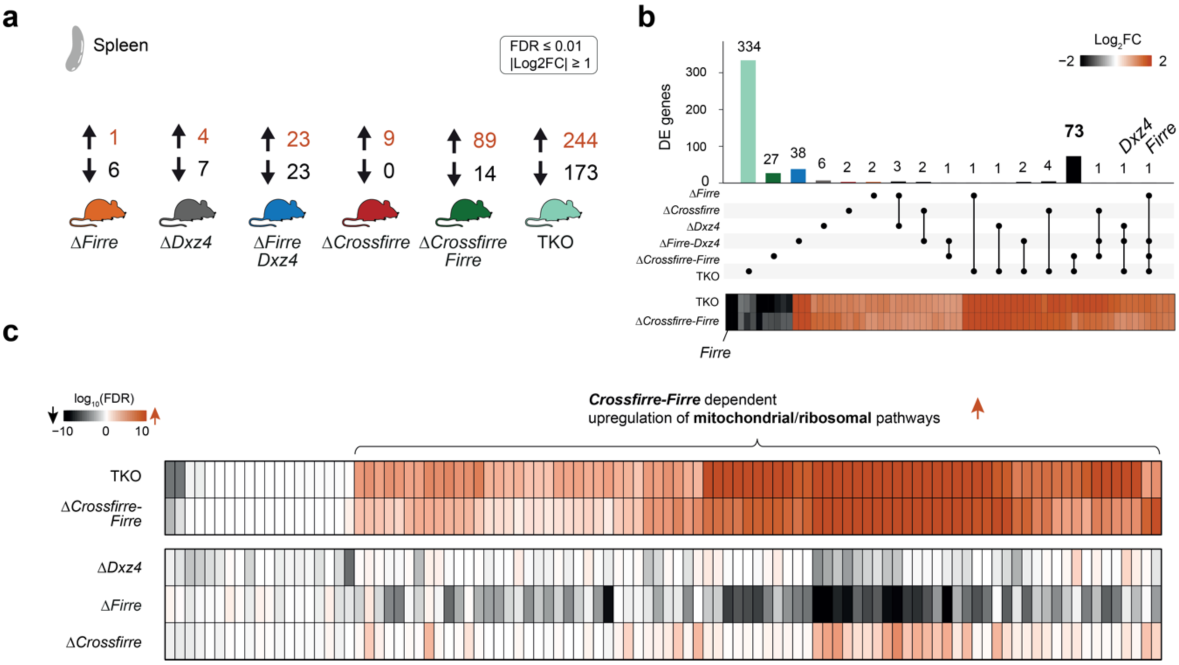


Figure 5.12 The combined deletion of *Crossfirre* and *Firre* drives the upregulation of mitochondrial and ribosomal pathways.
a, Differential gene expression analysis for female spleen samples of different mouse mutant models. The number of differentially expressed genes (DEGs) is displayed for various knockout mice, including the single, double, and triple deletions (DEseq2: FDR ≤ 0.01, |shrunk $\log_2$FC| ≥ 1). Genes are categorized as up- (orange) or downregulated (black).

b, Number of knockout-specific DEGs in the spleen and DEGs shared among mutants. The heatmap presents the $\log_2$ fold changes for genes shared between ∆*Crossfirre-Firre* and ∆*Crossfirre-Firre-Dxz4* (TKO).
c, Heatmap showing the $\log_{10}$(FDR) values for the top 100 informative enriched gene sets identified in the TKO. The heatmap shows upregulated and downregulated gene sets across the ∆*Crossfirre*, ∆*Firre*, and ∆*Dxz4* single deletions and the ∆*Crossfirre-Firre* and TKO models. The figure was taken from Hasenbein *et al.*, 2024[322].

5.1.7 Phenotyping pipeline uncovers sex-specific phenotypes in TKO mutants

To further investigate the phenotypic characteristics of the observed gene dysregulation of the *Crossfirre*, *Firre*, and *Dxz4* deletion, a cohort of control and TKO mice (females: *n* = 13 TKO, *n* = 15 WT, males: *n* = 13 TKO, *n* = 15 WT) was subjected to the comprehensive phenotyping pipeline of the GMC[324]. This analysis included multiple phenotypic screens with hundreds of tests covering the categories: immunology/allergy, behavior, biomarkers, cardiovascular, clinical chemistry, pathology, dysmorphology, eyes, metabolism, neurology, and nociception (**Figure 5.13a**). A detailed description of the phenotyping pipeline is available at https://www.mouseclinic.de.

A total of 28 knockout- and sex-specific phenotypes were identified by the phenotypic screen, encompassing immunology/allergy (*n* = 5), behavior (*n* = 2), neurology (*n* = 1), cardiovascular (*n* = 2), clinical chemistry (*n* = 8), dysmorphology (*n* = 2), metabolism (*n* = 3), and pathology (*n* = 5, **Figure 5.13b**).

Nine of the 28 observed phenotypes were TKO-specific, independent of sex. These included (i) an increased locomotor and (ii) exploratory activity, which was most pronounced during the first five minutes of the observation. In contrast, (iii) the acoustic startle reactivity was decreased compared to the control mice. In addition, TKO mutants showed (iv) altered red blood cell morphology with decreased mean corpuscular volume and increased mean corpuscular hemoglobin concentration. Mild effects were observed on (v) iron metabolism, including reduced plasma iron levels and calculated transferrin saturation. Additionally, (vi) pathology of the intestinal Peyer's patches revealed that mutant mice had increased secondary follicles (**Figure 5.13b-c**). Furthermore, consistent with the scRNA-seq analysis of the spleen, shifts were observed in the (vii) CD4/CD8 T cell ratios, as well as in the (viii) relative percentages of B cells and (ix) monocytes of the peripheral blood (**Figure 5.13b-c**).

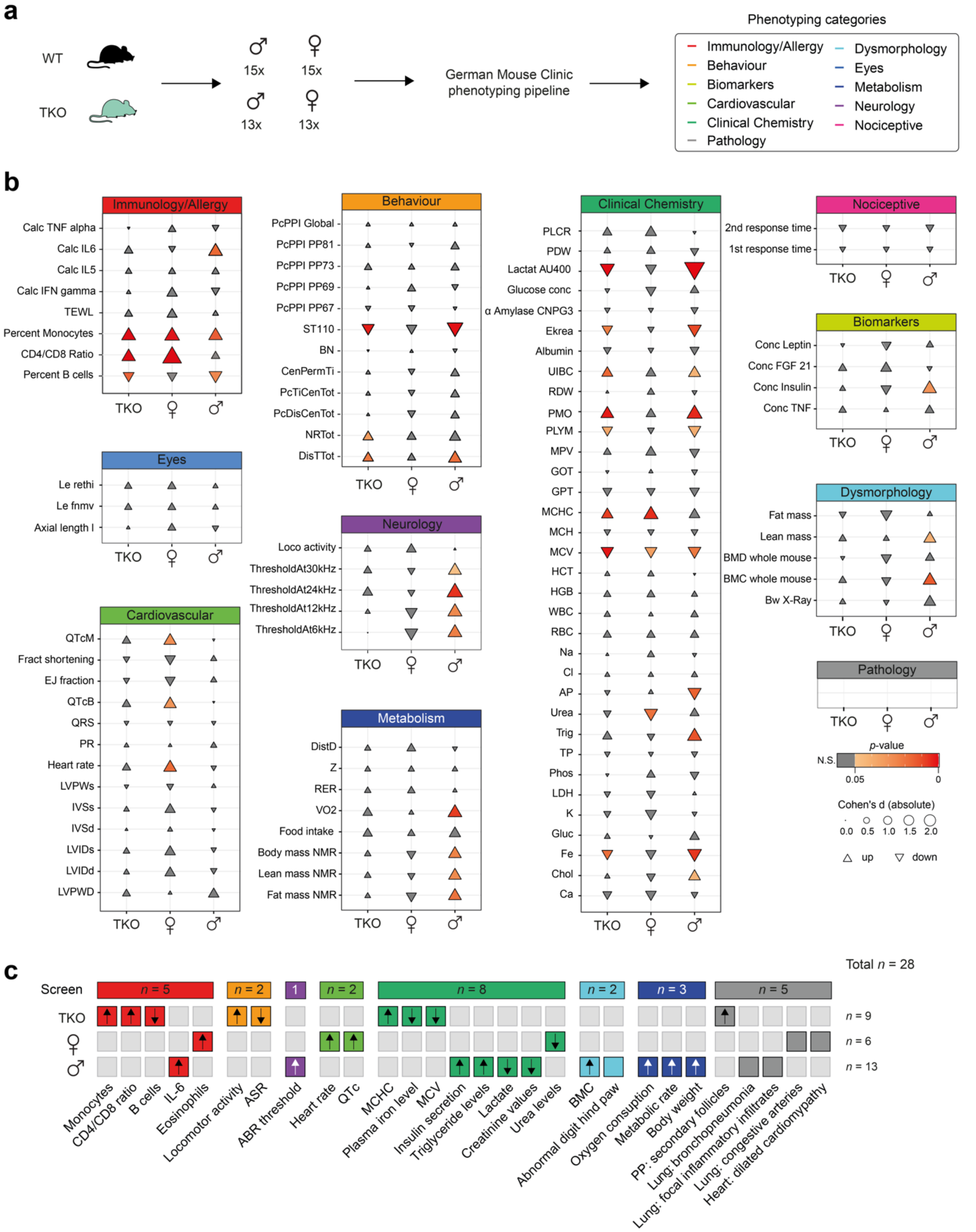


Figure 5.13 Comprehensive phenotyping pipeline identifies knockout- and sex-specific phenotypes.

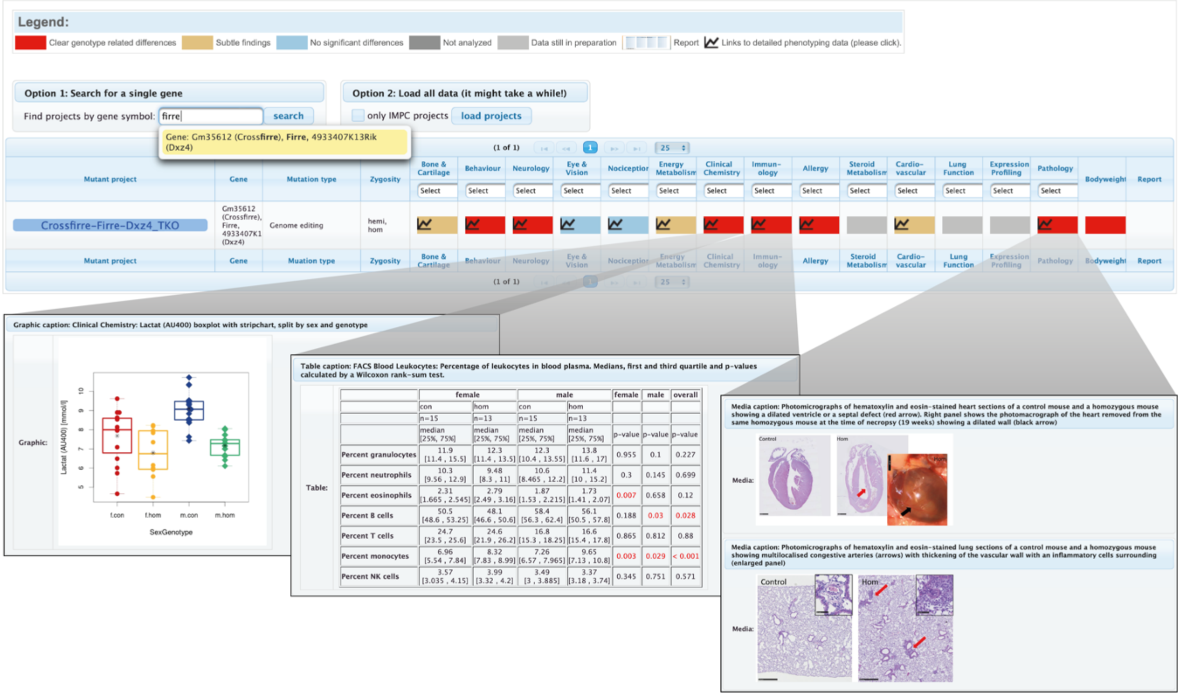
a, Schematic overview of the phenotyping process with the German Mouse Clinic (GMC). A total of 30 wildtype (male *n* = 15, female *n* = 15) and 26 ∆*Crossfirre-Firre-Dxz4* (TKO, male *n* = 13, female *n* = 13) mice were subjected to the phenotyping pipeline of the GMC. Various phenotyping screens were performed covering the categories: immunology/allergy, behavior, biomarkers, cardiovascular, clinical chemistry, pathology, dysmorphology, eyes, metabolism, neurology, and nociception.

b, Overview of the results of a set of parameter tests for each category. The parameters are selected to provide a summary overview of the phenotyping results of the GMC. Triangles indicate the direction of the effect sizes (Cohen's D), with color coding according to significance (*p*-value < 0.05). N.S.: not significant (t-test). An overview of the phenotyping screen abbreviations is provided in Appendix 10.2. **c**, Overview of all significant parameters by screening category for TKO (*n* = 9), female- (*n* = 6), and male-specific (*n* = 13) phenotypes. The color coding corresponds to the respective screening category, while the arrows denote the direction of effect sizes (Cohen's D). The figure was obtained from Hasenbein *et al.*, 2024[322].

In addition to the knockout-specific phenotypes, the GMC revealed sex-specific effects for TKO mice. A total of 13 phenotypes were identified as male-specific. Clinical chemistry detected (i) elevated insulin secretion and (ii) plasma triglyceride levels, along with (iii) reduced creatinine and (iv) lactate concentrations. The immunology screening found (v) higher levels of IL-6, a proinflammatory cytokine, in mutant males. In addition, males exhibited (vi) increased body weight and (vii) oxygen consumption, as well as (viii) a higher metabolic rate attributed to (ix) the increase in lean mass. The pathology screen identified (x) bronchopneumonia and mild inflammatory cell infiltrates in a subset of male mutants (1/5). The (xi) bone mineral content was increased, and (xii) 2/13 male mice showed abnormal hind paws digits. Notably, an (xiii) increased auditory brainstem response was identified in the neurology screen, indicating alterations in auditory processing (**Figure 5.13b-c**).

In contrast to the male-specific phenotypes, female mice exhibited six phenotypes, including (i) increased eosinophilic proportions and (ii) mean platelet volumes, while (iii) urea levels were decreased. Apparent changes in 1/5 female mice were detected by histopathological screens, including (iv) inflammatory cell infiltrates in the lungs and (v) congestive arteries with vessel wall thickening. These changes were mild and focal in the remaining females (4/5). Cardiological tests revealed subtle shifts (vi) characterized by higher heart rates and heart-rate-corrected QT intervals. Interestingly, one female mutant further exhibited dilated cardiomyopathy (**Figure 5.13b-c**).

In summary, the extensive phenotyping of the GMC highlights that *Crossfirre*, *Firre*, and *Dxz4* are involved in various physiological processes. The majority of phenotypes were sex-specific (female TKO-specific: 21.43%, *n* = 6; male TKO-specific: 46.43%, *n* = 13), while only a subset (32.14%, *n* = 9) was attributed to the TKO independent of sex. The discovery of sex-specific phenotypes underscores the highly sex-specific characteristics of these loci. A detailed overview of all phenotyping results and raw measurements can be found on the GMC Phenomap website (https://www.mouseclinic.de, **Figure 5.14**).



Legend:

🟥 Clear genotype related differences 🟨 Subtle findings 🟦 No significant differences ⬛ Not analyzed ⬜ Data still in preparation ▯ Report 📈 Links to detailed phenotyping data (please click).

Option 1: Search for a single gene

Find projects by gene symbol: firre search

Gene: Gm35612 (Crossfirre), Firre, 4933407K13Rik (Dxz4)

Option 2: Load all data (it might take a while!)

□ only IMPC projects load projects

(1 of 1) 1 25

Mutant project	Gene	Mutation type	Zygosity	Bone & Cartilage	Behaviour	Neurology	Eye & Vision	Nociceptior	Energy Metabolism	Clinical Chemistry	Immun-ology	Allergy	Steroid Metabolism	Cardio-vascular	Lung Function	Expression Profiling	Pathology	Bodyweight	Report
				Select	Select	Select	Select	Select	Select	Select	Select	Select	Select	Select	Select	Select	Select		
Crossfirre-Firre-Dxz4_TKO	Gm35612 (Crossfirre), Firre, 4933407K1 (Dxz4)	Genome editing	hemi, hom	🟨 📈	🟥 📈	🟥 📈	🟦 📈	🟦 📈	🟨 📈	🟥 📈	🟥 📈	🟥 📈	⬜	🟨 📈	⬜	⬜	🟥 📈	🟥	
Mutant project	Gene	Muation type	Zygosity	Bone & Cartilage	Behaviour	Neurology	Eye & Vision	Nociceptior	Energy Metabolism	Clinical Chemistry	Immun-ology	Allergy	Steroid Metabolism	Cardio-vascular	Lung Function	Expression Profiling	Pathology	Bodyweight	Report

(1 of 1) 1 25

Graphic caption: Clinical Chemistry: Lactat (AU400) boxplot with stripchart, split by sex and genotype

Table caption: FACS Blood Leukocytes: Percentage of leukocytes in blood plasma. Medians, first and third quartile and p-values calculated by a Wilcoxon rank-sum test.

Table:

	female		male		female	male	overall
	con	hom	con	hom			
	n=15	n=13	n=15	n=13			
	median [25%, 75%]	median [25%, 75%]	median [25%, 75%]	median [25%, 75%]	p-value	p-value	p-value
Percent granulocytes	11.9 [11.4 , 15.5]	12.3 [11.4 , 13.5]	12.3 [10.4 , 13.55]	13.8 [11.6 , 17]	0.955	0.1	0.227
Percent neutrophils	10.3 [9.56 , 12.9]	9.48 [8.3 , 11]	10.6 [8.465 , 12.2]	11.4 [10 , 15.2]	0.3	0.145	0.699
Percent eosinophils	2.31 [1.665 , 2.545]	2.79 [2.49 , 3.16]	1.87 [1.53 , 2.215]	1.73 [1.41 , 2.07]	0.007	0.658	0.12
Percent B cells	50.5 [48.6 , 53.25]	48.1 [46.6 , 50.6]	58.4 [56.3 , 62.4]	56.1 [50.5 , 57.8]	0.188	0.03	0.028
Percent T cells	24.7 [23.5 , 25.6]	24.6 [21.9 , 26.2]	16.8 [15.3 , 18.25]	16.6 [15.4 , 17.8]	0.865	0.812	0.88
Percent monocytes	6.96 [5.54 , 7.84]	8.32 [7.83 , 8.99]	7.26 [6.57 , 7.965]	9.65 [7.13 , 10.8]	0.003	0.029	< 0.001
Percent NK cells	3.57 [3.035 , 4.15]	3.99 [3.32 , 4.2]	3.49 [3 , 3.885]	3.37 [3.18 , 3.74]	0.345	0.751	0.571

Media caption: Photomicrographs of hematoxylin and eosin-stained heart sections of a control mouse and a homozygous mouse showing a dilated ventricle or a septal defect (red arrow). Right panel shows the photomacrograph of the heart removed from the same homozygous mouse at the time of necropsy (19 weeks) showing a dilated wall (black arrow)

Media caption: Photomicrographs of hematoxylin and eosin-stained lung sections of a control mouse and a homozygous mouse showing multilocalised congestive arteries (arrows) with thickening of the vascular wall with an inflammatory cells surrounding (enlarged panel)

Figure 5.14 Overview of the Phenomap webpage of the German Mouse Clinic.
The Phenomap resource comprises all raw measurements of the comprehensive TKO phenotyping for each category assessed. The data is publicly accessible at https://www.mouseclinic.de and can be found by searching for the *Crossfirre*, *Firre,* or *Dxz4* gene name.

5.2 Project 2: Decoding the targets and mechanisms of the non-coding genome through allele-specific genomics

Given that the experimental characterization of ncRNAs is laboratory extensive, the second project aimed to develop a novel bioinformatics framework to predict the target genes and mechanisms of ncRNAs.

5.2.1 Enrichment of allele-specific ncRNAs nearby allele-specific pcGenes

Previously, a comprehensive map of ASE was generated across multiple mouse organs and observed that the number of allele-specific ncRNAs correlated with the number of allele-specific pcGenes[78]. In this study, the allele-specific transcriptome was mapped across the major mouse organs including the brain, heart, lung, liver, kidney, and spleen, to further investigate this correlation. Organs were collected from nine-week-old F1 hybrids and used for RNA-seq (replicates *n* = 3; **Figure 5.15a**). Subsequently, ASE mapping was performed for loci that were consistently informative across replicates using the Allelome.PRO v2.0 pipeline.

On average, ASE was observed for 8.98% (*n* = 1039) of the informative genes per tissue (allelic ratio cutoff ≥ 0.7 or ≤ 0.3, **Figure 5.15b**). The highest proportion of allele-specific genes was found in the liver (10.5%, *n* = 1007), while the lowest amount was present in the brain (6.4%, *n* = 840, **Figure 5.15c**). Of these, an average of 2.13% were ncRNAs, with lncRNAs accounting for the majority of biotypes (69.4%, **Figure 5.15d**).

Next, the proportion of allele-specific ncRNAs and pcGenes was correlated across tissues, identifying a positive correlation (Spearman correlation: R = 0.66, *p*-value = 0.004, **Figure 5.15e**). The co-occurrence of allele-specific pcGenes and ncRNAs was quantified across various genomic window sizes to confirm this correlation. This analysis revealed a strong enrichment of allele-specific ncRNAs in the vicinity of allele-specific pcGenes within a distance of ±100kb (Wilcoxon test *p*-value = 0.002, **Figure 5.15f**). The finding that allele-specific ncRNAs often surround allele-specific pcGenes suggests a potential co-regulation and indicates that the allele-specific pattern could be used to predict the protein-coding targets of *cis*-acting ncRNAs. Furthermore, this approach allows to infer the mechanism of action based on the pattern of the allelic bias between ncRNA and pcGene towards the same (enhancing) or opposite (repressive) alleles (**Figure 5.15g**).

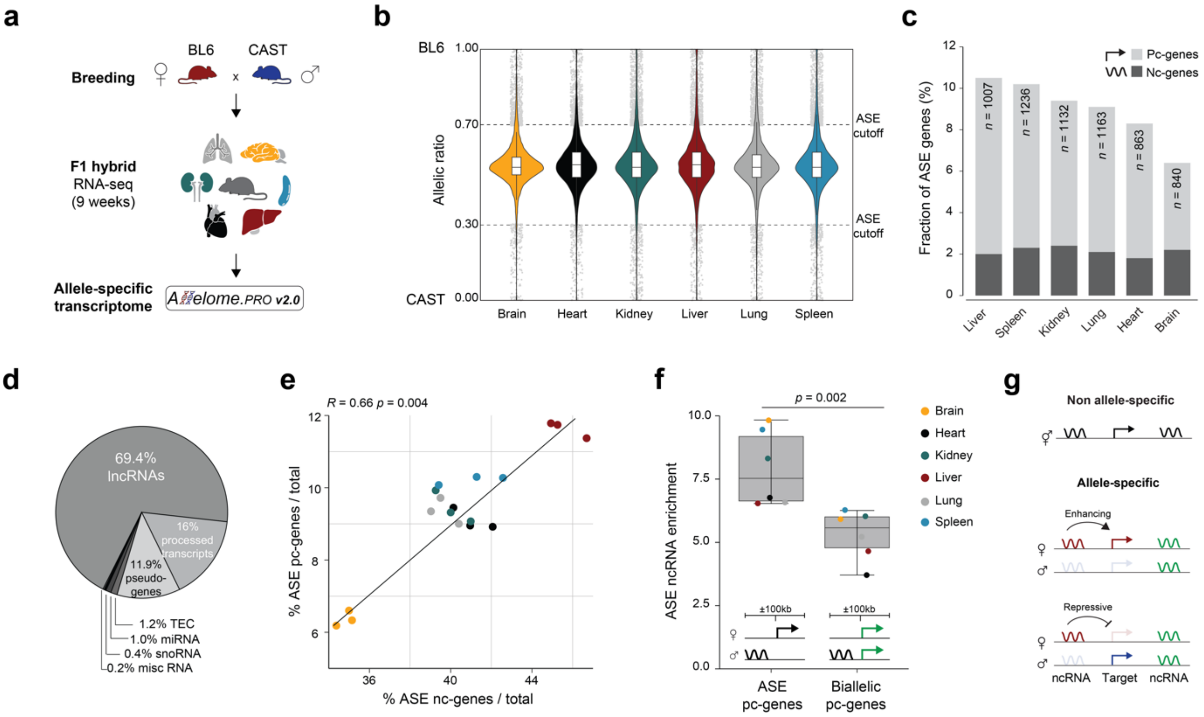


Figure 5.15 Allele-specific non-coding RNAs are enriched near allele-specific protein-coding genes.

a, Overview of the workflow to map the allele-specific transcriptome. BL6 females were crossed with CAST males. Nine-week-old female F1 organs (brain, spleen, liver, heart, kidney, lung) were harvested from three replicates (*n* = 18). RNA-seq was conducted and Allelome.PRO v2.0 was employed to generate an allele-specific bodymap.

b, Violin plot displaying the median allelic ratio for informative autosomal genes across replicates (*n* = 3, total reads ≥ 20). Colors represent different tissues. Allele-specific expression (ASE) was defined by an allelic ratio cutoff of ≥ 0.7 or ≤ 0.3. Genes beyond these cutoff values are indicated as dots. Boxplots indicate the interquartile range and median, while whiskers represent 1.5x the interquartile range.

c, Bar plot illustrating the fraction of allele-specific genes per tissue. Numbers denote the total count of biased genes. Light gray and dark gray colors represent protein-coding and non-coding genes, respectively. Pc: protein-coding, nc: non-coding.

d, Pie chart showing the biotype distribution of the ncRNAs with an allele-specific bias. Biotype information was available for 52.74% of the ncRNAs (*n* = 520). Misc RNA: miscellaneous RNAs without classification, TEC: to be experimentally confirmed.

e, Correlation plot showing the proportion of allele-specific ncRNAs against the fraction of allele-specific protein-coding genes for each replicate, normalized by the total amount of informative genes. Spearman correlation was calculated as a statistical test (*R* = 0.66, *p*-value = 0.004). The color coding is according to the tissue sample.

f, Boxplot displaying the enrichment of allele-specific ncRNAs around allele-specific and biallelic protein-coding genes within a ±100kb window. The colors represent the various tissues. Statistical significance was assessed using a Wilcoxon test (*p*-value = 0.002). Boxplots show the interquartile range, median, and whiskers range from maximum to minimum values.

g, Schematic overview of the mechanism prediction based on the allele-specific pattern. Depending on whether the allelic bias between ncRNA and pcGene was towards the same or opposite alleles, they were classified as enhancing or repressive, respectively. The figure was obtained from a submitted manuscript by the author of this thesis (see 9.2 Submitted manuscripts, 1.).

5.2.2 Validation of the Allelome.LINK pipeline using known targets of imprinted lncRNAs

Based on the findings that allele-specific ncRNAs and allele-specific pcGenes correlate with each other, the Allelome.LINK tool was developed to facilitate the target prediction of *cis*-acting ncRNAs. Allelome.LINK is a bioinformatics framework that builds upon the Allelome.PRO v2.0 pipeline to link regulatory loci to their potential target genes based on the ASE pattern. The mechanism is determined based on the allelic bias towards the same or opposite alleles (**Figure 5.16a**). Both, Allelome.PRO v2.0 and Allelome.LINK, offer straightforward execution through a simple one-line command, improving the accessibility for users of diverse backgrounds. The results from Allelome.LINK are provided in a tabular format sorted by a linkage score, along with a BEDPE file for genome browser visualization. A comprehensive overview of both tools is available in the **Appendix section 10.1**.

The pipeline was initially evaluated using the lncRNA *Xist*. *Xist* is the initiator of XCI, a process in which one of the two female X chromosomes undergoes epigenetic silencing to achieve dosage compensation between males and females[97,98]. Although XCI becomes random after embryonic implantation, it consistently results in the silencing of the paternal X chromosome in extraembryonic lineages[104]. In the placenta, Allelome.LINK correctly identified *Xist* as a repressive ncRNA for the majority of X-linked genes, with the exception of known escape genes, such as *Kdm6a*, *Eif2s3x*, *Jpx*, and *Kdm5c*, showcasing the robustness of the Allelome.LINK strategy (**Figure 5.16b**).

To further evaluate the efficacy of the pipeline, the target prediction for the lncRNA *Airn* was tested. *Airn* is an imprinted, paternally expressed lncRNA that silences target genes within the *Igf2r*/*Airn* cluster in a *cis*-dependent manner. In the placenta, the *Airn* locus is the largest imprinted region in mice[78]. Here, the paternally expressed lncRNA *Airn* represses multiple genes on the paternal allele, leading to maternal expression of the targets. Among these, seven genes (*Pde10a*, *Park2*, *Slc22a3*, Igf2r, *Dact2*, *Smoc2*, and *Thbs2*) were validated as repressive targets by reactivation of the silent allele upon deletion of the *Airn* promoter[78,174]. To assess the efficacy of the pipeline, Allelome.LINK was employed to analyze WT and *Airn* knockout placental RNA-seq datasets[78]. Allelome.PRO v2.0 accurately identified maternal bias for the known *Airn* targets and confirmed biallelic expression of the target genes upon *Airn* promoter deletion (**Figure 5.16c**).

Knowing the *cis*-acting targets of the lncRNA *Airn* in the placenta allowed the computation of the precision and recall for the *Airn* locus. This was done separately for each replicate and for pooled samples. The highest precision (85.7%) was obtained by pooling replicates, while the recall remained above 85% (**Figure 5.16d**). These results highlight the effectiveness of the Allelome.LINK pipeline in predicting the targets of ncRNAs.

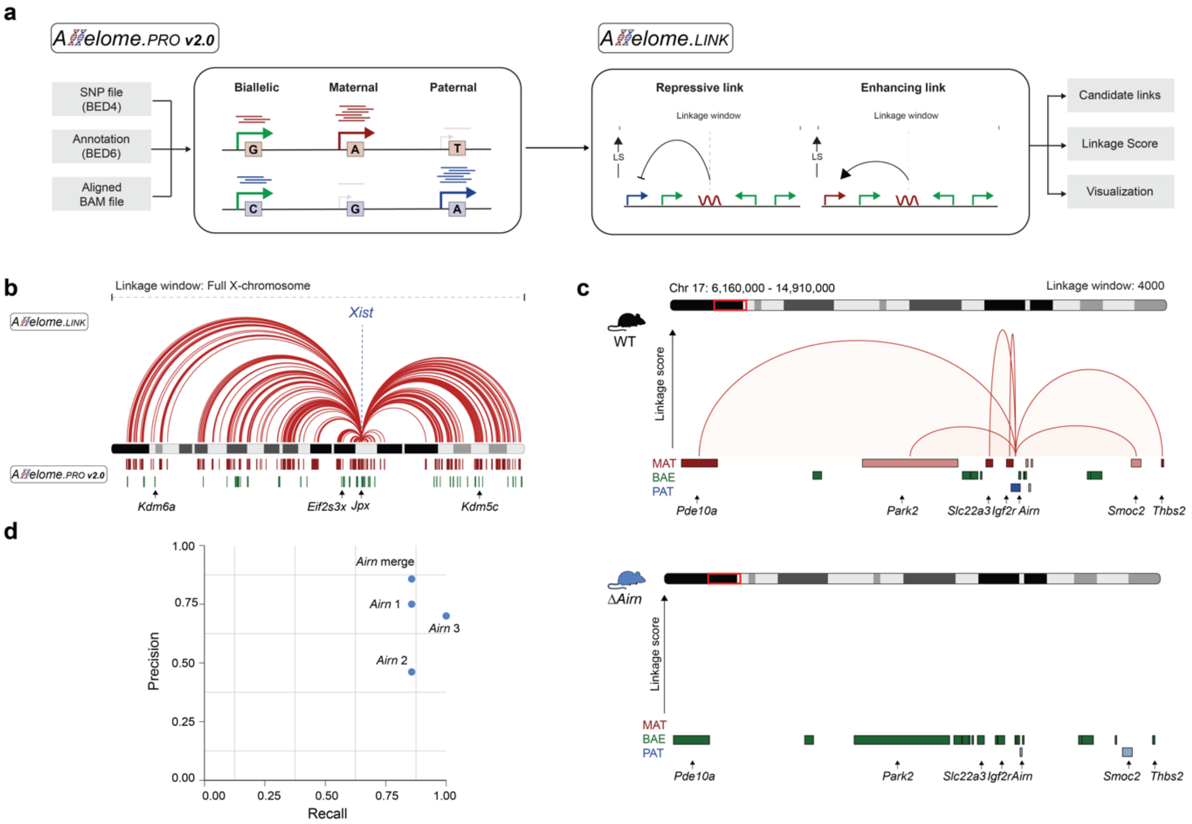


Figure 5.16 Workflow and validation of the Allelome.LINK pipeline.

a, Overview of the Allelome.PRO v2.0 and Allelome.LINK pipeline. Allelome.PRO v2.0 requires three input files: a SNP file, an annotation file, and an aligned sample BAM file. The allelic ratio is calculated for each locus based on the number of reads with SNPs from the maternal or paternal allele. The output of Allelome.PRO v2.0 is used as input for Allelome.LINK. The tool links allele-specific loci within user-defined genomic windows and calculates linkage scores. The output includes a list of candidate predictions and a genome browser file.

b, Genome browser output of the Allelome.LINK pipeline. Shown is the X chromosome with red arcs highlighting repressive interactions between *Xist* and protein-coding X-linked genes (total reads ≥ 20, window size: full chromosome, allelic ratio > 0.75 and < 0.25). Below the chromosome is the Allelome.PRO v2.0 output showing loci classified as maternally (red) or biallelic (green) expressed. Known escape genes are labeled. RNA-seq data from E12.5 placentas were used (CAST x FVB *n* = 2, FVB x CAST *n* = 2)[175].

c, Genome browser output of the Allelome.LINK pipeline for the *Igf2r/Airn* locus, showing the predicted interactions for the imprinted lncRNA *Airn*. Red arcs indicate repressive linkages, with arc height proportional to the linkage score. Below is the Allelome.PRO v2.0 output showing loci classified as maternal (MAT: red), biallelic (BAE: green) and paternal (PAT: blue). The upper panel shows results for wildtype mice (*n* = 3), and the lower panel for mice with an *Airn* promoter deletion (*n* = 3). Samples were pooled using the median (total reads ≥ 10, allelic ratio ≥ 0.7 or ≤ 0.3, window size: 4000kb) and were obtained from E12.5 placentas (*n* = 6)[78].

d, Precision-recall plot for the Allelome.LINK results of the *Igf2r/Airn* locus from E12.5 placentas (*n* = 3). Precision and recall were calculated per replicate and for pooled samples. The figure was sourced from a submitted manuscript by the author of this thesis (see 9.2 Submitted manuscripts, 1.).

5.2.3 Identification of 397 mouse ncRNA-target linkages and their mechanisms across organs

Following validation of the Allelome.LINK pipeline, the tool was used to predict the protein-coding targets of ncRNAs in a comprehensive set of mouse organs. Therefore, the mapped allele-specific transcriptome of nine-week-old animals was used, including samples from the brain, heart, lung, liver, kidney, and spleen (**Figure 5.17a**). On average, the approach identified 66.2 ncRNA-target associations per tissue, ranging from 50 linkages in the heart to 99 linkages in the spleen (**Figure 5.17b**). High-confident linkages were identified by using the linkage score. Notably, the known repressive interaction between *Airn* and *Igf2r* was among the top interactions in all tissues except the brain (**Figure 5.17c**). Using this approach, an average of 11.3% of the allele-specific ncRNAs per tissue were linked to their putative protein-coding target genes (**Figure 5.17d**). Interestingly, the analysis also revealed a predominance of tissue-specific linkages (62.2%, *n* = 247), while only 37.8% (*n* = 150) of the linkages were shared by two or more tissues (**Figure 5.17e**). Additionally, repressive interactions showed an even distribution of target distances, peaking at 32kb, while enhancing linkages were in close proximity (**Figure 5.17f-g**).

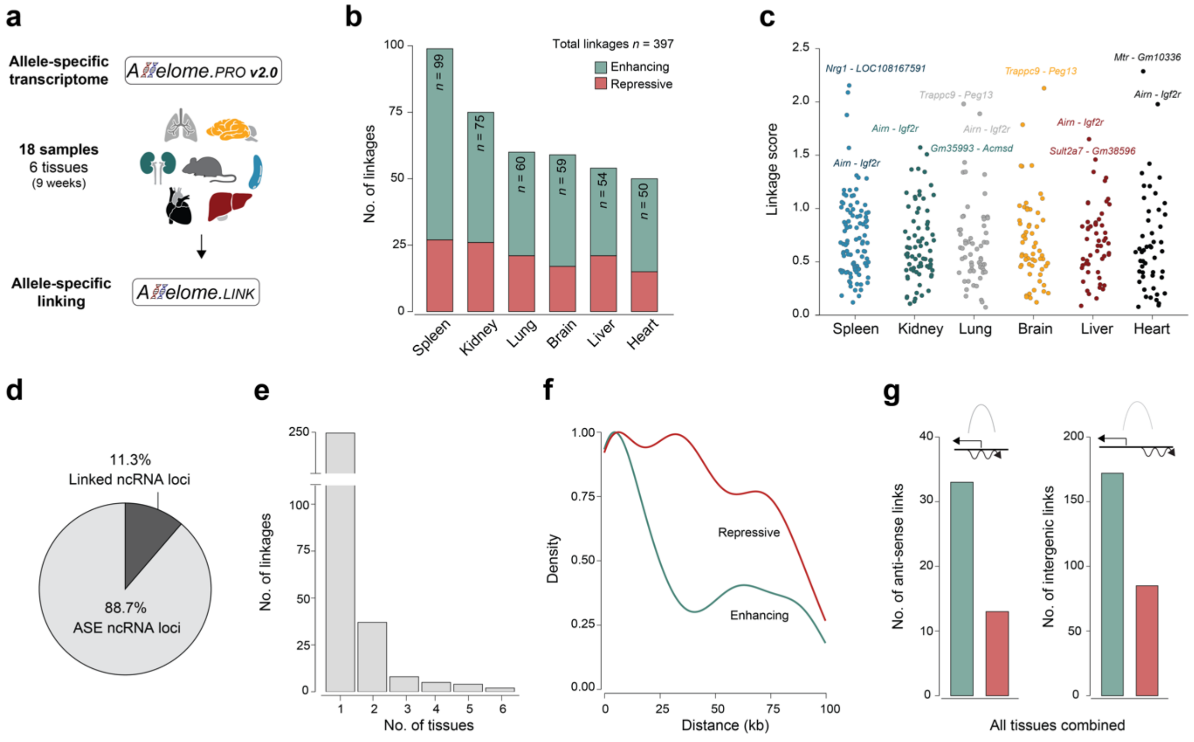


Figure 5.17 Identification of ncRNA-target linkages and their regulatory mechanisms.
a, Schematic overview of the workflow for predicting ncRNA-targets in mice. Allelome.PRO v2.0 was applied to a comprehensive set of organs from nine-week-old F1 mice (BL6 x CAST), including the brain, spleen, liver, heart, kidney, and lung (per replicate *n* = 3, total *n* = 18). The results were used as input for Allelome.LINK to predict candidate linkages.

b, Bar plot showing the number of predicted ncRNA-to-target linkages per tissue. Red bars illustrate repressive interactions, while green color marks enhancing linkages.
c, Manhattan plot displaying the linkage score of candidate linkages per tissue. Colors denote the different tissue samples.
d, Pie chart showing the mean fraction of linked ncRNAs per tissue relative to the total amount of ncRNAs with allele-specific expression.
e, Bar plot illustrating the number of linkages present in the number of tissues.
f, Density plot showing the distribution of linkage distances for enhancing (green) and repressive (red) interactions.
g, Bar plots displaying the number of anti-sense and intergenic linkages separated by enhancing (green) and repressive (red) interactions. The figure was taken from a submitted manuscript by the author of this thesis (see 9.2 Submitted Manuscripts, 1.).

Two examples of high-confident linkages are highlighted. The first example is a repressive anti-sense linkage identified in the kidney between the ncRNA *Gm35993* and the pcGene *Acmsd* (**Figure 5.18a**). Allele-specific read mapping of the underlying RNA-seq data confirmed the predominant expression of *Gm35993* from the maternal allele and the paternal expression of *Acmsd*, indicating a repressive association (**Figure 5.18a**). The second example is an intergenic repressive link in the liver, where *Gm38596* was predicted to repress *Sult2a7*. Allele-specific read mapping confirmed that low expression of *Gm38596* correlated with high expression of *Sult2a7*, while increased expression of *Gm38596* anti-correlated with the loss of *Sult2a7* expression (**Figure 5.18b**).

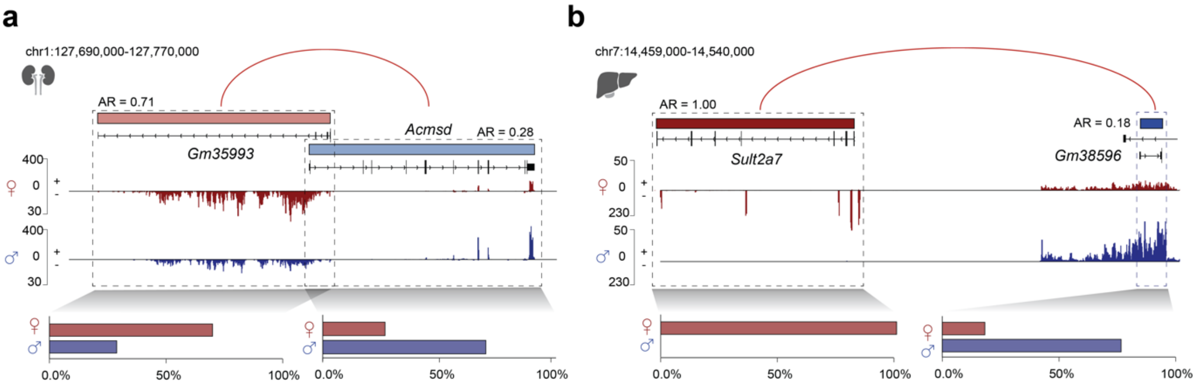


Figure 5.18 Examples of high-confident ncRNA-to-target predictions in mice.
a, Genome browser visualization of the repressive linkage between the ncRNA *Gm35993* and the protein-coding gene *Acmsd*, detected in the kidney. The red arc indicates a repressive interaction. The allelic bias is shown by gene color: red for maternal and blue for paternal expression, with the intensity reflecting the allelic bias. Sequencing tracks display the number of strand-specific reads mapped to the maternal (red) and paternal (blue) allele. The bar chart illustrates the quantification of sequencing reads per allele and gene.
b, Same as in (**a**), but for the repressive interaction between *Gm38596* and *Sult2a7* detected in the liver. The figure was obtained from a submitted manuscript by the author of this thesis (see 9.2 Submitted Manuscripts, 1.).

In total, the Allelome.LINK framework identified 397 ncRNA-to-target linkages across a comprehensive set of mouse organs (**Figure 5.19**). These results provide detailed insights into the tissue-specific nature of ncRNAs within the mouse genome. To support the exploration of all candidate linkages, an interactive resource was created using the Integrative Genomics Viewer (IGV)[329]. This database enables the dynamic visualization and analysis of the linkage data, allowing researchers to select candidates for further investigation and characterization in the tissue of interest. The URL links to access the interactive database are available at the https://github.com/AndergassenLab/Allelome.LINK. Detailed explanations of how to use the generated resource can be found in **Appendix section 10.3**.

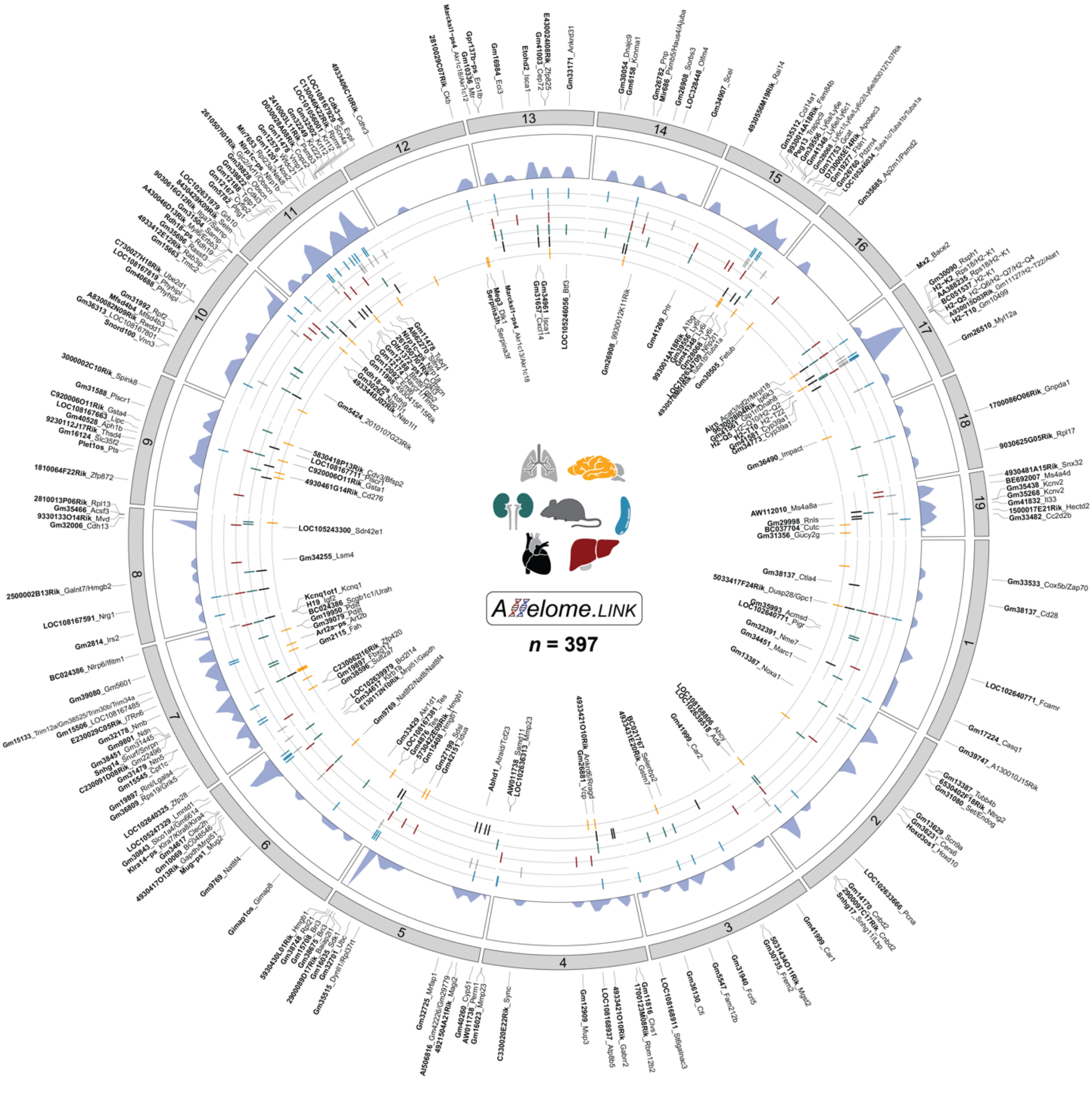


Figure 5.19 Comprehensive ncRNA-to-target resource for the major mouse organs.
Chord plot showing chromosome 1-19 with the predicted candidate linkages between ncRNA and protein-coding target identified in adult mice across six different tissues (spleen: blue, lung: gray,

liver: red, kidney: green, heart: black, brain: yellow). Linkages labeled on the outside are predicted to be enhancing, while linkages on the inside represent repressive interactions. The ncRNAs are highlighted in bold font. Dashes separate multiple protein-coding targets. The density plot shows the genomic distribution of linkages per chromosome. The figure was taken from a submitted manuscript by the author of this thesis (see 9.2 Submitted Manuscripts, 1.).

5.2.4 Leveraging the human genetic variation to uncover the gene targets of ncRNAs

Next, the Allelome.LINK framework was applied to human samples taking advantage of the GTEx resource[191]. The GTEx consortium collected RNA-seq data from up to 54 different tissues across 838 individuals, resulting in 15,253 samples and 153 million allele-specific haplotype measurements (**Figure 5.20a**). Due to the non-strand-specific nature of the GTEx RNA-seq data, overlapping genes were removed. Furthermore, a total read cutoff of ≥ 20 SNP-overlapping reads was required for a gene to be considered informative. Post filtering, 924,440 non-coding and 27,155,698 protein-coding allele-specific measurements remained in the dataset, comprising 1,825 ncRNAs and 6,281 pcGene loci (**Figure 5.20b**). The number of unique allele-specific genes generally increased with sample size, with an average of 3,580 ASE loci per tissue, including 312 ncRNAs and 3,268 pcGenes. The Kidney - Medulla, with data from only four individuals, had the fewest number of ASE genes (*n* = 310), while the lung, with data from 515 individuals, had the highest number of ASE gene loci (*n* = 4,983, **Figure 5.20c**).

Before predicting the protein-coding targets of ncRNAs, the co-occurrence of allelic ncRNAs and pcGenes was quantified to confirm the applicability of this approach in humans. Each tissue sample was screened for allele-specific ncRNAs, and the abundance of allele-specific pcGenes within a distance of ±100kb was compared to the abundance of biallelic pcGenes. Notably, this analysis revealed a significant enrichment of allele-specific ncRNAs around allele-specific pcGenes in half of the tissues examined (27 out of 54 tissues, **Figure 20d**).

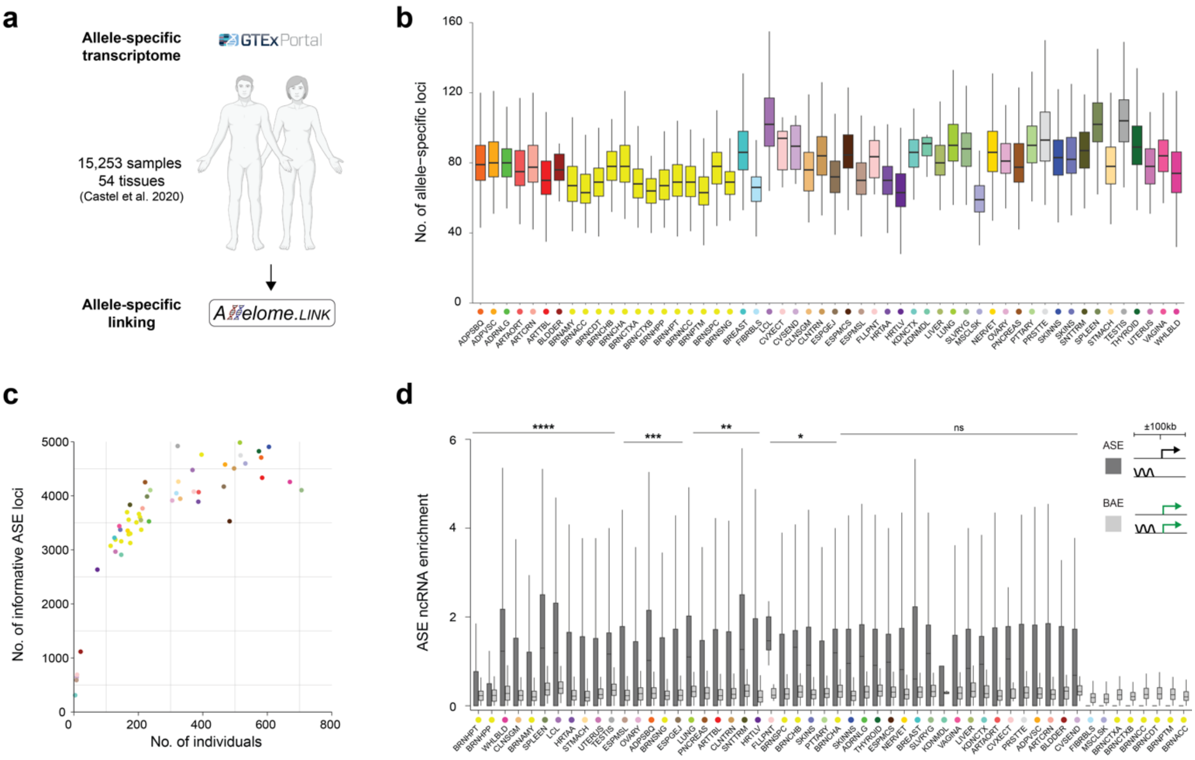


Figure 5.20 Allele-specific ncRNAs are enriched in proximity to allele-specific protein-coding genes in multiple human tissues.

a, Overview of the Allelome.LINK strategy on human samples. The allele-specific transcriptome of 15,253 samples, including 54 tissues and nearly 1,000 individuals, was obtained from the GTEx v8 release[191] and used to predict ncRNA-targets using Allelome.LINK.

b, Boxplot displaying the distribution of allele-specific loci per individual and tissue. The color code follows the GTEx color scheme for each tissue, and the tissue abbreviations are consistent with those used in the GTEx resource[190]. An overview of the GTEx tissue abbreviations is provided in Appendix 10.4.

c, Scatter plot illustrating the number of informative loci with allele-specific expression (ASE, allelic bias ≥ 0.7 or ≤ 0.3, total *n* = 193,327) and the number of individuals. Different colors indicate different tissues, consistent with the color coding used in panel (**b**).

d, Boxplot showing the enrichment of allele-specific ncRNAs in proximity to allele-specific (dark gray) and biallelic (light gray) protein-coding genes within ±100kb distance. Boxplots summarize samples per tissue. The significance level is indicated by the number of asterisks and was determined by Wilcoxon tests. Boxplots show the interquartile range, median, and whiskers range from maximum to minimum. ASE: allele-specific expression, BAE: biallelic expression. The figure was sourced from a submitted manuscript by the author of this thesis (see 9.2 Submitted Manuscripts, 1.).

Allelome.LINK was subsequently applied to all samples of the GTEx database. Due to the outbred nature of the human population, each individual possesses a unique set of genetic variants, resulting in a personalized allelic landscape. This diversity allows novel linkages to be identified with each sample. Although linkages can only be detected in samples where the ncRNA exhibits ASE, the regulatory relationship is expected to be common across samples (**Figure 5.21a**).

Each tissue sample revealed an average of one linkage per individual (**Figure 5.21b**). However, a significant proportion of these linkages were specific to individual samples (63.77%) rather than shared between individuals (36.23%, **Figure 5.21c**). This observation suggests that each individual contributes to the discovery of novel linkages. Notably, no saturation of novel linkages was observed in any tissue as the sample size increased (**Figure 5.21d**). Thus, the genetic variation present in humans provides substantial potential for discovering a large number of linkages as the sample sizes increase.

On average, 42.43 ncRNA-to-target linkages were identified per tissue in the human dataset, with the highest number of linkages observed in the Skin - Sun Exposed (lower leg, *n* = 95) and the Thyroid (*n* = 95), while the Kidney - Medulla showed the lowest number of ncRNA-to-target linkages, likely due to the small sample size (*n* = 4, **Figure 5.21e**). Next, the distribution of these ncRNA-to-target linkages was analyzed across all tissues and identified 530 unique linkages. Among these, 233 linkages (43.96%) were tissue-specific, while 297 linkages (56.04%) were shared by two or more sampling sites (**Figure 5.21f**). Notably, 80.47% (*n* = 239) of the shared linkages were present in fewer than 10 tissues (**Figure 5.21f-g**). Linkages identified in more than 38 tissues belong exclusively to the human leukocyte antigen (HLA) cluster, genes known to show high genetic variability leading to ASE[330]. The finding of a higher frequency of tissue-specific linkages compared to shared linkages across tissues aligns with the observations in mice and highlights the tissue-specific nature of ncRNAs.

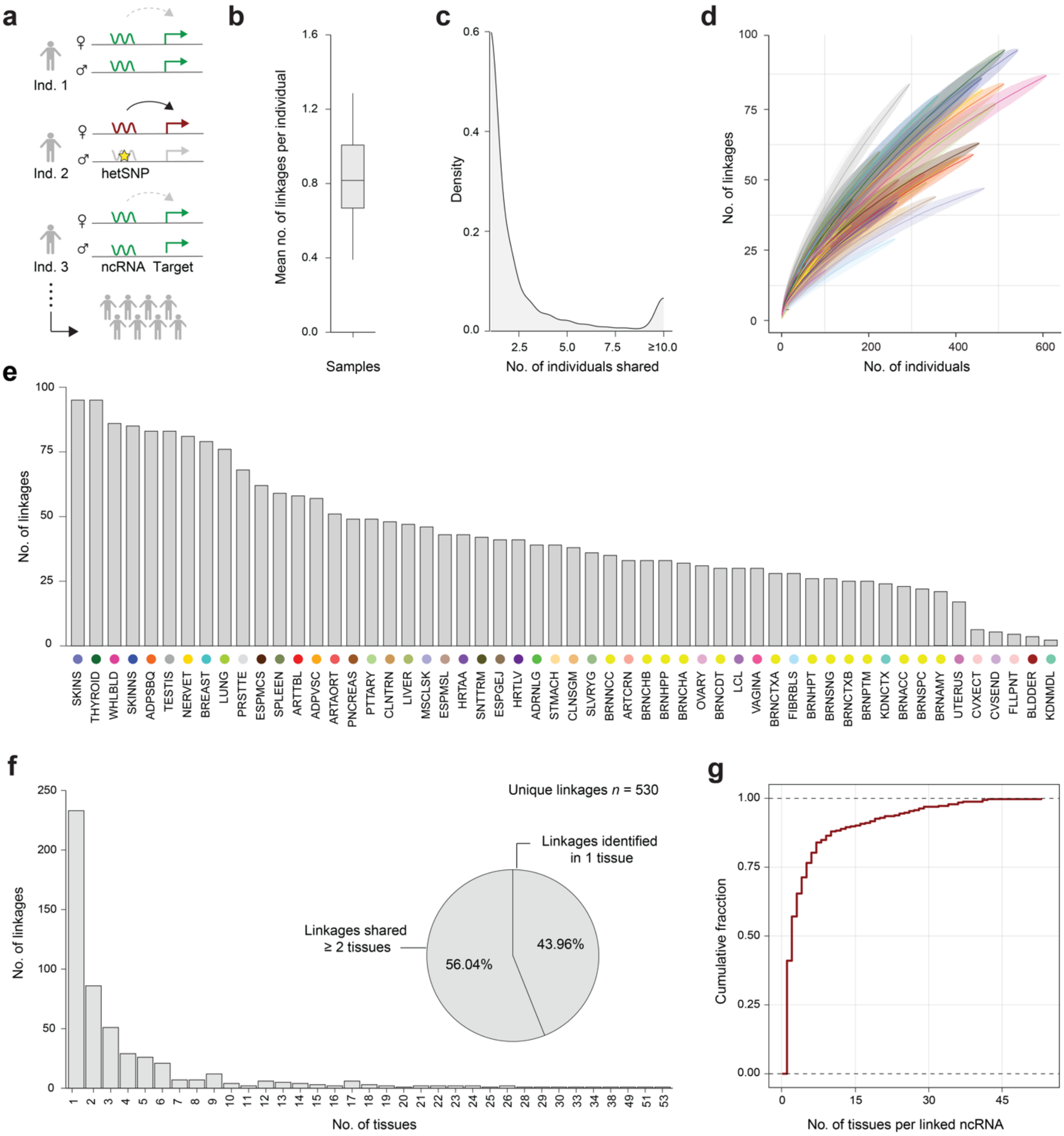


Figure 5.21 Properties of ncRNA-to-target linkages of outbred human samples.

a, Overview of the Allelome.LINK framework for outbred samples with different genotypes. The ncRNA-to-target interaction is detected in individual 2 due to a heterozygous SNP (hetSNP) that results in allele-specific expression of the ncRNA. Although this regulatory association is expected to be present across all samples, it is often masked by the biallelic expression of the ncRNA.

b, Boxplot demonstrating the mean number of linkages per individual across tissues. The interquartile range around the median is shown. Whiskers range from maximum to minimum.

c, Density plot displaying the distribution of linkages shared across different numbers of individuals.

d, Saturation curve representing the average number of linkages relative to the number of individuals. Lines display mean values and shaded areas the standard deviations calculated from random sampling (iterations *n* = 1,000). Different colors correspond to various tissues.

e, Bar plot showing the number of identified linkages per tissue. The color codes and abbreviations correspond to the respective tissues, consistent with the GTEx database[190].

f, Bar plot showing the number of linkages shared in a given number of sampling sites. The pie chart illustrates the fraction of linkages that are tissue-specific and present in ≥ 2 tissues.
g, Cumulative fraction plot of the number of tissues in which a ncRNA was linked. The figure was modified from a submitted manuscript by the author of this thesis (see 9.2 Submitted Manuscripts, 1.).

In summary, 2,291 linkages were identified across all tissues, representing 17.75% of the informative ncRNAs (*n* = 324) that were successfully assigned to their targets (**Figure 5.22**). The entire resource, including all linkages, is available via URL links listed in https://github.com/AndergassenLab/Allelome.LINK. A detailed explanation for using the generated resource can be retrieved from **Appendix section 10.3**.

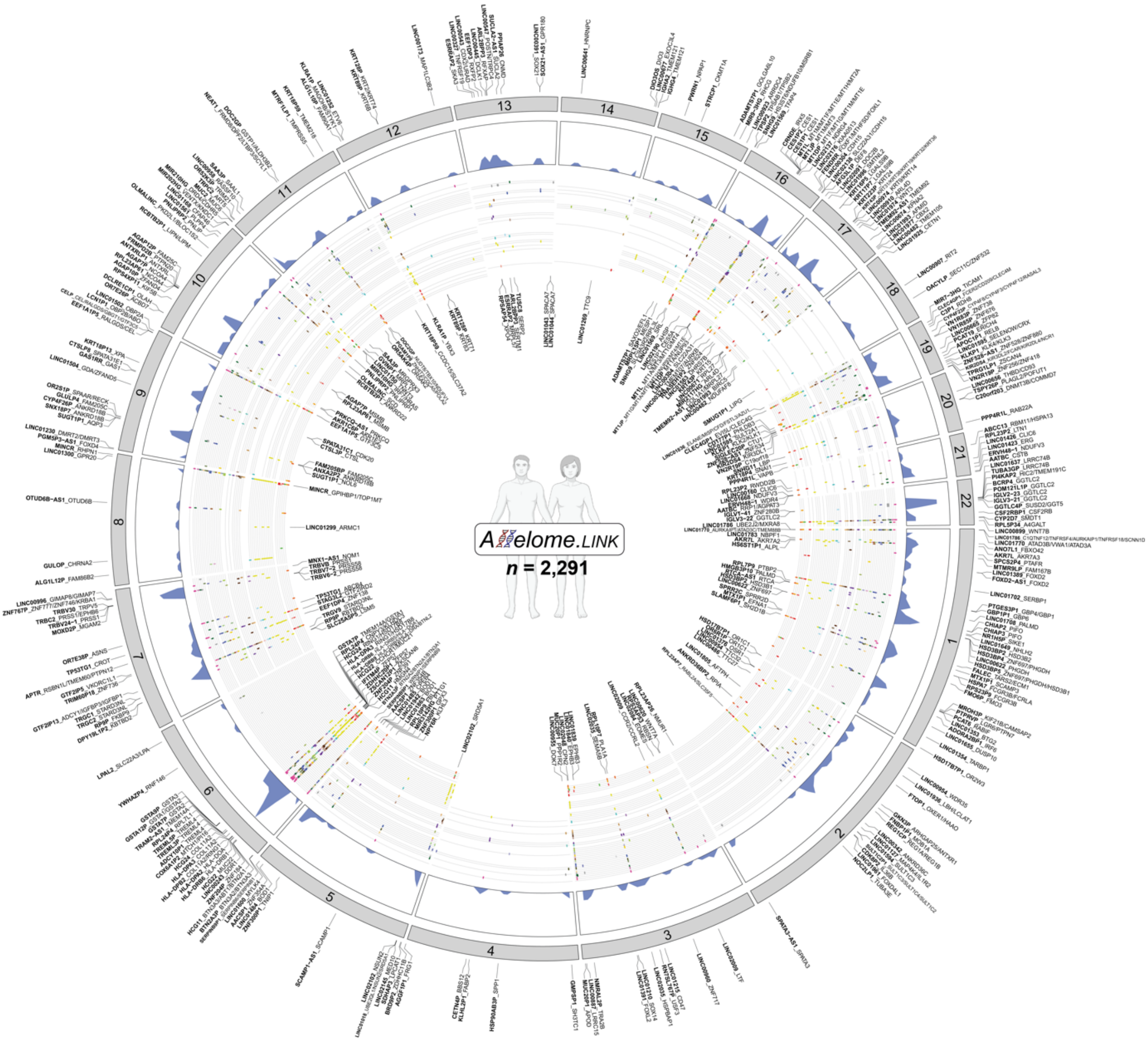


Figure 5.22 Comprehensive ncRNA-to-target resource for 54 different human tissues.
Chord plot showing the predicted ncRNA-to-target linkages identified across 54 human tissues for chromosome 1-22. The outer labels indicate linkages that were more frequently detected as enhancing interactions, while the inner labels depict predominantly repressive linkages. Bold font highlights ncRNAs. Dashes separate multiple protein-coding targets. The density plot indicates the

genomic distribution of linkages per chromosome. Individual tracks and colors denote the tissues in which a linkage was detected. The figure was obtained from a submitted manuscript by the author of this thesis (see 9.2 Submitted Manuscripts, 1.).

5.2.5 Most linkages identified in humans are supported by eQTL data

Sample-matched eQTL data from the GTEx database was used to validate the predicted ncRNA-to-target linkages in human tissues. This dataset contains 21,412,255 fine-mapped eQTLs across 49 of the 54 tissues analyzed[318]. eQTL data was unavailable for tissue samples of the Kidney - Medulla, Fallopian Tube, Cervix - Endocervix, Cervix - Ectocervix, and Bladder. The validation rates ranged up to 100% for small sample groups such as the Bladder (**Figure 5.23a**). On average, 77.47% (standard deviation = 9.83) of the linkages were confirmed by the eQTL data, of which 18.72% were specifically validated by eQTLs from the same tissue type (**Figure 5.23a**). This substantial validation rate underscores the robustness and efficacy of the allele-specific approach for ncRNA-target prediction, providing valuable insights into the regulatory landscape of the non-coding genome.

In addition, the accuracy of predicted regulatory mechanisms (enhancing or repressive) was assessed by evaluating the mechanism assignments for linkages shared by a large proportion of individuals. This approach helped to determine the consistency of enhancing and repressive mechanism assignments. The mechanism prediction was first tested for linkages within the HLA cluster. Given the high variability in ASE among HLA genes[330], a random distribution of mechanisms was anticipated across individuals, which was tested for samples of the Whole Blood. Indeed, 49.25% of the linkages were identified as enhancing and 50.75% as repressive for a total of 361 HLA gene interactions (**Figure 5.23b**). Moreover, Allelome.LINK was applied to samples of the Heart - Left Ventricle (*n* = 386), Pancreas (*n* = 305), Adrenal gland (*n* = 233), and Muscle - Skeletal (*n* = 706), where the imprinted status of *DLK1* and *MEG3* was previously confirmed by allelic expression data[151]. In a mouse embryonic stem cell system, the lncRNA *Meg3* was shown to repress *Dlk1* expression in *cis*[331]. Remarkably, the interaction was consistently identified as repressive in 79.46% of the 564 individuals where the linkage was detected (**Figure 5.23c**). This result highlights the reliability of the Allelome.LINK framework in accurately assigning the regulatory mechanisms across individuals.

In conclusion, the high validation rate of predicted ncRNA-to-target interactions and the accuracy in mechanism assignment underscore the effectiveness of the Allelome.LINK framework in identifying regulatory ncRNA-to-target interactions.

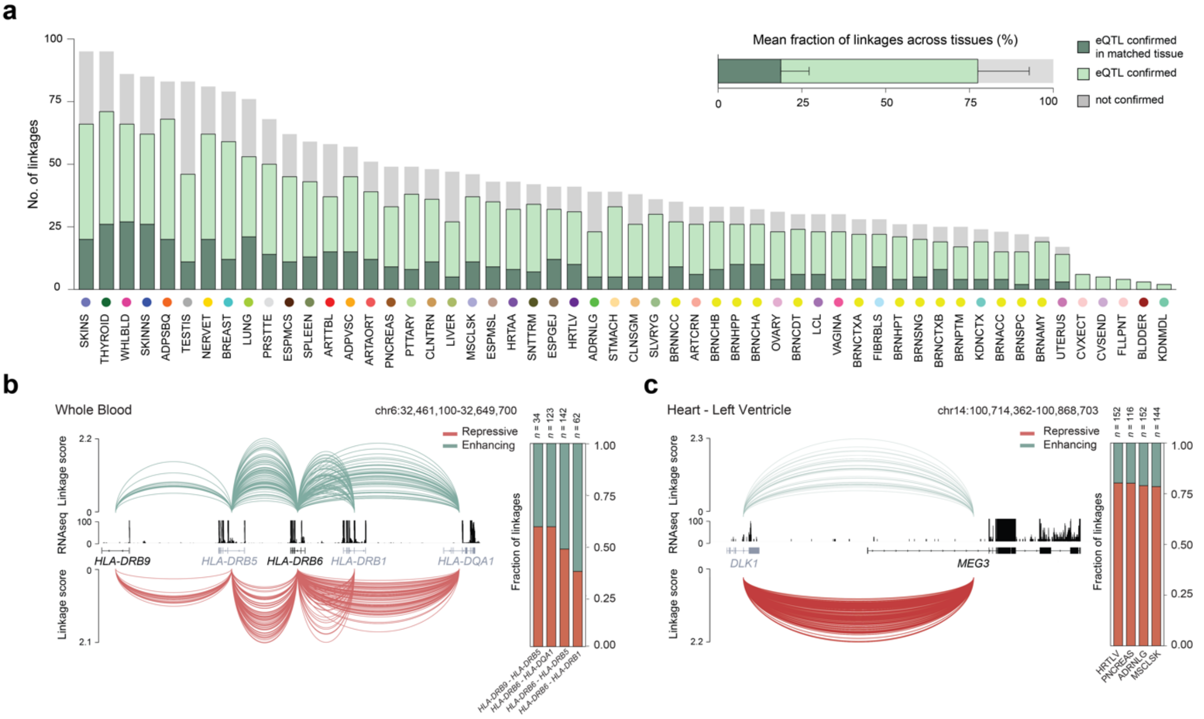


Figure 5.23 Validation of the ncRNA-to-target linkages and the mechanism assignment.
a, Bar plot showing the number of validated linkages confirmed by eQTL analysis across various tissues. Each bar represents the number of validated linkages for a specific tissue. The horizontal bar plot shows the mean fraction and standard deviation of all linkages summarized across tissues. Linkages confirmed by eQTLs from the same tissue are depicted in dark green, while those validated by eQTLs from different tissues are shown in light green.
b, Genome browser visualization of the Allelome.LINK output for Whole Blood samples (*n* = 611). Shown are the linkage predictions between the genes: *HLA-DRB9*, *HLA-DRB5*, *HLA-DRB6*, *HLA-DRB1,* and *HLA-DQA1*. Enhancing interactions are represented by green arcs (*n* = 178), and repressive interactions are depicted by red arcs (*n* = 183), with arc height proportional to the linkage score. Black gene names highlight ncRNAs, while gray color indicates protein-coding genes. The RNA-seq track shows gene expression levels of a representative tissue sample from the GTEx database. The bar plot shows the proportion of enhancing (green) and repressive (red) linkages for each interaction.
c, Genome browser visualization of the Allelome.LINK output as in (**b**) but for the imprinted interaction between *MEG3* and *DLK1* of Heart - Left Ventricle samples. The bar chart displays the faction of enhancing (green) and repressive (red) *MEG3-DLK1* linkages for samples from Heart - Left Ventricle, Pancreas, Adrenal gland, and Muscle - Skeletal. The figure was taken from a submitted manuscript by the author of this thesis (see 9.2 Submitted Manuscripts, 1.).

5.2.6 Elucidating high-confident ncRNA-targets by assessing the mechanism assignment

The mechanism fraction for each linkage was determined by calculating the proportion of individuals within each tissue sample where a given linkage was classified as either enhancing or repressive. This approach allowed the evaluation of the consistency of mechanism assignments across samples within each tissue type (**Figure 24a**). A consistency threshold of

≥ 75% was applied to filter for high-confident linkages, meaning that a linkage had to be classified as enhancing or repressive in at least 75% of the tissue samples where it was detected. This process resulted in 35.0% repressive (*n* = 802) and 48.1% enhancing (*n* = 1102) linkages. Notably, of the 16.89% (*n* = 387) of linkages with random mechanism assignment, 36.69% involved the HLA genes (**Figure 5.24a**).

While linkages identified in single individuals provide valuable regulatory insights, those observed consistently across multiple samples with identical mechanisms represent more robust findings due to their reproducibility. Therefore, linkages with consistent mechanism predictions in more than 10 individuals were filtered, uncovering 24 high-confident linkages (repressive *n* = 9, enhancing *n* = 15, **Figure 5.24a**).

One of these high-confident linkages is the interaction between the lncRNA *FENDRR* and the pcGene *FOXF1*. In line with previous studies, a regulatory interaction between these gene loci was identified across multiple samples[332,333]. This linkage was consistently classified as enhancing in 94.55% (*n* = 52) of the *FENDRR-FOXF1* linkages detected in Cells - Cultured Fibroblast samples (**Figure 5.24a-b**). Another example of a high-confident linkage was the positive regulatory relationship of the pcGene *TREML4* with the ncRNAs *TREML3P* and *TREML5P* in Whole Blood samples. These interactions were consistently identified as enhancing in 117 (*TREML3P* - *TREML4*) and 103 (*TREML5P* - *TREML4*) individuals (**Figure 5.24a-b**). As repressive examples, the linkages between the ncRNA *SERPINB9P1* and the predicted target *SERPINB9* in the Esophagus - Mucosa (*n* = 21), as well as the ncRNA *MIR2117HG* and the pcGene *ARL4D* in the Skin - Sun-Exposed Lower Leg (*n* = 65), were highlighted. Both linkages were identified in a large subset of individuals and showed consistent mechanism assignment (**Figure 5.24a-b**).

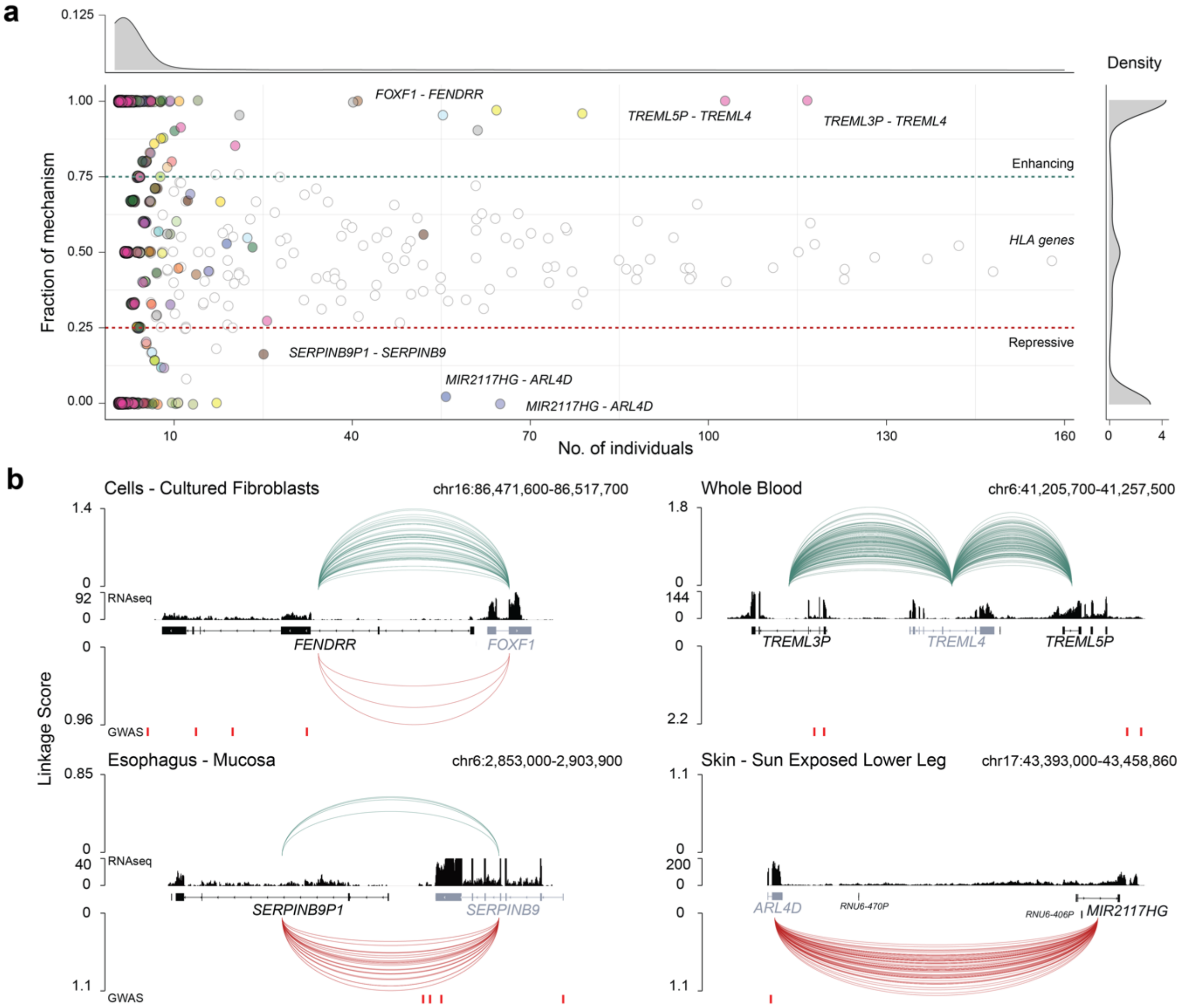


Figure 5.24 Analysis of linkage mechanisms and high-confidence examples.
a, Scatter plot illustrating the proportion of enhancing and repressive mechanisms for each linkage relative to the number of individuals exhibiting that linkage. A value of 1 means that the linkage is enhancing in all individuals where it was detected, while a value of 0 means that the linkage is repressive in all individuals. Transparent dots represent linkages involving genes from the *HLA* gene cluster. Colored dots represent the different tissues. The upper density plot shows the distribution of linkages across individuals per tissue, and the right density plot depicts the distribution of linkage mechanisms, ranging from 100% repressive to 100% enhancing.
b, Examples of high-confident linkages. Green arcs denote enhancing interactions, while red arcs indicate repressive interactions, with arc height proportional to the linkage score. Gene names in black denote ncRNAs, while gray names indicate protein-coding genes. The RNA-seq track displays gene expression levels for a representative example of that tissue from the GTEx database. Red lines mark GWAS hits sourced from the GWAS catalog[334]. The figure was retrieved from a submitted manuscript by the author of this thesis (see 9.2 Submitted Manuscripts, 1.).

5.2.7 Assignment of ncRNA-overlapping GWAS variants to their protein-coding targets via identified linkages

As a final step, the generated ncRNA-to-target linkages were leveraged to assign GWAS variants in the non-coding genome to their potential protein-coding targets. To accomplish this, all GWAS variants present in the NHGRI-EBI catalog were downloaded and overlapped with the informative ncRNA loci (*n* = 1,059 variants)[334]. The number of linkages with GWAS variants varied between tissues, ranging from 1 in the Kidney - Medulla to 42 in the Skin - Sun Exposed Lower leg (median *n* = 15, **Figure 5.25a**). Overall, 36.73% (*n* = 119) of the linked ncRNAs were associated with at least one GWAS variant (**Figure 5.25b**).

Notably, using the Allelome.LINK resource, a fraction of 30.59% of the informative non-coding GWAS variants were assigned to a pcGene, allowing researchers to gain important insights into their potential functional consequences (*n* = 324, **Figure 5.25b**). All candidate linkages, including their associated GWAS variants, are available via URL links listed in https://github.com/AndergassenLab/Allelome.LINK, providing a valuable resource for further exploration by the research community (**Appendix 10.3**). This extensive resource contributes to the understanding of the functional role of trait- and disease-related variants within the non-coding genome. As the availability of GWAS datasets and sequencing information continues to grow, this approach will continue to decode the target genes of non-coding risk variants.

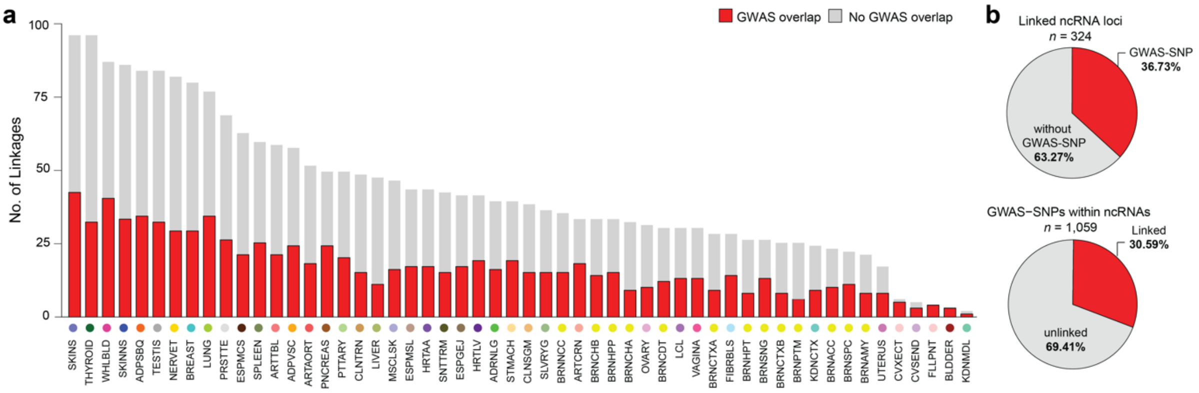


Figure 5.25 Linking non-coding GWAS variants to their potential protein-coding targets.
a, Bar plot displaying the number of linkages identified per tissue. Red bars indicate the number of linkages where the ncRNA harbors a GWAS variant.
b, The upper pie chart shows the fraction of ncRNAs that could be linked (*n* = 324) and contain a GWAS-SNP in the ncRNA (red) versus the fraction of linked ncRNAs without a GWAS variant (gray). The lower pie chart depicts the percentage of GWAS variants overlapping informative ncRNAs (*n* = 1,059), separated by linked (red) and unlinked (gray) ncRNAs. The figure was obtained from a submitted manuscript by the author of this thesis (see 9.2 Submitted Manuscripts, 1.).

6 Discussion

6.1 Project 1: Investigating the *in vivo* contribution of the *Crossfirre* locus alone and in combination with *Firre* and *Dxz4*

The present study investigated the *in vivo* contribution of the previously uncharacterized lncRNA *Crossfirre*, alone and in combination with *Firre* and *Dxz4*. Using one of the largest cohorts of genetically modified X-linked mouse models, combined with multi-omics approaches and extensive phenotyping, the project uncovered their functional role at the molecular and phenotypic level.

6.1.1 Deletion of the top female-specific loci has no effect on development and XCI biology

Over the past decade, several studies have demonstrated that the *Crossfirre-Firre* and *Dxz4* loci possess female-specific signatures of open chromatin that are absent in males[78,121,124,128,129]. However, it remained unclear whether such female-specific patterns occur frequently throughout the genome. This study addressed this question for the first time by systematically comparing the epigenetic landscape between males and females and found that this pattern represents the topmost female-specific accessible regions genome-wide. Despite the distinct female-specific pattern, the results of this thesis revealed that mutant mice lacking the *Crossfirre* locus, alone or in combination with *Firre* and *Dxz4*, exhibited no adverse effects on development or fertility, with the offspring being viable and showing the expected sex ratios and litter sizes. Similarly, genetic deletion models have previously demonstrated that the *Firre* and *Dxz4* loci are dispensable for mouse development, as the offspring of mutant mice were both viable and fertile[127,131]. These observations lead to the conclusion that, despite their highly female-specific epigenetic properties, *Crossfirre*, *Firre* and *Dxz4* are not essential for either development or fertility.

Moreover, the results of this study confirmed previous findings demonstrating that the female-specific chromatin accessibility pattern originates from the Xi[78,124,126]. Several studies have shown that these open chromatin regions correspond to Xi-specific CTCF binding sites, which contribute to the perinucleolar localization of the Xi[124,126]. Knockdown experiments of *Firre* resulted in the loss of the perinucleolar localization and lower H3K27me3 levels, highlighting a functional importance of this locus in maintaining both the Xi epigenetic landscape and nuclear positioning[124,126].

In addition to the female-specific features of the *Firre* and *Dxz4* loci, Andergassen *et al.* performed an extensive allele-specific analysis and identified *Crossfirre* as the only imprinted X-linked gene in somatic tissues, predominantly expressed from the maternal X chromosome[78]. This finding was observed in RNA-seq data from the brain, where *Crossfirre* is expressed at comparatively higher levels, and was confirmed in mouse embryonic fibroblasts by maternal H3K4me3 enrichment at the *Crossfirre* promoter[78]. Thus, it has been speculated that *Crossfirre* may act as a key genomic regulator of imprinted XCI, marking the maternal X chromosome. Prior to this study, the functional contribution of *Crossfirre* to XCI biology was entirely unknown. Despite its imprinted expression, the present study found that the deletion of *Crossfirre* has no effect on imprinted XCI, independent of whether the deletion was on Xa (maternal X chromosome) or Xi (paternal X chromosome), or whether it was deleted individually or in combination with *Firre* and *Dxz4*. Given the Xi-specific epigenetic signatures of *Firre* and *Dxz4*, along with their role in folding of the Barr body, a functional role for these loci in XCI has been hypothesized. Yet, several studies have shown that despite the loss of the Xi-structure, *Firre* and *Dxz4* are dispensable for XCI biology in cell culture models[121,125,126,128,129]. A knockout study by Andergassen *et al.* has further investigated the effects of *Firre* and *Dxz4 in vivo*[127]. Consistent with the results of this thesis, the authors did not observe a significant enrichment of differentially expressed genes on the X chromosome or changes in the ASE patterns of X-linked genes in the placenta, concluding that the deletion of *Firre* and *Dxz4* does not affect imprinted XCI *in vivo*.

The authors of the same study also investigated the effects of the loss of these loci on random XCI by analyzing XCI skewing ratios in brain tissues of mutant mice[127]. In line with the findings of this thesis, no alterations in the XCI skewing were observed, suggesting that random XCI is not affected by the absence of *Firre* and *Dxz4*. However, the conclusive effects on random XCI have not been addressed at the gene-level. This gap arises from the complexity of studying XCI at the whole-organ level, in particular due to the random XCI status where cells alternate between having the maternal or paternal X chromosome active. By developing a novel approach that includes allele-specific single-cell sorting, this study addressed the challenge and investigated the consequences of the deletion of the Xa-specific lncRNA expression and the Xi-specific epigenetic characteristics on random XCI *in vivo*. This analysis conclusively demonstrated that the deletions of *Crossfirre, Firre*, and *Dxz4* do not affect random XCI maintenance in the adult spleen.

The *Crossfirre* knockout further marked the first deletion on an X-linked LINE cluster attached to the *Crossfirre* locus. The deletion of this element was of particular interest as LINE clusters have been suggested to prevent local gene escape and contribute to the maintenance of

XCI[141,327]. The lack of effect on the expression of neighboring genes further challenges the presumed role of LINE clusters in XCI stability, suggesting that LINE clusters may be dispensable for the maintenance of XCI.

In conclusion, the comprehensive allele-specific analyses of this study demonstrate that *Crossfirre*, *Firre*, and *Dxz4* are dispensable for both imprinted and random XCI *in vivo*. These findings provide important insights into long-standing questions in the field and address previous assumptions about the role of these loci in XCI biology.

6.1.2 *Crossfirre* and *Firre* combined have a synergistic role in autosomal gene regulation

In contrast to the dispensability of these loci in XCI biology, transcriptomic analysis of the TKO mouse models revealed large-scale autosomal gene dysregulation, a finding that was common across organs except for the brain.

Previously, *Firre* and *Dxz4* have been reported to play a functional role in autosomal gene regulation in an organ-specific manner[127,131]. In mice lacking these loci, spleen tissue exhibited autosomal gene dysregulation associated with chromosome segregation and structure gene sets. A single deletion model of *Firre* confirmed this locus as the primary driver of these effects[127]. Interestingly, the additional knockout of *Crossfirre* to the *Firre-Dxz4* double deletion in the present study resulted in a more than 11.4-fold increase in the number of dysregulated genes. These findings suggest a significant effect of *Crossfirre* in addition to *Firre* and *Dxz4* on autosomal gene regulation. Comparisons of multiple knockout strains, including single and double deletions, further provide evidence that the combined knockout of *Crossfirre* and *Firre* contributes to the autosomal gene dysregulation. These findings indicate a synergistic role for *Crossfirre* and *Firre* in autosomal gene regulation. Synergistic effects between genes are known as epistasis or gene-gene interactions and describe a functional relationship of genes[335]. An example for synergistic gene induction is provided by Goldstein *et al.*, where the authors show that STAT3 binding to the target sites is enhanced by NF-κB. In this process, the activity of NF-κB primes enhancers, facilitating STAT3 binding to chromatin and driving synergistic gene expression[336]. However, the mechanism by which *Crossfirre* and *Firre* may interact and affect autosomal gene expression is unknown, representing a significant limitation of the study that requires further investigation.

To regulate genes on autosomes, *Crossfirre* and *Firre* would have to function in *trans*. This assumption aligns with the proposed mechanism for *Firre*, which serves as a platform for *trans*-chromosomal interactions, bringing together at least three gene loci located on different chromosomes around the transcription site of *Firre*[72]. Additional evidence for a *trans*-acting *Firre* RNA comes from a study by Lewandowski *et al.*[131]. The authors showed that *Firre*-mediated hematopoietic defects in knockout mice could be rescued by transgenic expression of *Firre*[131]. A *trans*-acting role for *Firre* has also been described in Patski cell lines, where Fang *et al.* found that *Firre* RNA transcribed from the Xa preserves H3K27me3 enrichment on the Xi[126]. However, further investigation is needed to understand how *Crossfirre* contributes to these findings. The same study observed that the knockdown of *Firre* leads to the upregulation of *Crossfirre*, suggesting that *Firre* represses the antisense transcript[126]. In contrast, the present study found that the deletion of *Crossfirre* does not exert regulatory control over the expression of *Firre*, indicating that *Crossfirre* does not play a significant role in controlling *Firre* expression. Given the complexity of lncRNA-mediated gene regulation, the potential mechanisms by which both loci affect autosomal gene expression are diverse and could further include indirect *trans*-regulatory effects, such as those mediated by interactions with RBPs or small RNA pathways, including piRNA-driven gene regulation or miRNA sponging[40,337]. Future studies are required to elucidate the exact mechanisms underlying the observed gene dysregulation and whether these function via RNA or DNA elements.

6.1.3 Deletion of X-linked lncRNAs revealed knockout and sex-specific phenotypes

TKO mice further underwent an extensive phenotypic screening at the GMC to elucidate the phenotypic consequences upon deleting these gene loci. The GMC provides standardized and unbiased phenotyping services for mouse mutant lines, assessing over 550 disease-relevant parameters[207,324]. Gene loci, especially lncRNAs, often exert pleiotropic effects with different functions during developmental stages or across different tissues[207]. Thus, large-scale phenotyping of knockout mouse models is essential to robustly detect phenotypic consequences[207].

The phenotypic characterization of mouse models lacking *Crossfirre*, *Firre*, and *Dxz4* uncovered genotype effects on traits related to immunology, behavior, clinical chemistry, dysmorphology, metabolism, and pathology. These findings support a functional role of these loci in various biological processes. Transcriptomic analysis of TKO mice identified an upregulation of mitochondrial and ribosomal gene sets, suggesting implications in energy metabolism[338,339]. This is consistent with several phenotypes detected by the GMC, including lower plasma cholesterol concentrations and urea levels, phenotypic traits associated with shifts in protein metabolism[340]. Moreover, the observed upregulation of mitochondrial and ribosomal gene sets is supported by the finding that TKO mutants showed higher levels of triglycerides, as well as lower creatinine levels and lactate concentrations[341-343]. Interestingly, the decrease in lactate was one of the most pronounced effects observed by the GMC. Lactate has been reported to have important functional roles in cellular metabolism, serving as a product of the glycolysis pathway and substate for mitochondrial respiration[341]. Apart these functions, lactate has further gained recognition as signaling molecule between different tissues and organs to facilitate metabolic adaptation in response to changing conditions[344]. Combined, the increase in mitochondrial and ribosomal activity, along with the identified metabolic phenotypes support alterations in the energy metabolism of mutant mice. The origin of these metabolic shifts remains unknown, but may arise in response to altered cellular processes or metabolic demands[338,339].

Beyond energy metabolism, analysis of the peripheral blood showed apparent differences in monocyte and eosinophil proportions, as well as B cell and CD4/CD8 T cell ratios. These findings are in agreement with the previously described molecular function of *Firre* in hematopoiesis, where Lewandowski *et al.* reported alterations in the blood cell composition upon the deletion of *Firre*[131]. The same study further showed that *Firre*-overexpressing mice had elevated levels of proinflammatory cytokines and decreased survival when exposed to lipopolysaccharide[131]. Interestingly, phenotyping of the TKO males also revealed increased levels of the proinflammatory cytokine IL-6, confirming a link between the TKO and immune response. These findings are further supported by studies of human *FIRRE* that demonstrate a feedback loop between *FIRRE* and the NF-κB signaling pathway[132,345]. Combined with the results of the thesis, a conserved functional role for *Firre/FIRRE* in regulating inflammatory responses in both mice and humans is indicated.

Interestingly, male mutants further displayed altered behavior and deficits characteristic for sensory processing disorders, such as impaired hearing[346]. These findings may be supported by two human case studies that report male children with a *FIRRE* locus duplication[136,137]. The patients showed mild to severe intellectual disability with clinical phenotypes of neurodevelopmental delay associated to the genetic alteration[136,137]. One of the human studies further observed cardiac abnormalities that may correspond with the subtle cardiovascular phenotypes observed in female mice, including higher heart rates and heart-rate-corrected QT intervals, as well as signs of dilated cardiomyopathy[136]. Although these findings need to be further investigated in future studies, the results support the disease relevance of *Firre* and a possible conservation of its pathological effects across species.

Among all phenotypes identified in the TKO mice, a significant number was specific to sex. The presence of a high number of sex-specific phenotypes indicates potential sex-specific functions for *Crossfirre*, *Firre*, and *Dxz4*. This is in line with the findings, that these loci were identified as the topmost female-specific chromatin accessibility regions. However, it is important to note that sexual dimorphisms between males and females can influence the prevalence of phenotypic traits. A study by Karp *et al.*, which analyzed 234 traits across 40,192 mutant mice, revealed that a large proportion of phenotypes are impacted by sex[347]. Consequently, it remains to be investigated whether the observed sex-specific TKO phenotypes result from direct sex-specific functions of *Crossfirre*, *Firre*, and *Dxz4* or manifest as a result of the sexual dimorphism between males and females.

Considering that all three lncRNAs are located on the X chromosome, phenotypes observed independently of sex or in male mutant mice suggest a functional role of the RNA of these loci from the Xa. In contrast, the involvement of Xi-specific epigenetic signatures cannot be excluded for female-specific phenotypes. Determining how these sex-specific loci function differently in males and females remains a subject for further investigation. Evidence supporting the differential effects of these loci based on whether the deletion occurs on Xa or Xi comes from the scRNA-seq analysis of the spleen. Consistent with the immunological findings from the phenotyping screen, a significant reduction in B cell proportions was observed in cells lacking *Crossfirre*, *Firre*, and *Dxz4* expression when the deletion was present on Xa. Conversely, cells with the deletion on Xi maintained normal B cell proportions but showed a significant reduction in CD4 T cells. These results underscore the distinct functional roles of these loci depending on whether they are located on Xa or Xi, potentially explaining the emergence of sex-specific phenotypes.

6.1.4 Outlook

The multi-omics characterization of different mutant mouse models lacking *Crossfirre*, *Firre*, and *Dxz4*, as well as the comprehensive phenotypic analysis, lays a robust foundation for future studies investigating the interplay of these X-linked loci and their role in autosomal gene regulation. However, further research is needed to dissect the specific functional implications of the synergistic gene regulation of *Crossfirre-Firre* and the pleiotropic phenotypic effects observed in the TKO mutants.

One limitation of the current study is that the impact of the lack of these loci on XCI was only investigated at one developmental timepoint (E12.5) and in multiple adult organs of the same age (6-weeks). Therefore, effects on the maintenance of Xi repression during aging cannot be excluded. Additionally, imprinted genes are often functional during embryonic development and can have subtle effects during the transient embryonic growth period, which can be compensated until birth[348]. Although these experiments were beyond the scope of the current study, they may warrant investigation in the future.

In addition, the present study did not investigate the regulatory mechanism through which *Crossfirre* and *Firre* influence autosomal gene expression. Given the complexity of lncRNA-mediated gene regulation, the potential mechanisms are diverse and could include direct or indirect *trans*-regulatory pathways. Therefore, further studies are needed to determine the underlying molecular mechanisms and whether these function via RNA and/or DNA elements.

This study further found sex-specific effects of *Crossfirre*, *Firre*, and *Dxz4* with different phenotypic consequences in males and females. While the Xi-specific characteristics of these loci are specific to females, a higher frequency of phenotypes was found in males. This raises the question of the distinct functional properties of these loci between the sexes. Further investigation is needed to unravel the functional properties that mediate the sex-specific functions and to elucidate the different functional mechanisms that can arise from Xi and Xa.

Finally, several findings from the study point towards the disease relevance of the TKO model, including sensory processing disorders or subtle cardiac phenotypes. Interestingly, the dysregulation of *FIRRE* in humans has been linked to intellectual disability in male patients with a *FIRRE* locus duplication[136,137]. Together, these findings indicate a disease relevance of the *Firre* locus in both mice and humans. Further research is required to explore these implications, particularly by challenging the mouse mutants in disease models. Such models could help to delineate the molecular pathways affected by these loci and how their dysregulation contributes to disease phenotypes.

6.1.5 Summary

In summary, this project provides the first *in vivo* characterization of the previously uncharacterized lncRNA *Crossfirre*, both individually and in combination with *Firre* and *Dxz4*. By leveraging one of the largest cohorts of genetically modified X-linked mouse models, combined with multi-omics approaches and extensive phenotyping, the study uncovered their functional roles at both the molecular and phenotypic levels. Interestingly, the study identified these loci as the topmost female-specific chromatin accessibility regions. Despite the imprinting of *Crossfirre* and the unique female-specific characteristics, these X-linked loci were found to be dispensable for XCI biology. In contrast, the study identified that *Crossfirre* and *Firre* function synergistically in autosomal gene regulation, affecting mitochondrial and ribosomal pathways. Finally, mouse models lacking all three loci underwent a comprehensive phenotypic screening at the GMC, revealing knockout- and sex-specific effects, shedding light on the *in vivo* roles of *Crossfirre*, *Firre*, and *Dxz4*. The resulting dataset provides a solid basis for further studies exploring these X-linked loci.

6.2 Project 2: Decoding the targets and mechanisms of the non-coding genome through allele-specific genomics

Despite advances in ncRNA research, the understanding of their functional roles and regulatory targets remains incomplete. Experimental characterization of ncRNAs using mouse models is considered the gold-standard but is laboratory extensive. Thus, computational tools are needed to predict the target genes and mechanisms of ncRNAs a priori to facilitate the selection of candidates for experimental validation. So far, traditional approaches, such as genotype correlation studies, rely on large sample sizes and have been unable to identify a high number of ncRNA-targets due to their dynamic expression patterns. This project aimed to unravel the *cis*-regulatory ncRNAs of mice and humans by developing a novel bioinformatics framework to predict their targets and mechanisms, based on the ASE patterns.

6.2.1 Leveraging the allele-specific information to identify the regulatory targets of ncRNAs

Previously, Andergassen *et al.* performed comprehensive mapping of the allele-specific landscape across different mouse tissues and observed that the number of allele-specific ncRNAs and pcGenes correlated[78]. The present study confirmed this observation by systematically assessing the frequency of allele-specific gene loci, which identified that allele-specific ncRNAs are significantly enriched around allele-specific pcGenes in mice. To further validate these findings in humans, the GTEx database was used and revealed similar trends, supporting the hypothesis of the co-regulation between adjacent allele-specific loci. However, a significant enrichment was detected in only half of the human tissues tested. A possible explanation for several non-significant samples is that overlapping gene loci of the GTEx data were not considered. Due to the non-strand-specific nature of the data, overlapping gene loci had to be removed to avoid the detection of false positives. In mice, 15.18% of the linkages were found to overlap antisense transcripts. Consequently, the lack of strand-specificity of the GTEx data represents a significant limitation of the study and may have hindered the identification of a significant fraction of relevant *cis*-acting ncRNAs.

Over the past decades, the computational identification of ncRNA-targets has remained challenging, mainly due to the dynamic expression pattern of ncRNAs[33,42]. The developed Allelome.LINK approach leverages the allele-specific information, providing a promising framework to address this issue by comparing the expression levels of alleles within the same sample. This approach provides a highly controlled system that avoids compensatory effects or dynamic expression patterns. To date, ASE has been used to identify regulatory variants in the genomes of multiple organisms, including yeast, plants, mice, flies, wasps, birds, and humans[349]. However, rather than investigating the effects of genetic variants, this study examined the co-occurrence of ASE gene loci, providing insights beyond variant-specific effects. In addition to the Allelome.LINK approach, a recent study by Goede *et al*. also suggested that ASE patterns might be used to detect *cis*-regulatory interactions between genes[350]. The study identified local patterns of high ASE sharing between lncRNAs and nearby genes and identified 137 ASE sharing events. For these cases, the authors also implied potential *cis*-regulatory relationships between the nearby allele-specific genes[350]. However, the study examined ASE at the population level to identify associations that were consistently observed across the cohort. In contrast, Allelome.LINK performs individual-level analyses to capture ASE variability specific to particular samples. This approach provides a more nuanced understanding of ASE patterns and offers insights into personalized regulatory mechanisms that may be missed in population-wide summaries. Several studies have shown that allele-specific ncRNAs can induce ASE in their regulatory target genes[142,187,351]. Thus, investigating the ASE pattern of ncRNA and target further allow to infer the underlying mechanism of the regulatory relationship. An enhancing ncRNA that is higher expressed on one allele is expected to result in increased expression of the target gene on the same allele. In contrast, repressive effects should be reflected as anti-correlation between the alleles.

While one limitation of Allelome.LINK is that the detection of targets is restricted to *cis*-acting ncRNAs, this characteristic simplifies the process of identifying direct regulatory targets. Differential gene expression analysis following a gene knockout often reveals numerous dysregulated genes, making the interpretation of the results challenging[206,352]. These include all genes affected by the knockout in a direct and indirect manner. In contrast, in an allele-specific model, effects that occur in *trans* are masked, as these effects affect both alleles equally[353]. As a result, allele-specific models provide a robust framework for identifying the primary *cis*-interactions of ncRNAs and their targets. Additionally, one of the key advantages of the Allelome.LINK approach is the ability to identify regulatory associations based on transcriptomic data. This facilitates the candidate selection by providing prior knowledge of the specific tissues in which a ncRNA regulates a target gene.

6.2.2 Predicting ncRNA-targets in mice

The present study identified 397 ncRNA-to-target linkages in the major organs of F1 hybrid mice between the BL6 and CAST strains. The usage of F1 hybrids provides a robust approach for identifying gene loci with ASE, as the SNP information of the inbred strains are well defined[189]. Furthermore, F1 mouse hybrids allow the inclusion of biological replicates with the same genetic background. As a higher number of replicates has been shown to increase the power of ASE detection, the ability to use replicates increases the robustness of ASE detection for both ncRNAs and targets[354]. This finding was further confirmed by our study, highlighting that the highest precision and recall for Allelome.LINK were obtained for pooled samples.

The predicted interactions identified by Allelome.LINK are based on statistical correlations between the ncRNAs and potential target genes. Therefore, these correlations do not necessarily imply causation. However, the significant enrichment of allele-specific ncRNAs nearby allele-specific pcGenes supports the likelihood of co-regulatory associations. It is important to note that the Allelome.LINK strategy was designed to predict regulatory interactions, providing a valuable foundation for further investigation. Future studies can build upon these findings and leverage this resource for candidate selection to perform functional assays and refine our understanding of regulatory ncRNAs.

The present study provided functional validation of Allelome.LINK by using a genetic knockout mouse model for the lncRNA *Airn*[78]. The *Airn* cluster has been shown to provide a powerful model for studying the regulatory targets of a lncRNA with ASE[78,79,148,174,175]. However, to further investigate the associations identified by Allelome.LINK, different knockout strategies have to be used in future experiments to validate the predicted regulatory relationships and explore their mechanisms beyond *cis*-acting effects. While transcriptomic data suggest RNA-based mechanisms, the possibility that regulatory DNA elements or the transcription of the ncRNA itself drive the regulatory effect on the target cannot be excluded[101,355]. Further studies are needed to disentangle these mechanisms, including whole-gene ablation to confirm regulatory relationships, followed by strategies such as polyadenylation-terminator insertion or promoter deletion to distinguish between DNA elements, transcription, and the ncRNA transcript[101]. Replacement of the gene body with reporter genes can further identify effects due to the act of transcription and promoter activity. *Trans* effects can be tested by rescue experiments following ncRNA deletion, while frameshift or start codon mutations can be used to investigate small functional peptides within ncRNAs[101]. Although Allelome.LINK identifies regulatory ncRNAs that act primarily in *cis*, it may be worthwhile to test for additional functional effects in *trans*, as regulatory effects can be affected by multiple *cis* and *trans* effects[356]. For example, the lncRNA *Tug1*, which is essential for male fertility, acts in *cis* by regulating neighboring genes via DNA elements and in *trans*, through RNA-based mechanisms[357]. These findings highlight the relevance of testing gene loci for multiple regulatory mechanisms.

In contrast to knockout studies, analyses of parent-hybrid trios with data collected from the parental lines and the F1 hybrid allow the influence of *cis*- over *trans*-regulatory effects to be inferred. Comparative analysis of parent-hybrid trios allows the identification of both regulatory relationships within F1 hybrids[358]. Sequencing of parental F0 strains offers the advantage of detecting regulatory variation by differential gene expression analysis. *Trans*-acting regulatory effects manifest as expression differences between the parental strains without ASE in the F1 hybrid, whereas *cis*-acting effects are reflected in both the allelic ratio of the F1 hybrids and in the differential expression between the parental strains[358]. This approach provides valuable insights into the regulatory mechanisms underlying gene expression variation.

The ability to replicate biological samples in mice further offers the advantage of studying F1 hybrids across various conditions, including disease models or different stages during development and aging. In outbred populations such as humans, genetic variation makes it challenging to distinguish genotype-driven effects from true context-specific interactions, as collecting samples from the same individual under different environmental contexts is often impractical. The use of genetically identical F1 hybrids eliminates this variability and facilitates the identification of context-dependent regulatory interactions. Thus, this strategy provides a robust model for studying regulatory shifts across different conditions.

The integration of multiple strains into the analytical framework could further be used to cross-validate existing regulatory linkages and to identify novel associations by mimicking outbred populations and genetic diversity. A recent study by Tsouris *et al.* employed a large diallel panel comprising 323 hybrid yeast strains to analyze the impact of different genetic variants on gene expression[358]. This concept could be translated to the allele-specific analysis, to pinpoint the causal variants leading to ASE in the identified regulatory ncRNAs. Comparing the ASE patterns of different F1 strains enables to classify strains with and without ASE for a particular locus. This allows the identification of genetic variants present in ASE-positive strains but absent in ASE-negative strains, leading to the identification of candidate variants. Subsequently, statistical approaches such as regression models can be applied to establish causal relationships between these variants and ASE. To gain further mechanistic insights, these variants could then be investigated using epigenetic data to determine whether these variants are located within open chromatin regions or known DNA regulatory sites[358]. Motif analysis can then assess whether these variants disrupt transcription factor binding, providing a detailed understanding of their functional impact. Thus, these analyses could elucidate the underlying molecular variants responsible for ASE.

6.2.3 Predicting ncRNA-targets in humans

After predicting ncRNA-targets in mice, the Allelome.LINK approach was applied to the GTEx database and identified 2,291 linkages across 54 different tissues and 838 human individuals.

Interestingly, most ncRNA-to-target associations were detected in single individuals. Unlike inbred mouse strains, the human population is outbred, exhibiting high genetic diversity. Thus, each individual possesses a unique set of genotypes. Considering that genetic variation can lead to ASE, the human population has the potential to reveal a vast array of ASE gene loci[142,176]. The observation that more ASE loci are identified as the number of individuals increases supports this notion. As a result, novel linkages were identified with each sample with no saturation observed in any tissue. Thus, the human genetic variation provides a substantial potential for the discovery of a large number of linkages. Although the regulatory effects can only be detected in individuals harboring the specific set of variants leading to ASE, the underlying associations are expected to be common across individuals.

To further evaluate the performance of the pipeline, sample-matched eQTL data were retrieved from the GTEx database and used to validate the predicted ncRNA-target linkages in human tissues[318]. Remarkably, 77.47% of the linkages were confirmed by the eQTL data across tissues, with 18.72% of the eQTLs originating from the same tissue as the predicted linkage. This substantial validation rate highlights the effectiveness of the allele-specific approach for predicting ncRNA-targets. In addition, the accuracy of the mechanism prediction was assessed by evaluating the mechanism assignments for the imprinted interaction between *DLK1* and *MEG3*. Testing the mechanism assignment for this imprinted interaction is particularly useful as the expected repressive relationship is independent of the DNA sequence[151]. Remarkably, the interaction was correctly identified as repressive in 79.46% of the samples tested, highlighting the reliability of the allele-specific framework in accurately assigning regulatory mechanisms. About 20% of the linkages were misclassified as enhancing, which could be due to errors in phasing. Phasing is a statistical method that assigns alleles to haplotypes based on linkage disequilibrium. However, as the distance between loci increases, phasing tends to become less accurate and more prone to errors[359]. Advances in phasing algorithms will further improve the mechanism assignment of the Allelome.LINK strategy.

Linkages that were identified in many samples and were consistently classified as enhancing or repressive were highlighted as high-confident linkages. While linkages identified in individual samples are expected to reflect common mechanisms, the high-confident linkages could be replicated across multiple samples and thus represent more robust findings due to their reproducibility. It is noted that the prevalence of a consistent ASE pattern may further indicate biological relevance[142]. Given that ASE is primarily driven by heterozygous variants, their widespread presence in a population implies that balancing selection favors the heterozygous state[360]. Consequently, these loci are of particular interest and warrant further investigation.

6.2.4 GWAS integration links non-coding variants to pcGenes

Over the past few decades, GWAS have significantly advanced our understanding of the genetic basis of disease, uncovering hundreds of thousands of risk variants[214]. However, the vast majority of these variants are located in the non-coding genome, making it difficult to interpret their functional implications[361]. Many risk variants identified by GWAS closely co-localize with regulatory regions, such as ncRNA loci, yet the limited functional understanding, including their regulatory targets and mechanisms, poses a significant challenge in understanding how these variants contribute to disease[362].

This study successfully predicted the protein-coding targets for 30.59% of the variants that overlapped informative ncRNA loci (*n* = 324). Although this number represents only a fraction of the total GWAS-identified SNPs, the future integration of more human sequencing datasets will continue to increase the number of non-coding risk variants that can be assigned to their respective protein-coding targets. To incorporate additional GWAS variants that do not overlap with linked ncRNAs, variants and ncRNA loci can be linked via eQTL co-localization[350]. If an eQTL associated with the expression of a particular ncRNA significantly co-localizes with a disease-associated variant, this SNP could be assigned to the ncRNA, which can be linked to the target gene via Allelome.LINK. This approach will further increase the number of non-coding GWAS variants linked to their protein-coding target genes.

All ncRNA-target predictions, including the inferred mechanisms and GWAS information, are accessible via intuitive genome browser visualization and can be accessed via URL links listed in the GitHub repository at https://github.com/AndergassenLab/Allelome.LINK. This dataset provides a valuable resource to select candidate ncRNAs for the tissue of interest. As the availability of sequencing data and risk variants continues to expand, our strategy offers promising avenues for future research.

6.2.5 Outlook

The present findings provide the basis for numerous future research projects that will help to elucidate the non-coding genome. The extensive resource, comprising ncRNA-target predictions for the major mouse organs and 54 different human tissues, serves as an ideal starting point for the community to select candidates for further investigation and validation. This resource will help to prioritize functional and disease-relevant ncRNAs that can be investigated using wet-lab experiments. The user-friendly nature of Allelome.LINK, combined with the growing pool of GWAS variants and sequencing data, will continue to identify novel regulatory interactions and shed light on the functional implications of the non-coding genome.

The highly dynamic expression pattern of ncRNAs also provides the opportunity to discover novel linkages in different tissues, cell types, or conditions. Thus, as more data is integrated, the pool of ncRNA-target predictions will continue to grow. Because bulk RNA-seq data comprises mixed cell types, ASE patterns deriving from different cell types could be masked[142]. To address this limitation, the updated Allelome.PRO v2.0 can be used on single-cell data, allowing Allelome.LINK to identify cell-type-specific interactions. Additionally, integrating data from different developmental stages or conditions, such as aging or disease, can reveal condition-specific ncRNA-target interactions. Interestingly, it is assumed that the presence of ASE increases with age[142]. By stratifying sequencing data based on condition, such as age or disease, this would not only allow the identification of novel linkages but could potentially highlight linkages or ASE patterns with biological relevance to a specific condition.

The vast amount of data offered by the GTEx database provides further opportunities. RNA-seq data can be used to calculate gene co-expression networks to assign linked ncRNAs and targets to pathways or cell types, providing additional insights into their functional roles[350]. Moreover, outlier enrichment analysis could be used to investigate the expression of ncRNA and target in the tissue and individual where the linkage was identified[363]. Due to the presence of ASE, the expression profiles of both interaction partners are likely to differ significantly from that of the remaining population[350]. Correlation of the outlier expression of ncRNA and target could validate the identified interaction in the respective individual and tissue and provide an approach to confirm linkages *in silico*. Finally, comparative analyses of genetic variants could reveal the specific variations responsible for the observed expression changes[350]. This could provide further insights into the genetics of ncRNA regulation and its potential relevance to disease.

It is worth noting that the Allelome.LINK approach can be extended by integrating sample-matched genomic data, such as those obtained from ATAC- or ChIP-sequencing. The additional integration of such sequencing data would allow the prediction of target genes associated with DNA regulatory elements, for example enhancer elements. This integration enables a comprehensive understanding of the regulatory mechanisms within the non-coding genome, providing insights into the complex network of interactions that control gene expression. Additionally, since many GWAS risk variants are located in DNA regulatory elements, including enhancers or repressors[364], these regulatory-target predictions can subsequently be applied to link disease variants that overlap DNA elements to their associated target genes. Thus, the Allelome.LINK strategy has the potential to unravel the intricacies of a majority of the *cis*-acting non-coding genome and its implications for complex diseases.

6.2.6 Summary

The present study found a significant enrichment of allele-specific ncRNAs nearby allele-specific pcGenes in both mice and humans, supporting the concept of co-regulatory associations. This discovery led to the development of Allelome.LINK, a novel bioinformatics framework that uses the allele-specific information to predict the target genes and mechanisms of *cis*-acting ncRNAs. Applying Allelome.LINK to the major mouse organs and human samples resulted in the identification of 397 mouse and 2,291 human ncRNA-target pairs and their predicted mode-of-actions. Following extensive validation, the integration of GWAS data allowed a substantial proportion of ncRNA-overlapping risk variants to be mapped to their respective protein-coding targets. With the increasing availability of sequencing data, this strategy has the potential to elucidate the targets and mechanisms of a majority of the *cis*-acting non-coding genome.

7 References

1 Pang, B., van Weerd, J. H., Hamoen, F. L. & Snyder, M. P. Identification of non-coding silencer elements and their regulation of gene expression. *Nat Rev Mol Cell Biol* 24, 383-395 (2023). https://doi.org/10.1038/s41580-022-00549-9

2 Nurk, S. *et al.* The complete sequence of a human genome. *Science* 376, 44-53 (2022). https://doi.org/10.1126/science.abj6987

3 ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74 (2012). https://doi.org/10.1038/nature11247

4 Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* 583, 699-710 (2020). https://doi.org/10.1038/s41586-020-2493-4

5 Kopp, F. & Mendell, J. T. Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell* 172, 393-407 (2018). https://doi.org/https://doi.org/10.1016/j.cell.2018.01.011

6 Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res* 22, 1748-1759 (2012). https://doi.org/10.1101/gr.136127.111

7 Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat Rev Genet* 17, 93-108 (2016). https://doi.org/10.1038/nrg.2015.17

8 ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 306, 636-640 (2004). https://doi.org/10.1126/science.1105136

9 Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* 507, 462-470 (2014). https://doi.org/10.1038/nature13182

10 Yuan, J. *et al.* A compendium of genetic variations associated with promoter usage across 49 human tissues. *Nature Communications* 15, 8758 (2024). https://doi.org/10.1038/s41467-024-53131-6

11 Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* 15, 272-286 (2014). https://doi.org/10.1038/nrg3682

12 Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* 20, 207-220 (2019). https://doi.org/10.1038/s41576-018-0089-8

13 Calo, E. & Wysocka, J. Modification of Enhancer Chromatin: What, How, and Why? *Molecular Cell* 49, 825-837 (2013). https://doi.org/https://doi.org/10.1016/j.molcel.2013.01.038

14 Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat Rev Genet* 19, 789-800 (2018). https://doi.org/10.1038/s41576-018-0060-8

15 Nemeth, K., Bayraktar, R., Ferracin, M. & Calin, G. A. Non-coding RNAs in disease: from mechanisms to therapeutics. *Nat Rev Genet* 25, 211-232 (2024). https://doi.org/10.1038/s41576-023-00662-1

16 Kapranov, P., Willingham, A. T. & Gingeras, T. R. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* 8, 413-423 (2007). https://doi.org/10.1038/nrg2083

17 Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet* 19, 535-548 (2018). https://doi.org/10.1038/s41576-018-0017-y

18 Poliseno, L., Lanza, M. & Pandolfi, P. P. Coding, or non-coding, that is the question. *Cell Res* 34, 609-629 (2024). https://doi.org/10.1038/s41422-024-00975-8

19 Villa, T., Pleiss, J. A. & Guthrie, C. Spliceosomal snRNAs: Mg<sup>2+</sup>-Dependent Chemistry at the Catalytic Core? *Cell* 109, 149-152 (2002). https://doi.org/10.1016/S0092-8674(02)00726-2

20 Jiao, L. *et al.* Ribosome biogenesis in disease: new players and therapeutic targets. *Signal Transduction and Targeted Therapy* 8, 15 (2023). https://doi.org/10.1038/s41392-022-01285-4

21 Coller, J. & Ignatova, Z. tRNA therapeutics for genetic diseases. *Nature Reviews Drug Discovery* 23, 108-125 (2024). https://doi.org/10.1038/s41573-023-00829-9

22 Pinzaru, A. M. & Tavazoie, S. F. Transfer RNAs as dynamic and critical regulators of cancer progression. *Nature Reviews Cancer* 23, 746-761 (2023). https://doi.org/10.1038/s41568-023-00611-4

23 Huang, Z.-h., Du, Y.-p., Wen, J.-t., Lu, B.-f. & Zhao, Y. snoRNAs: functions and mechanisms in biological processes, and roles in tumor pathophysiology. *Cell Death Discovery* 8, 259 (2022). https://doi.org/10.1038/s41420-022-01056-8

24 Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75, 843-854 (1993). https://doi.org/10.1016/0092-8674(93)90529-y

25 Good, D. J. Non-Coding RNAs in Human Health and Diseases. *Genes* 14 (2023).

26 Mizuno, T., Chou, M. Y. & Inouye, M. A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). *Proc Natl Acad Sci U S A* 81, 1966-1970 (1984). https://doi.org/10.1073/pnas.81.7.1966

27 Corcoran, C. P. *et al.* Superfolder GFP reporters validate diverse new mRNA targets of the classic porin regulator, MicF RNA. *Mol Microbiol* 84, 428-445 (2012). https://doi.org/10.1111/j.1365-2958.2012.08031.x

28 Chen, L.-L. & Kim, V. N. Small and long non-coding RNAs: Past, present, and future. *Cell* 187, 6451-6485 (2024). https://doi.org/10.1016/j.cell.2024.10.024

29 Kim, V. N., Han, J. & Siomi, M. C. Biogenesis of small RNAs in animals. *Nature Reviews Molecular Cell Biology* 10, 126-139 (2009). https://doi.org/10.1038/nrm2632

30 Shang, R., Lee, S., Senavirathne, G. & Lai, E. C. microRNAs in action: biogenesis, function and regulation. *Nat Rev Genet* 24, 816-833 (2023). https://doi.org/10.1038/s41576-023-00611-y

31 Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D. & Zamore, P. D. PIWI-interacting RNAs: small RNAs with big functions. *Nature Reviews Genetics* 20, 89-108 (2019). https://doi.org/10.1038/s41576-018-0073-3

32 Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* 309, 1559-1563 (2005). https://doi.org/10.1126/science.1112014

33 Mattick, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat Rev Mol Cell Biol* 24, 430-447 (2023). https://doi.org/10.1038/s41580-022-00566-8

34 Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916-919 (2002). https://doi.org/10.1126/science.1068597

35 Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563-573 (2002). https://doi.org/10.1038/nature01266

36 Rinn, J. L. & Chang, H. Y. Long Noncoding RNAs: Molecular Modalities to Organismal Functions. *Annu Rev Biochem* 89, 283-308 (2020). https://doi.org/10.1146/annurev-biochem-062917-012708

37 Fang, S. *et al.* NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res* 46, D308-d314 (2018). https://doi.org/10.1093/nar/gkx1107

38 Volders, P.-J. *et al.* LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Research* 47, D135-D139 (2019). https://doi.org/10.1093/nar/gky1031

39 Mattick, J. S. & Rinn, J. L. Discovery and annotation of long noncoding RNAs. *Nat Struct Mol Biol* 22, 5-7 (2015). https://doi.org/10.1038/nsmb.2942

40 Statello, L., Guo, C. J., Chen, L. L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol* 22, 96-118 (2021). https://doi.org/10.1038/s41580-020-00315-9

41 Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 22, 1775-1789 (2012). https://doi.org/10.1101/gr.132159.111

42 Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* 571, 510-514 (2019). https://doi.org/10.1038/s41586-019-1341-x

43 Lakhotia, S. C. Long non-coding RNAs coordinate cellular responses to stress. *Wiley Interdiscip Rev RNA* 3, 779-796 (2012). https://doi.org/10.1002/wrna.1135

44 Khan, M. R., Xiang, S., Song, Z. & Wu, M. The p53-inducible long noncoding RNA TRINGS protects cancer cells from necrosis under glucose starvation. *Embo j* 36, 3483-3500 (2017). https://doi.org/10.15252/embj.201696239

45 Connerty, P., Lock, R. B. & de Bock, C. E. Long Non-coding RNAs: Major Regulators of Cell Stress in Cancer. *Front Oncol* 10, 285 (2020). https://doi.org/10.3389/fonc.2020.00285

46 Liu, K. *et al.* Long non-coding RNAs regulate drug resistance in cancer. *Mol Cancer* 19, 54 (2020). https://doi.org/10.1186/s12943-020-01162-0

47 Huarte, M. *et al.* A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* 142, 409-419 (2010). https://doi.org/10.1016/j.cell.2010.06.040

48 Rothschild, G. *et al.* Noncoding RNA transcription alters chromosomal topology to promote isotype-specific class switch recombination. *Sci Immunol* 5 (2020). https://doi.org/10.1126/sciimmunol.aay5864

49 Fanucchi, S. *et al.* Immune genes are primed for robust transcription by proximal long noncoding RNAs located in nuclear compartments. *Nat Genet* 51, 138-150 (2019). https://doi.org/10.1038/s41588-018-0298-2

50 Vollmers, A. C. *et al.* A conserved long noncoding RNA, GAPLINC, modulates the immune response during endotoxic shock. *Proc Natl Acad Sci U S A* 118 (2021). https://doi.org/10.1073/pnas.2016648118

51 Atianand, M. K. *et al.* A Long Noncoding RNA lincRNA-EPS Acts as a Transcriptional Brake to Restrain Inflammation. *Cell* 165, 1672-1685 (2016). https://doi.org/10.1016/j.cell.2016.05.075

52 Zhang, P., Cao, L., Fan, P., Mei, Y. & Wu, M. LncRNA-MIF, a c-Myc-activated long non-coding RNA, suppresses glycolysis by promoting Fbxw7-mediated c-Myc degradation. *EMBO Rep* 17, 1204-1220 (2016). https://doi.org/10.15252/embr.201642067

53 Zheng, X. *et al.* LncRNA wires up Hippo and Hedgehog signaling to reprogramme glucose metabolism. *Embo j* 36, 3325-3335 (2017). https://doi.org/10.15252/embj.201797609

54 McClintock, M. A. *et al.* RNA-directed activation of cytoplasmic dynein-1 in reconstituted transport RNPs. *Elife* 7 (2018). https://doi.org/10.7554/eLife.36312

55 Ma, Y., Zhang, J., Wen, L. & Lin, A. Membrane-lipid associated lncRNA: A new regulator in cancer signaling. *Cancer Lett* 419, 27-29 (2018). https://doi.org/10.1016/j.canlet.2018.01.008

56 Wang, F. *et al.* The long noncoding RNA Synage regulates synapse stability and neuronal function in the cerebellum. *Cell Death Differ* 28, 2634-2650 (2021). https://doi.org/10.1038/s41418-021-00774-3

57 Samaddar, S. & Banerjee, S. Far from the nuclear crowd: Cytoplasmic lncRNA and their implications in synaptic plasticity and memory. *Neurobiol Learn Mem* 185, 107522 (2021). https://doi.org/10.1016/j.nlm.2021.107522

58 Wei, W. *et al.* ADRAM is an experience-dependent long noncoding RNA that drives fear extinction through a direct interaction with the chaperone protein 14-3-3. *Cell Rep* 38, 110546 (2022). https://doi.org/10.1016/j.celrep.2022.110546

59 Chen, L., Zhu, Q. H. & Kaufmann, K. Long non-coding RNAs in plants: emerging modulators of gene activity in development and stress responses. *Planta* 252, 92 (2020). https://doi.org/10.1007/s00425-020-03480-5

60 Wierzbicki, A. T., Blevins, T. & Swiezewski, S. Long Noncoding RNAs in Plants. *Annu Rev Plant Biol* 72, 245-271 (2021). https://doi.org/10.1146/annurev-arplant-093020-035446

61 Dueva, R. *et al.* Neutralization of the Positive Charges on Histone Tails by RNA Promotes an Open Chromatin Structure. *Cell Chem Biol* 26, 1436-1449.e1435 (2019). https://doi.org/10.1016/j.chembiol.2019.08.002

62 Yap, K. L. *et al.* Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol Cell* 38, 662-674 (2010). https://doi.org/10.1016/j.molcel.2010.03.021

63 Jain, A. K. *et al.* LncPRESS1 Is a p53-Regulated LncRNA that Safeguards Pluripotency by Disrupting SIRT6-Mediated De-acetylation of Histone H3K56. *Mol Cell* 64, 967-981 (2016). https://doi.org/10.1016/j.molcel.2016.10.039

64 Chammas, P., Mocavini, I. & Di Croce, L. Engaging chromatin: PRC2 structure meets function. *British Journal of Cancer* 122, 315-328 (2020). https://doi.org/10.1038/s41416-019-0615-2

65 Boque-Sastre, R. *et al.* Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. *Proc Natl Acad Sci U S A* 112, 5785-5790 (2015). https://doi.org/10.1073/pnas.1421197112

66 Mondal, T. *et al.* MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA-DNA triplex structures. *Nat Commun* 6, 7743 (2015). https://doi.org/10.1038/ncomms8743

67 Schmitz, K. M., Mayer, C., Postepska, A. & Grummt, I. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev* 24, 2264-2269 (2010). https://doi.org/10.1101/gad.590910

68 Leisegang, M. S., Warwick, T., Stötzel, J. & Brandes, R. P. RNA-DNA triplexes: molecular mechanisms and functional relevance. *Trends in Biochemical Sciences* 49, 532-544 (2024). https://doi.org/https://doi.org/10.1016/j.tibs.2024.03.009

69 Engreitz, J. M. *et al.* RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell* 159, 188-199 (2014). https://doi.org/10.1016/j.cell.2014.08.018

70 Tripathi, V. *et al.* SRSF1 regulates the assembly of pre-mRNA processing factors in nuclear speckles. *Mol Biol Cell* 23, 3694-3706 (2012). https://doi.org/10.1091/mbc.E12-03-0206

71 Yang, L. *et al.* ncRNA- and Pc2 methylation-dependent gene relocation between nuclear structures mediates gene activation programs. *Cell* 147, 773-788 (2011). https://doi.org/10.1016/j.cell.2011.08.054

72 Hacisuleyman, E. *et al.* Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nat Struct Mol Biol* 21, 198-206 (2014). https://doi.org/10.1038/nsmb.2764

73 Latos, P. A. *et al.* Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science* 338, 1469-1472 (2012). https://doi.org/10.1126/science.1228110

74 Stojic, L. *et al.* Transcriptional silencing of long noncoding RNA GNG12-AS1 uncouples its transcriptional and product-related functions. *Nat Commun* 7, 10406 (2016). https://doi.org/10.1038/ncomms10406

75 Thebault, P. *et al.* Transcription regulation by the noncoding RNA SRG1 requires Spt2-dependent chromatin deposition in the wake of RNA polymerase II. *Mol Cell Biol* 31, 1288-1300 (2011). https://doi.org/10.1128/mcb.01083-10

76 Rom, A. *et al.* Regulation of CHD2 expression by the Chaserr long noncoding RNA gene is essential for viability. *Nat Commun* 10, 5092 (2019). https://doi.org/10.1038/s41467-019-13075-8

77 Schertzer, M. D. *et al.* lncRNA-Induced Spread of Polycomb Controlled by Genome Architecture, RNA Abundance, and CpG Island DNA. *Mol Cell* 75, 523-537.e510 (2019). https://doi.org/10.1016/j.molcel.2019.05.028

78 Andergassen, D. *et al.* Mapping the mouse Allelome reveals tissue-specific regulation of allelic expression. *Elife* 6 (2017). https://doi.org/10.7554/eLife.25125

79 Nagano, T. *et al.* The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* 322, 1717-1720 (2008). https://doi.org/10.1126/science.1163802

80 Sleutels, F., Zwart, R. & Barlow, D. P. The non-coding Air RNA is required for silencing autosomal imprinted genes. *Nature* 415, 810-813 (2002). https://doi.org/10.1038/415810a

81 Santoro, F. & Pauler, F. M. Silencing by the imprinted Airn macro lncRNA: transcription is the answer. *Cell Cycle* 12, 711-712 (2013). https://doi.org/10.4161/cc.23860

82 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455-461 (2014). https://doi.org/10.1038/nature12787

83 Hon, C. C. *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* 543, 199-204 (2017). https://doi.org/10.1038/nature21374

84 Kim, Y. J., Xie, P., Cao, L., Zhang, M. Q. & Kim, T. H. Global transcriptional activity dynamics reveal functional enhancer RNAs. *Genome Res* 28, 1799-1811 (2018). https://doi.org/10.1101/gr.233486.117

85 Jiao, W. *et al.* HPSE enhancer RNA promotes cancer progression through driving chromatin looping and regulating hnRNPU/p300/EGR1/HPSE axis. *Oncogene* 37, 2728-2745 (2018). https://doi.org/10.1038/s41388-018-0128-0

86 Yin, Q. F. *et al.* Long noncoding RNAs with snoRNA ends. *Mol Cell* 48, 219-230 (2012). https://doi.org/10.1016/j.molcel.2012.07.033

87 Wu, H. *et al.* Unusual Processing Generates SPA LncRNAs that Sequester Multiple RNA Binding Proteins. *Mol Cell* 64, 534-548 (2016). https://doi.org/10.1016/j.molcel.2016.10.007

88 Guo, C. J. *et al.* Distinct Processing of lncRNAs Contributes to Non-conserved Functions in Stem Cells. *Cell* 181, 621-636.e622 (2020). https://doi.org/10.1016/j.cell.2020.03.006

89 Miller, M. A. & Olivas, W. M. Roles of Puf proteins in mRNA degradation and translation. *Wiley Interdiscip Rev RNA* 2, 471-492 (2011). https://doi.org/10.1002/wrna.69

90 Lee, S. *et al.* Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* 164, 69-80 (2016). https://doi.org/10.1016/j.cell.2015.12.017

91 Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature* 470, 284-288 (2011). https://doi.org/10.1038/nature09701

92 Cesana, M. *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 147, 358-369 (2011). https://doi.org/10.1016/j.cell.2011.09.028

93 Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146, 353-358 (2011). https://doi.org/10.1016/j.cell.2011.07.014

94 Brannan, C. I., Dees, E. C., Ingram, R. S. & Tilghman, S. M. The product of the H19 gene may function as an RNA. *Mol Cell Biol* 10, 28-36 (1990). https://doi.org/10.1128/mcb.10.1.28-36.1990

95 Brockdorff, N. *et al.* The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell* 71, 515-526 (1992). https://doi.org/10.1016/0092-8674(92)90519-i

96 Brown, C. J. *et al.* The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* 71, 527-542 (1992). https://doi.org/10.1016/0092-8674(92)90520-m

97 Lyon, M. F. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature* 190, 372-373 (1961). https://doi.org/10.1038/190372a0

98 Loda, A., Collombet, S. & Heard, E. Gene regulation in time and space during X-chromosome inactivation. *Nat Rev Mol Cell Biol* 23, 231-249 (2022). https://doi.org/10.1038/s41580-021-00438-7

99 Marahrens, Y., Panning, B., Dausman, J., Strauss, W. & Jaenisch, R. Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes Dev* 11, 156-166 (1997). https://doi.org/10.1101/gad.11.2.156

100 Yang, L., Kirby, J. E., Sunwoo, H. & Lee, J. T. Female mice lacking Xist RNA show partial dosage compensation and survive to term. *Genes Dev* 30, 1747-1760 (2016). https://doi.org/10.1101/gad.281162.116

101 Andergassen, D. & Rinn, J. L. From genotype to phenotype: genetics of mammalian long non-coding RNAs in vivo. *Nat Rev Genet* 23, 229-243 (2022). https://doi.org/10.1038/s41576-021-00427-8

102 Okamoto, I., Otte, A. P., Allis, C. D., Reinberg, D. & Heard, E. Epigenetic dynamics of imprinted X inactivation during early mouse development. *Science* 303, 644-649 (2004). https://doi.org/10.1126/science.1092727

103 Inoue, A., Jiang, L., Lu, F. & Zhang, Y. Genomic imprinting of Xist by maternal H3K27me3. *Genes Dev* 31, 1927-1932 (2017). https://doi.org/10.1101/gad.304113.117

104 Takagi, N. & Sasaki, M. Preferential inactivation of the paternally derived X chromosome in the extraembryonic membranes of the mouse. *Nature* 256, 640-642 (1975). https://doi.org/10.1038/256640a0

105 Mak, W. *et al.* Reactivation of the paternal X chromosome in early mouse embryos. *Science* 303, 666-669 (2004). https://doi.org/10.1126/science.1092674

106 Lee, J. T. & Lu, N. Targeted mutagenesis of Tsix leads to nonrandom X inactivation. *Cell* 99, 47-57 (1999). https://doi.org/10.1016/s0092-8674(00)80061-6

107 Ohhata, T. & Wutz, A. Reactivation of the inactive X chromosome in development and reprogramming. *Cell Mol Life Sci* 70, 2443-2461 (2013). https://doi.org/10.1007/s00018-012-1174-3

108 Sugimoto, M. & Abe, K. X chromosome reactivation initiates in nascent primordial germ cells in mice. *PLoS Genet* 3, e116 (2007). https://doi.org/10.1371/journal.pgen.0030116

109 Wutz, A. Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nature Reviews Genetics* 12, 542-553 (2011). https://doi.org/10.1038/nrg3035

110 Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* 341, 1237973 (2013). https://doi.org/10.1126/science.1237973

111 Wutz, A., Rasmussen, T. P. & Jaenisch, R. Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nature Genetics* 30, 167-174 (2002). https://doi.org/10.1038/ng820

112 Chu, C. *et al.* Systematic discovery of Xist RNA binding proteins. *Cell* 161, 404-416 (2015). https://doi.org/10.1016/j.cell.2015.03.025

113 McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* 521, 232-236 (2015). https://doi.org/10.1038/nature14443

114 Żylicz, J. J. *et al.* The Implication of Early Chromatin Changes in X Chromosome Inactivation. *Cell* 176, 182-197.e123 (2019). https://doi.org/https://doi.org/10.1016/j.cell.2018.11.041

115 Bousard, A. *et al.* The role of Xist-mediated Polycomb recruitment in the initiation of X-chromosome inactivation. *EMBO reports* 20, e48019 (2019). https://doi.org/https://doi.org/10.15252/embr.201948019

116 Marks, H. *et al.* Dynamics of gene silencing during X inactivation using allele-specific RNA-seq. *Genome Biology* 16, 149 (2015). https://doi.org/10.1186/s13059-015-0698-x

117 Carrel, L. & Willard, H. F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434, 400-404 (2005). https://doi.org/10.1038/nature03479

118 Tukiainen, T. *et al.* Landscape of X chromosome inactivation across human tissues. *Nature* 550, 244-248 (2017). https://doi.org/10.1038/nature24265

119 Carrel, L. & Brown, C. J. When the Lyon(ized chromosome) roars: ongoing expression from an inactive X chromosome. *Philos Trans R Soc Lond B Biol Sci* 372 (2017). https://doi.org/10.1098/rstb.2016.0355

120 Barr, M. L. & Bertram, E. G. A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature* 163, 676 (1949). https://doi.org/10.1038/163676a0

121 Giorgetti, L. *et al.* Structural organization of the inactive X chromosome in the mouse. *Nature* 535, 575-579 (2016). https://doi.org/10.1038/nature18589

122 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665-1680 (2014). https://doi.org/10.1016/j.cell.2014.11.021

123 Deng, X. *et al.* Bipartite structure of the inactive mouse X chromosome. *Genome Biol* 16, 152 (2015). https://doi.org/10.1186/s13059-015-0728-8

124 Yang, F. *et al.* The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome Biol* 16, 52 (2015). https://doi.org/10.1186/s13059-015-0618-0

125 Darrow, E. M. *et al.* Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture. *Proc Natl Acad Sci U S A* 113, E4504-4512 (2016). https://doi.org/10.1073/pnas.1609643113

126 Fang, H. *et al.* Trans- and cis-acting effects of Firre on epigenetic features of the inactive X chromosome. *Nat Commun* 11, 6053 (2020). https://doi.org/10.1038/s41467-020-19879-3

127 Andergassen, D. *et al.* In vivo Firre and Dxz4 deletion elucidates roles for autosomal gene regulation. *Elife* 8 (2019). https://doi.org/10.7554/eLife.47214

128 Froberg, J. E., Pinter, S. F., Kriz, A. J., Jégu, T. & Lee, J. T. Megadomains and superloops form dynamically but are dispensable for X-chromosome inactivation and gene escape. *Nat Commun* 9, 5004 (2018). https://doi.org/10.1038/s41467-018-07446-w

129 Bonora, G. *et al.* Orientation-dependent Dxz4 contacts shape the 3D structure of the inactive X chromosome. *Nature Communications* 9, 1445 (2018). https://doi.org/10.1038/s41467-018-03694-y

130 Barutcu, A. R., Maass, P. G., Lewandowski, J. P., Weiner, C. L. & Rinn, J. L. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nat Commun* 9, 1444 (2018). https://doi.org/10.1038/s41467-018-03614-0

131 Lewandowski, J. P. *et al.* The Firre locus produces a trans-acting RNA molecule that functions in hematopoiesis. *Nat Commun* 10, 5137 (2019). https://doi.org/10.1038/s41467-019-12970-4

132 Lu, Y. *et al.* The NF-κB-Responsive Long Noncoding RNA FIRRE Regulates Posttranscriptional Regulation of Inflammatory Gene Expression through Interacting with hnRNPU. *J Immunol* 199, 3571-3582 (2017). https://doi.org/10.4049/jimmunol.1700091

133 Sun, L. *et al.* Long noncoding RNAs regulate adipogenesis. *Proc Natl Acad Sci U S A* 110, 3387-3392 (2013). https://doi.org/10.1073/pnas.1222643110

134 Hacisuleyman, E., Shukla, C. J., Weiner, C. L. & Rinn, J. L. Function and evolution of local repeats in the Firre locus. *Nat Commun* 7, 11021 (2016). https://doi.org/10.1038/ncomms11021

135 Much, C. *et al.* The temporal dynamics of lncRNA Firre-mediated epigenetic and transcriptional regulation. *Nature Communications* 15, 6821 (2024). https://doi.org/10.1038/s41467-024-50402-0

136 Abe, Y. *et al.* Xq26.1-26.2 gain identified on array comparative genomic hybridization in bilateral periventricular nodular heterotopia with overlying polymicrogyria. *Developmental Medicine & Child Neurology* 56, 1221-1224 (2014). https://doi.org/https://doi.org/10.1111/dmcn.12553

137 Miolo, G. *et al.* Identification of a De Novo Xq26.2 Microduplication Encompassing FIRRE Gene in a Child with Intellectual Disability. *Diagnostics (Basel)* 10 (2020). https://doi.org/10.3390/diagnostics10121009

138 Haga, Y. *et al.* Increased expression of long non-coding RNA FIRRE promotes hepatocellular carcinoma by HuR-CyclinD1 axis signaling. *J Biol Chem* 300, 107247 (2024). https://doi.org/10.1016/j.jbc.2024.107247

139 Shi, X. *et al.* LncRNA FIRRE is activated by MYC and promotes the development of diffuse large B-cell lymphoma via Wnt/β-catenin signaling pathway. *Biochem Biophys Res Commun* 510, 594-600 (2019). https://doi.org/10.1016/j.bbrc.2019.01.105

140 Wang, Y. *et al.* LncRNA FIRRE functions as a tumor promoter by interaction with PTBP1 to stabilize BECN1 mRNA and facilitate autophagy. *Cell Death & Disease* 13, 98 (2022). https://doi.org/10.1038/s41419-022-04509-1

141 Chow, J. C. *et al.* LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* 141, 956-969 (2010). https://doi.org/10.1016/j.cell.2010.04.042

142 Cleary, S. & Seoighe, C. Perspectives on Allele-Specific Expression. *Annu Rev Biomed Data Sci* 4, 101-122 (2021). https://doi.org/10.1146/annurev-biodatasci-021621-122219

143 Fan, J., Wang, X., Xiao, R. & Li, M. Detecting cell-type-specific allelic expression imbalance by integrative analysis of bulk and single-cell RNA sequencing data. *PLoS Genet* 17, e1009080 (2021). https://doi.org/10.1371/journal.pgen.1009080

144 Qi, G. *et al.* Single-cell allele-specific expression analysis reveals dynamic and cell-type-specific regulatory effects. *Nature Communications* 14, 6317 (2023). https://doi.org/10.1038/s41467-023-42016-9

145 McGrath, J. & Solter, D. Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell* 37, 179-183 (1984). https://doi.org/10.1016/0092-8674(84)90313-1

146 Surani, M. A. H., Barton, S. C. & Norris, M. L. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. *Nature* 308, 548-550 (1984). https://doi.org/10.1038/308548a0

147 Peters, J. The role of genomic imprinting in biology and disease: an expanding view. *Nature Reviews Genetics* 15, 517-530 (2014). https://doi.org/10.1038/nrg3766

148 Barlow, D. P., Stöger, R., Herrmann, B. G., Saito, K. & Schweifer, N. The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. *Nature* 349, 84-87 (1991). https://doi.org/10.1038/349084a0

149 DeChiara, T. M., Robertson, E. J. & Efstratiadis, A. Parental imprinting of the mouse insulin-like growth factor II gene. *Cell* 64, 849-859 (1991). https://doi.org/10.1016/0092-8674(91)90513-x

150 Bartolomei, M. S., Zemel, S. & Tilghman, S. M. Parental imprinting of the mouse H19 gene. *Nature* 351, 153-155 (1991). https://doi.org/10.1038/351153a0

151 Baran, Y. *et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Res* 25, 927-936 (2015). https://doi.org/10.1101/gr.192278.115

152 Court, F. *et al.* Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome Res* 24, 554-569 (2014). https://doi.org/10.1101/gr.164913.113

153 Smith, Z. D. *et al.* DNA methylation dynamics of the human preimplantation embryo. *Nature* 511, 611-615 (2014). https://doi.org/10.1038/nature13581

154 Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B. & Bestor, T. H. Dnmt3L and the establishment of maternal genomic imprints. *Science* 294, 2536-2539 (2001). https://doi.org/10.1126/science.1065848

155 Kaneda, M. *et al.* Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature* 429, 900-903 (2004). https://doi.org/10.1038/nature02633

156 Morgan, H. D., Santos, F., Green, K., Dean, W. & Reik, W. Epigenetic reprogramming in mammals. *Hum Mol Genet* 14 Spec No 1, R47-58 (2005). https://doi.org/10.1093/hmg/ddi114

157 Hanna, C. W. & Kelsey, G. The specification of imprints in mammals. *Heredity (Edinb)* 113, 176-183 (2014). https://doi.org/10.1038/hdy.2014.54

158 Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* 366, 362-365 (1993). https://doi.org/10.1038/366362a0

159 Ooi, S. K. T., O'Donnell, A. H. & Bestor, T. H. Mammalian cytosine methylation at a glance. *Journal of Cell Science* 122, 2787-2791 (2009). https://doi.org/10.1242/jcs.015123

160 Lee, J. *et al.* Erasing genomic imprinting memory in mouse clone embryos produced from day 11.5 primordial germ cells. *Development* 129, 1807-1817 (2002). https://doi.org/10.1242/dev.129.8.1807

161 Barlow, D. P. Genomic imprinting: a mammalian epigenetic discovery model. *Annu Rev Genet* 45, 379-403 (2011). https://doi.org/10.1146/annurev-genet-110410-132459

162 Ferguson-Smith, A. C. Genomic imprinting: the emergence of an epigenetic paradigm. *Nat Rev Genet* 12, 565-575 (2011). https://doi.org/10.1038/nrg3032

163 Bartolomei, M. S. & Ferguson-Smith, A. C. Mammalian genomic imprinting. *Cold Spring Harb Perspect Biol* 3 (2011). https://doi.org/10.1101/cshperspect.a002592

164 Bressler, J. *et al.* The SNRPN promoter is not required for genomic imprinting of the Prader-Willi/Angelman domain in mice. *Nat Genet* 28, 232-240 (2001). https://doi.org/10.1038/90067

165 Fitzpatrick, G. V., Soloway, P. D. & Higgins, M. J. Regional loss of imprinting and growth deficiency in mice with a targeted deletion of KvDMR1. *Nat Genet* 32, 426-431 (2002). https://doi.org/10.1038/ng988

166 Lin, S.-P. *et al.* Asymmetric regulation of imprinting on the maternal and paternal chromosomes at the Dlk1-Gtl2 imprinted cluster on mouse chromosome 12. *Nature Genetics* 35, 97-102 (2003). https://doi.org/10.1038/ng1233

167 Shiura, H. *et al.* Paternal deletion of Meg1/Grb10 DMR causes maternalization of the Meg1/Grb10 cluster in mouse proximal Chromosome 11 leading to severe pre- and postnatal growth retardation. *Hum Mol Genet* 18, 1424-1438 (2009). https://doi.org/10.1093/hmg/ddp049

168 Thorvaldsen, J. L., Duran, K. L. & Bartolomei, M. S. Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2. *Genes Dev* 12, 3693-3702 (1998). https://doi.org/10.1101/gad.12.23.3693

169 Williamson, C. M. *et al.* Identification of an imprinting control region affecting the expression of all transcripts in the Gnas cluster. *Nature Genetics* 38, 350-355 (2006). https://doi.org/10.1038/ng1731

170 Wutz, A. *et al.* Imprinted expression of the Igf2r gene depends on an intronic CpG island. *Nature* 389, 745-749 (1997). https://doi.org/10.1038/39631

171 Mancini-Dinardo, D., Steele, S. J., Levorse, J. M., Ingram, R. S. & Tilghman, S. M. Elongation of the Kcnq1ot1 transcript is required for genomic imprinting of neighboring genes. *Genes Dev* 20, 1268-1282 (2006). https://doi.org/10.1101/gad.1416906

172 Lerchner, W. & Barlow, D. P. Paternal repression of the imprinted mouse Igf2r locus occurs during implantation and is stable in all tissues of the post-implantation mouse embryo. *Mech Dev* 61, 141-149 (1997). https://doi.org/10.1016/s0925-4773(96)00630-2

173 Yamasaki, Y. *et al.* Neuron-specific relaxation of Igf2r imprinting is associated with neuron-specific histone modifications and lack of its antisense transcript Air. *Hum Mol Genet* 14, 2511-2520 (2005). https://doi.org/10.1093/hmg/ddi255

174 Wutz, A. *et al.* Non-imprinted Igf2r expression decreases growth and rescues the Tme mutation in mice. *Development* 128, 1881-1887 (2001). https://doi.org/10.1242/dev.128.10.1881

175 Andergassen, D. *et al.* The Airn lncRNA does not require any DNA elements within its locus to silence distant imprinted genes. *PLoS Genet* 15, e1008268 (2019). https://doi.org/10.1371/journal.pgen.1008268

176 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 506-511 (2013). https://doi.org/10.1038/nature12531

177 Yang, H. H. *et al.* Influence of genetic background and tissue types on global DNA methylation patterns. *PLoS One* 5, e9355 (2010). https://doi.org/10.1371/journal.pone.0009355

178 Orjuela, S., Machlab, D., Menigatti, M., Marra, G. & Robinson, M. D. DAMEfinder: a method to detect differential allele-specific methylation. *Epigenetics Chromatin* 13, 25 (2020). https://doi.org/10.1186/s13072-020-00346-8

179 Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res* 22, 860-869 (2012). https://doi.org/10.1101/gr.131201.111

180 Leung, D. *et al.* Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature* 518, 350-354 (2015). https://doi.org/10.1038/nature14217

181 Kervestin, S. & Jacobson, A. NMD: a multifaceted response to premature translational termination. *Nat Rev Mol Cell Biol* 13, 700-712 (2012). https://doi.org/10.1038/nrm3454

182 Hug, N., Longman, D. & Cáceres, J. F. Mechanism and regulation of the nonsense-mediated decay pathway. *Nucleic Acids Res* 44, 1483-1495 (2016). https://doi.org/10.1093/nar/gkw010

183 Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E. T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet* 7, e1002144 (2011). https://doi.org/10.1371/journal.pgen.1002144

184 Sheinberger, J. *et al.* CD-tagging-MS2: detecting allelic expression of endogenous mRNAs and their protein products in single cells. *Biol Methods Protoc* 2, bpx004 (2017). https://doi.org/10.1093/biomethods/bpx004

185 Kim, J. & Bartel, D. P. Allelic imbalance sequencing reveals that single-nucleotide polymorphisms frequently alter microRNA-directed repression. *Nat Biotechnol* 27, 472-477 (2009). https://doi.org/10.1038/nbt.1540

186 Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am J Hum Genet* 102, 11-26 (2018). https://doi.org/10.1016/j.ajhg.2017.11.002

187 Messemaker, T. C. *et al.* Allele-specific repression of Sox2 through the long non-coding RNA Sox2ot. *Sci Rep* 8, 386 (2018). https://doi.org/10.1038/s41598-017-18649-4

188 Robert, F. & Pelletier, J. Exploring the Impact of Single-Nucleotide Polymorphisms on Translation. *Front Genet* 9, 507 (2018). https://doi.org/10.3389/fgene.2018.00507

189 Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 477, 289-294 (2011). https://doi.org/10.1038/nature10413

190 GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318-1330 (2020). https://doi.org/10.1126/science.aaz1776

191 Castel, S. E., Aguet, F., Mohammadi, P., Ardlie, K. G. & Lappalainen, T. A vast resource of allelic expression data spanning human tissues. *Genome Biol* 21, 234 (2020). https://doi.org/10.1186/s13059-020-02122-z

192 Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biology* 16, 195 (2015). https://doi.org/10.1186/s13059-015-0762-6

193 Manske, H. M. & Kwiatkowski, D. P. SNP-o-matic. *Bioinformatics* 25, 2434-2435 (2009). https://doi.org/10.1093/bioinformatics/btp403

194 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21 (2013). https://doi.org/10.1093/bioinformatics/bts635

195 Miao, Z., Alvarez, M., Pajukanta, P. & Ko, A. ASElux: an ultra-fast and accurate allelic reads counter. *Bioinformatics* 34, 1313-1320 (2018). https://doi.org/10.1093/bioinformatics/btx762

196 Pandey, R. V., Franssen, S. U., Futschik, A. & Schlötterer, C. Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Mol Ecol Resour* 13, 740-745 (2013). https://doi.org/10.1111/1755-0998.12110

197 Lee, C., Kang, E. Y., Gandal, M. J., Eskin, E. & Geschwind, D. H. Profiling allele-specific gene expression in brains from individuals with autism spectrum disorder reveals preferential minor allele usage. *Nat Neurosci* 22, 1521-1532 (2019). https://doi.org/10.1038/s41593-019-0461-9

198 Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* 7, 522 (2011). https://doi.org/10.1038/msb.2011.54

199 Wood, D. L. *et al.* Recommendations for Accurate Resolution of Gene and Isoform Allele-Specific Expression in RNA-Seq Data. *PLoS One* 10, e0126911 (2015). https://doi.org/10.1371/journal.pone.0126911

200 Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. *Nat Commun* 7, 12817 (2016). https://doi.org/10.1038/ncomms12817

201 Andergassen, D. *et al.* Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data. *Nucleic Acids Res* 43, e146 (2015). https://doi.org/10.1093/nar/gkv727

202 Dong, L., Wang, J. & Wang, G. BYASE: a Python library for estimating gene and isoform level allele-specific expression. *Bioinformatics* 36, 4955-4956 (2020). https://doi.org/10.1093/bioinformatics/btaa636

203 Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J. M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res* 21, 1728-1737 (2011). https://doi.org/10.1101/gr.119784.110

204 McCoy, R. C., Wakefield, J. & Akey, J. M. Impacts of Neanderthal-Introgressed Sequences on the Landscape of Human Gene Expression. *Cell* 168, 916-927.e912 (2017). https://doi.org/10.1016/j.cell.2017.01.038

205 Quek, X. C. *et al.* lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Research* 43, D168-D173 (2015). https://doi.org/10.1093/nar/gku988

206 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102, 15545-15550 (2005). https://doi.org/10.1073/pnas.0506580102

207 Fuchs, H. *et al.* Mouse phenotyping. *Methods* 53, 120-135 (2011). https://doi.org/10.1016/j.ymeth.2010.08.006

208 Zhang, B. *et al.* The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. *Cell Rep* 2, 111-123 (2012). https://doi.org/10.1016/j.celrep.2012.06.003

209 Nakagawa, S. *et al.* Malat1 is not an essential component of nuclear speckles in mice. *Rna* 18, 1487-1499 (2012). https://doi.org/10.1261/rna.033217.112

210 Eißmann, M. *et al.* Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. *RNA Biology* 9, 1076-1087 (2012). https://doi.org/10.4161/rna.21089

211 Arun, G., Aggarwal, D. & Spector, D. L. MALAT1 Long Non-Coding RNA: Functional Implications. *Noncoding RNA* 6 (2020). https://doi.org/10.3390/ncrna6020022

212 Zhang, X., Hamblin, M. H. & Yin, K. J. The long noncoding RNA Malat1: Its physiological and pathophysiological functions. *RNA Biol* 14, 1705-1714 (2017). https://doi.org/10.1080/15476286.2017.1358347

213 Zhang, J. & Zhao, H. eQTL studies: from bulk tissues to single cells. *J Genet Genomics* 50, 925-933 (2023). https://doi.org/10.1016/j.jgg.2023.05.003

214 Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* 1, 59 (2021). https://doi.org/10.1038/s43586-021-00056-9

215 Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* 51, D977-d985 (2023). https://doi.org/10.1093/nar/gkac1010

216 Giral, H., Landmesser, U. & Kratzer, A. Into the Wild: GWAS Exploration of Non-coding RNAs. *Front Cardiovasc Med* 5, 181 (2018). https://doi.org/10.3389/fcvm.2018.00181

217 Ning, S. *et al.* LincSNP 2.0: an updated database for linking disease-associated SNPs to human long non-coding RNAs and their TFBSs. *Nucleic Acids Res* 45, D74-d78 (2017). https://doi.org/10.1093/nar/gkw945

218 Aho, A. V., Kernighan, B. W. & Weinberger, P. J. *The AWK programming language*. (Addison-Wesley Longman Publishing Co., Inc., 1987).

219 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841-842 (2010). https://doi.org/10.1093/bioinformatics/btq033

220 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9, 357-359 (2012). https://doi.org/10.1038/nmeth.1923

221 Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 8, 14049 (2017). https://doi.org/10.1038/ncomms14049

222 Hostetter, M. a. K., David A and Seed, Cotton and Terman, Chris and Ward, Stephen. Curl: a gentle slope language for the Web. *World wide web journal* 2, 121-134 (1997).

223 Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* 44, W160-W165 (2016). https://doi.org/10.1093/nar/gkw257

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

225 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297-1303 (2010). https://doi.org/10.1101/gr.107524.110

226 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166-169 (2015). https://doi.org/10.1093/bioinformatics/btu638

227 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9, R137 (2008). https://doi.org/10.1186/gb-2008-9-9-r137

228 Wall, L. a. C., Tom and Orwant, Jon. Programming perl. *O'Reilly Media, Inc.* (2000).

229 Van Rossum, G. & Drake, F. L., Jr. *Python reference manual*. (Centrum voor Wiskunde en Informatica Amsterdam, 1995).

230 R Core Team. R: A Language and Environment for Statistical Computing. (2021).

231 Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* 28 16, 2184-2185 (2012).

232 Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* 10 (2021). https://doi.org/10.1093/gigascience/giab008

233 Krueger, F. & Andrews, S. R. SNPsplit: Allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Res* 5, 1479 (2016). https://doi.org/10.12688/f1000research.9037.2

234 Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* 26, 2204-2207 (2010). https://doi.org/10.1093/bioinformatics/btq351

235 Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* 27, 2156-2158 (2011). https://doi.org/10.1093/bioinformatics/btr330

236 Pagès, H., Carlson, M., Falcon, S. & Li, N. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. (2023). https://doi.org/10.18129/B9.bioc.AnnotationDbi

237 Morgan, M. & Rainer, J. AnnotationFilter: Facilities for Filtering Bioconductor Annotation Resources. (2023). https://doi.org/10.18129/B9.bioc.AnnotationFilter

238 Morgan, M. & Shepherd, L. AnnotationHub: Client to access AnnotationHub resources. (2024). https://doi.org/10.18129/B9.bioc.AnnotationHub

239 Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526-528 (2019). https://doi.org/10.1093/bioinformatics/bty633

240 Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* 35, 2084-2092 (2019). https://doi.org/10.1093/bioinformatics/bty895

241 Eklund, A. & Trimble, J. beeswarm: The Bee Swarm Plot, an Alternative to Stripchart. (2021). https://doi.org/10.32614/CRAN.package.beeswarm

242 Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 3439-3440 (2005).

243 Pagès, H. BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. (2024). https://doi.org/10.18129/B9.bioc.BSgenome

244 Gu, Z. CePa: Centrality-Based Pathway Enrichment. (2022). https://doi.org/10.1093/bioinformatics/btt008>.

245 Zhu, L. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* 11, 237 (2010). https://doi.org/10.1186/1471-2105-11-237

246 Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* 30, 2811-2812 (2014).

247 Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology* 16, 284-287 (2012). https://doi.org/10.1089/omi.2011.0118

248 Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* (2016). https://doi.org/10.1093/bioinformatics/btw313

249 Wilke, C. O. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. (2024). https://doi.org/10.32614/CRAN.package.cowplot

250 Barrett, T. *et al.* data.table: Extension of `data.frame`. (2024). https://doi.org/10.32614/CRAN.package.data.table

251 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014). https://doi.org/10.1186/s13059-014-0550-8

252 Wickham, H., Hester, J., Chang, W. & Bryan, J. devtools: Tools to Make Developing R Packages Easier. (2022). https://doi.org/10.32614/CRAN.package.devtools

253 Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. dplyr: A Grammar of Data Manipulation. (2023). https://doi.org/10.32614/CRAN.package.dplyr

254 Blighe, K., Rana, S. & Lewis, M. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. (2023). https://doi.org/10.18129/B9.bioc.EnhancedVolcano

255 Yu, G. enrichplot: Visualization of Functional Enrichment Result. (2023). https://doi.org/10.18129/B9.bioc.enrichplot

256 Rainer, J., Gatto, L. & Weichenberger, C. X. ensembldb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics* (2019). https://doi.org/10.1093/bioinformatics/btz031

257 Larsson, J. eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. (2024). https://doi.org/10.32614/CRAN.package.eulerr

258 Klaus, B. & Strimmer, K. fdrtool: Estimation of (Local) False Discovery Rates and Higher Criticism. (2024). https://doi.org/10.32614/CRAN.package.fdrtool

259 Warnes, G. R. *et al.* gdata: Various R Programming Tools for Data Manipulation. (2023). https://doi.org/10.32614/CRAN.package.gdata

260 Arora, S., Morgan, M., Carlson, M. & Pagès, H. GenomeInfoDb: Utilities for manipulating chromosome names, including modifying them to follow a particular naming style. (2024). https://doi.org/10.18129/B9.bioc.GenomeInfoDb

261 Bioconductor Core Team. GenomeInfoDbData: Species and taxonomy ID look up tables used by GenomeInfoDb. (2023). https://doi.org/10.18129/B9.bioc.GenomeInfoDbData

262 Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology* 9 (2013). https://doi.org/10.1371/journal.pcbi.1003118

263 Clarke, E., Sherrill-Mix, S. & Dawson, C. ggbeeswarm: Categorical Scatter (Violin Point) Plots. (2023). https://doi.org/10.32614/CRAN.package.ggbeeswarm

264 Xu, S. *et al.* Ggtree: A serialized data object for visualization of a phylogenetic tree and annotation data. *iMeta* 1, e56 (2022). https://doi.org/10.1002/imt2.56

265 Attali, D. & Baker, C. ggExtra: Add Marginal Histograms to 'ggplot2', and More 'ggplot2' Enhancements. (2023). https://doi.org/10.32614/CRAN.package.ggExtra

266 Pedersen, T. L. ggforce: Accelerating 'ggplot2'. (2024). https://doi.org/10.32614/CRAN.package.ggforce

267 Yu, G. & Xu, S. ggfun: Miscellaneous Functions for 'ggplot2'. (2024). https://doi.org/10.32614/CRAN.package.ggfun

268 Briatte, F. ggnetwork: Geometries to Plot Networks with 'ggplot2'. (2024). https://doi.org/10.32614/CRAN.package.ggnetwork

269 Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (2016).

270 Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. (2023). https://doi.org/10.32614/CRAN.package.ggpubr

271 Pedersen, T. L. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. (2024). https://doi.org/10.32614/CRAN.package.ggraph

272 Petukhov, V., van den Brand, T. & Biederstedt, E. ggrastr: Rasterize Layers for 'ggplot2'. (2023). https://doi.org/10.32614/CRAN.package.ggrastr

273 Slowikowski, K. ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. (2024). https://doi.org/10.32614/CRAN.package.ggrepel

274 Wilke, C. O. ggridges: Ridgeline Plots in 'ggplot2'. (2024). https://doi.org/10.32614/CRAN.package.ggridges

275 Xiao, N. ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for 'ggplot2'. (2024). https://doi.org/10.32614/CRAN.package.ggsci

276 Ahlmann-Eltze, C. & Patil, I. ggsignif: R Package for Displaying Significance Brackets for 'ggplot2'. *PsyArxiv* (2021). https://doi.org/10.31234/osf.io/7awm6

277 Yu, G., Smith, D., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 8, 28-36 (2017). https://doi.org/10.1111/2041-210X.12628

278 Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26, 976-978 (2010). https://doi.org/10.1093/bioinformatics/btq064

279 Warnes, G. R. *et al.* gplots: Various R Programming Tools for Plotting Data. (2024). https://doi.org/10.32614/CRAN.package.gplots

280 Murrell, P. gridBase: Integration of base and grid graphics. (2014).

281 Auguie, B. gridExtra: Miscellaneous Functions for "Grid" Graphics. (2017). https://doi.org/10.32614/CRAN.package.gridExtra

282 Murrell, P. & Wen, Z. gridGraphics: Redraw Base Graphics Using 'grid' Graphics. (2020). https://doi.org/10.32614/CRAN.package.gridGraphics

283 Warnes, G. R. *et al.* gtools: Various R Programming Tools. (2023). https://doi.org/10.32614/CRAN.package.gtools

284 Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* Complex Systems, 1695 (2006).

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

286 Mays, J. karyotapR: DNA Copy Number Analysis for Genome-Wide Tapestri Panels. (2023). https://doi.org/10.32614/CRAN.package.karyotapR

287 Kelly, S. T. leiden: R implementation of the Leiden algorithm. (2023). https://doi.org/10.32614/CRAN.package.leiden

288 Ewing, B. leidenbase: R and C/C++ Wrappers to Run the Leiden find_partition() Function. (2024). https://doi.org/10.32614/CRAN.package.leidenbase

289 Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res* 40, e133 (2012). https://doi.org/10.1093/nar/gks461

290 Bengtsson, H. matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors). (2024). https://doi.org/10.32614/CRAN.package.matrixStats

291 Bredikhin, D., Kats, I. & Fang, Z. MuDataSeurat: MuData Serialization for Seurat. (2023).

292 Carlson, M. org.Hs.eg.db: Genome wide annotation for Human. (2023). https://doi.org/10.18129/B9.bioc.org.Hs.eg.db

293 Kolde, R. pheatmap: Pretty Heatmaps. (2019). https://doi.org/10.32614/CRAN.package.pheatmap

294 Wickham, H. The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software* 40, 1-29 (2011).

295 Urbanek, S. png: Read and write PNG images. (2022). https://doi.org/10.32614/CRAN.package.png

296 Storey, J. D., Bass, A. J., Dabney, A. & Robinson, D. qvalue: Q-value estimation for false discovery rate control. (2023). https://doi.org/10.18129/B9.bioc.qvalue

297 Neuwirth, E. RColorBrewer: ColorBrewer Palettes. (2022). https://doi.org/10.32614/CRAN.package.RColorBrewer

298 Wickham, H., Hester, J. & Bryan, J. readr: Read Rectangular Text Data. (2024). https://doi.org/10.32614/CRAN.package.readr

299 Wickham, H. & Bryan, J. readxl: Read Excel Files. (2023). https://doi.org/10.32614/CRAN.package.readxl

300 Bryan, J., Hester, J., Robinson, D., Wickham, H. & Dervieux, C. reprex: Prepare Reproducible Example Code via the Clipboard. (2024). https://doi.org/10.32614/CRAN.package.reprex

301 Morgan, M., Pagès, H., Obenchain, V. & Hayden, N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import. (2023). https://doi.org/10.18129/B9.bioc.Rsamtools

302 Wickham, H., Pedersen, T. L. & Seidel, D. scales: Scale Functions for Visualization. (2023). https://doi.org/10.32614/CRAN.package.scales

303 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* 184, 3573-3587.e3529 (2021). https://doi.org/https://doi.org/10.1016/j.cell.2021.04.048

304 Gu, Z. & Hübschmann, D. simplifyEnrichment: A Bioconductor Package for Clustering and Visualizing Functional Enrichment Results. *Genomics Proteomics Bioinformatics* 21, 190-202 (2023). https://doi.org/10.1016/j.gpb.2022.04.008

305 Wickham, H. stringr: Simple, Consistent Wrappers for Common String Operations. (2023). https://doi.org/10.32614/CRAN.package.stringr

306 Morgan, M., Obenchain, V., Hester, J. & Pagès, H. SummarizedExperiment: SummarizedExperiment container. (2023). https://doi.org/10.18129/B9.bioc.SummarizedExperiment

307 Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* 20, 296 (2019). https://doi.org/10.1186/s13059-019-1874-1

308 Müller, K. & Wickham, H. tibble: Simple Data Frames. (2023). https://doi.org/10.32614/CRAN.package.tibble

309 Pedersen, T. L. tidygraph: A Tidy API for Graph Manipulation. (2024). https://doi.org/10.32614/CRAN.package.tidygraph

310 Wickham, H., Vaughan, D. & Girlich, M. tidyr: Tidy Messy Data. (2024). https://doi.org/10.32614/CRAN.package.tidyr

311 Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* 4, 1686 (2019). https://doi.org/10.21105/joss.01686

312 Gehlenborg, N. UpSetR: A More Scalable Alternative to Venn and Euler Diagrams for Visualizing Intersecting Sets. (2019). https://doi.org/10.32614/CRAN.package.UpSetR

313 Chen, H. VennDiagram: Generate High-Resolution Venn and Euler Plots. (2022). https://doi.org/10.32614/CRAN.package.VennDiagram

314 Ooms, J. writexl: Export Data Frames to Excel 'xlsx' Format. (2024). https://doi.org/10.32614/CRAN.package.writexl

315 Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific Reports* 9, 9354 (2019). https://doi.org/10.1038/s41598-019-45839-z

316 Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* 47, D766-d773 (2019). https://doi.org/10.1093/nar/gky955

317 Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30, 207-210 (2002). https://doi.org/10.1093/nar/30.1.207

318 Wen, X., Pique-Regi, R. & Luca, F. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS Genet* 13, e1006646 (2017). https://doi.org/10.1371/journal.pgen.1006646

319 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 44, D733-745 (2016). https://doi.org/10.1093/nar/gkv1189

320 Xu, J. *et al.* Landscape of monoallelic DNA accessibility in mouse embryonic stem cells and neural progenitor cells. *Nat Genet* 49, 377-386 (2017). https://doi.org/10.1038/ng.3769

321 Liu, C. *et al.* An ATAC-seq atlas of chromatin accessibility in mouse tissues. *Sci Data* 6, 65 (2019). https://doi.org/10.1038/s41597-019-0071-0

322 Hasenbein, T. P. *et al.* X-linked deletion of Crossfirre, Firre, and Dxz4 in vivo uncovers diverse phenotypes and combinatorial effects on autosomes. *Nature Communications* 15, 10631 (2024). https://doi.org/10.1038/s41467-024-54673-5

323 Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* 153, 910-918 (2013). https://doi.org/10.1016/j.cell.2013.04.025

324 Fuchs, H. *et al.* Understanding gene functions and disease mechanisms: Phenotyping pipelines in the German Mouse Clinic. *Behav Brain Res* 352, 187-196 (2018). https://doi.org/10.1016/j.bbr.2017.09.048

325 Andergassen, D., Smith, Z. D., Kretzmer, H., Rinn, J. L. & Meissner, A. Diverse epigenetic mechanisms maintain parental imprints within the embryonic and extraembryonic lineages. *Dev Cell* 56, 2995-3005.e2994 (2021). https://doi.org/10.1016/j.devcel.2021.10.010

326 Borensztein, M. *et al.* Xist-dependent imprinted X inactivation and the early developmental consequences of its failure. *Nat Struct Mol Biol* 24, 226-233 (2017). https://doi.org/10.1038/nsmb.3365

327 Lyon, M. F. Do LINEs have a role in X-chromosome inactivation? *J Biomed Biotechnol* 2006, 59746 (2006). https://doi.org/10.1155/jbb/2006/59746

328 Calaway, J. D. *et al.* Genetic architecture of skewed X inactivation in the laboratory mouse. *PLoS Genet* 9, e1003853 (2013). https://doi.org/10.1371/journal.pgen.1003853

329 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* 29, 24-26 (2011). https://doi.org/10.1038/nbt.1754

330 Trowsdale, J. & Knight, J. C. Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet* 14, 301-323 (2013). https://doi.org/10.1146/annurev-genom-091212-153455

331 Sanli, I. *et al.* Meg3 Non-coding RNA Expression Controls Imprinting by Preventing Transcriptional Upregulation in cis. *Cell Rep* 23, 337-348 (2018). https://doi.org/10.1016/j.celrep.2018.03.044

332 Grote, P. *et al.* The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev Cell* 24, 206-214 (2013). https://doi.org/10.1016/j.devcel.2012.12.012

333 Ferrer, J. & Dimitrova, N. Transcription regulation by long non-coding RNAs: mechanisms and disease relevance. *Nat Rev Mol Cell Biol* 25, 396-415 (2024). https://doi.org/10.1038/s41580-023-00694-9

334 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 47, D1005-d1012 (2019). https://doi.org/10.1093/nar/gky1120

335 Phillips, P. C. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics* 9, 855-867 (2008). https://doi.org/10.1038/nrg2452

336 Goldstein, I., Paakinaho, V., Baek, S., Sung, M.-H. & Hager, G. L. Synergistic gene expression during the acute phase response is characterized by transcription factor assisted loading. *Nature Communications* 8, 1849 (2017). https://doi.org/10.1038/s41467-017-02055-5

337 Wang, C. & Lin, H. Roles of piRNAs in transposon and pseudogene regulation of germline mRNAs and lncRNAs. *Genome Biology* 22, 27 (2021). https://doi.org/10.1186/s13059-020-02221-x

338 Carthew, R. W. Gene Regulation and Cellular Metabolism: An Essential Partnership. *Trends Genet* 37, 389-400 (2021). https://doi.org/10.1016/j.tig.2020.09.018

339 Shore, D. & Albert, B. Ribosome biogenesis and the cellular energy economy. *Curr Biol* 32, R611-r617 (2022). https://doi.org/10.1016/j.cub.2022.04.083

340 Kondo, Y. *et al.* Moderate protein intake percentage in mice for maintaining metabolic health during approach to old age. *Geroscience* 45, 2707-2726 (2023). https://doi.org/10.1007/s11357-023-00797-3

341 Li, X. *et al.* Lactate metabolism in human health and disease. *Signal Transduct Target Ther* 7, 305 (2022). https://doi.org/10.1038/s41392-022-01151-3

342 Kazak, L. & Cohen, P. Creatine metabolism: energy homeostasis, immunity and cancer biology. *Nat Rev Endocrinol* 16, 421-436 (2020). https://doi.org/10.1038/s41574-020-0365-5

343 Brosnan, J. T. & Brosnan, M. E. Creatine metabolism and the urea cycle. *Mol Genet Metab* 100 Suppl 1, S49-52 (2010). https://doi.org/10.1016/j.ymgme.2010.02.020

344 Brooks, G. A. The Science and Translation of Lactate Shuttle Theory. *Cell Metabolism* 27, 757-785 (2018). https://doi.org/https://doi.org/10.1016/j.cmet.2018.03.008

345 Zang, Y., Zhou, X., Wang, Q., Li, X. & Huang, H. LncRNA FIRRE/NF-kB feedback loop contributes to OGD/R injury of cerebral microglial cells. *Biochem Biophys Res Commun* 501, 131-138 (2018). https://doi.org/10.1016/j.bbrc.2018.04.194

346 Alkhamra, R. A. & Abu-Dahab, S. M. N. Sensory processing disorders in children with hearing impairment: Implications for multidisciplinary approach and early intervention. *Int J Pediatr Otorhinolaryngol* 136, 110154 (2020). https://doi.org/10.1016/j.ijporl.2020.110154

347 Karp, N. A. *et al.* Prevalence of sexual dimorphism in mammalian phenotypic traits. *Nature Communications* 8, 15475 (2017). https://doi.org/10.1038/ncomms15475

348 Moore, G. E. *et al.* The role and interaction of imprinted genes in human fetal growth. *Philos Trans R Soc Lond B Biol Sci* 370, 20140074 (2015). https://doi.org/10.1098/rstb.2014.0074

349 Hill, M. S., Vande Zande, P. & Wittkopp, P. J. Molecular and evolutionary processes generating variation in gene expression. *Nat Rev Genet* 22, 203-215 (2021). https://doi.org/10.1038/s41576-020-00304-w

350 de Goede, O. M. *et al.* Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. *Cell* 184, 2633-2648.e2619 (2021). https://doi.org/https://doi.org/10.1016/j.cell.2021.03.050

351 Võsa, U., Esko, T., Kasela, S. & Annilo, T. Altered Gene Expression Associated with microRNA Binding Site Polymorphisms. *PLoS One* 10, e0141351 (2015). https://doi.org/10.1371/journal.pone.0141351

352 Rosati, D. *et al.* Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Comput Struct Biotechnol J* 23, 1154-1168 (2024). https://doi.org/10.1016/j.csbj.2024.02.018

353 Panten, J. *et al.* The dynamic genetic determinants of increased transcriptional divergence in spermatids. *Nature Communications* 15, 1272 (2024). https://doi.org/10.1038/s41467-024-45133-1

354 Sherbina, K., León-Novelo, L. G., Nuzhdin, S. V., McIntyre, L. M. & Marroni, F. Power calculator for detecting allelic imbalance using hierarchical Bayesian model. *BMC Research Notes* 14, 436 (2021). https://doi.org/10.1186/s13104-021-05851-x

355 Kornienko, A. E., Guenzl, P. M., Barlow, D. P. & Pauler, F. M. Gene regulation by the act of long non-coding RNA transcription. *BMC Biol* 11, 59 (2013). https://doi.org/10.1186/1741-7007-11-59

356 Long, J. *et al.* Long noncoding RNA Tug1 regulates mitochondrial bioenergetics in diabetic nephropathy. *J Clin Invest* 126, 4205-4218 (2016). https://doi.org/10.1172/jci87927

357 Lewandowski, J. P. *et al.* The Tug1 lncRNA locus is essential for male fertility. *Genome Biol* 21, 237 (2020). https://doi.org/10.1186/s13059-020-02081-5

358 Tsouris, A., Brach, G., Schacherer, J. & Hou, J. Non-additive genetic components contribute significantly to population-wide gene expression variation. *Cell Genomics* 4, 100459 (2024). https://doi.org/https://doi.org/10.1016/j.xgen.2023.100459

359 Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics* 12, 703-714 (2011). https://doi.org/10.1038/nrg3054

360 Hedrick, P. W. Balancing selection. *Current Biology* 17, R230-R231 (2007). https://doi.org/https://doi.org/10.1016/j.cub.2007.01.012

361 Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics & Chromatin* 8, 57 (2015). https://doi.org/10.1186/s13072-015-0050-4

362 Kumar, V. *et al.* Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet* 9, e1003201 (2013). https://doi.org/10.1371/journal.pgen.1003201

363 Ferraro, N. M. *et al.* Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science* 369 (2020). https://doi.org/10.1126/science.aaz5900

364 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190-1195 (2012). https://doi.org/10.1126/science.1222794

8 Acknowledgements

First, I would like to sincerely thank Dr. Daniel Andergassen for the opportunity to conduct my thesis in his lab. Thank you for the chance to work on the projects and for the ongoing support throughout my research. I am grateful for your excellent scientific guidance and for creating a work environment where everyone feels comfortable.

Further, I would like to thank Prof. Dr. Dr. Stefan Engelhardt for the possibility of working in the Institute of Pharmacology and Toxicology. Moreover, I would like to thank Prof. Dr. Dr. Stefan Engelhardt and Prof. Dr. Martin Hrabě de Angelis for being part of my thesis committee and for the valuable feedback and guidance throughout the projects.

I would also like to thank Prof. Dr. John Rinn for the great collaboration, insightful discussions, and thoughtful feedback on the projects, which significantly enriched my work. Moreover, I want to thank the entire team of the German Mouse Clinic for their expert support and the productive collaboration associated with Project 1.

Special thanks also go to Sarah Hoelzl who was an invaluable colleague, not only because she took care of all the wet-lab procedures, but becoming a friend. Further, I am thankful to Marion for her excellent genotyping skills, as well as Niklas and Federico for their valuable feedback and making the time fun and enjoyable. Further, I would like to thank all my colleagues at the Andergassen Lab and the Institute of Pharmacology and Toxicology for fostering a nice and friendly working environment.

Additionally, I would like to thank Michaela Hennig, Violetta Cavic, and Elif Berger for their excellent support with all administrative matters. I would also like to thank Anton Bomhard, Julia Auerswald, Selahattin Sahiner, Antonio Sterle, and Mehmet Durmaz for their mouse handling work, as well as Josef Reischenbeck for the great technical support.

Finally, I would like to thank the people who mean the most to me. To my family: I am thankful for your endless support and for each and every one of you: my grandparents, my grandma, my mother, my father, my sister, my brother-in-law, my nephews. To my extended family: Henni, Aileen, Elli, Daniel Kolbe, Joanna, thank you for the happiness that true friendship brings. To my wonderful wife: Thank you for being my home. Thank you for believing, when it was hard to believe. I am deeply grateful for living this life with you.

9 Publications

9.1 Peer-reviewed publications

1. **Tim P. Hasenbein***, Sarah Hoelzl*, Zachary D. Smith, Chiara Gerhardinger, Marion O. C. Gonner, Antonio Aguilar-Pimentel, Oana V. Amarie, Lore Becker, Julia Calzada-Wack, Nathalia R. V. Dragano, Patricia da Silva-Buttkus, Lillian Garrett, Sabine M. Hölter, Markus Kraiger, Manuela A. Östereicher, Birgit Rathkolb, Adrián Sanz-Moreno, Nadine Spielmann, Wolfgang Wurst, Valerie Gailus-Durner, Helmut Fuchs, Martin Hrabě de Angelis, Alexander Meissner, Stefan Engelhardt, John L. Rinn† and Daniel Andergassen† *X-linked deletion of Crossfirre, Firre, and Dxz4 in vivo uncovers diverse phenotypes and combinatorial effects on autosomes.* Nature Communications (2024)

2. Sarah Hoelzl, **Tim P. Hasenbein**, Stefan Engelhardt, Daniel Andergassen *Aging promotes reactivation of the Barr body at distal chromosome regions.* Nature Aging (accepted in principle)

3. Karin Ziegler, Manuel Zeitler, Sandro Meunier, Inga Sinicina, **Tim P. Hasenbein**, Daniel Andergassen, Anton Bomhard, Reginald van der Kwast, and Stefan Engelhardt *Ganglionic inflammation in a patient with takotsubo syndrome.* Circulation (accepted)

9.2 Submitted manuscripts

1. **Tim P. Hasenbein**, Sarah Hoelzl, Stefan Engelhardt, and Daniel Andergassen *Allele-specific genomics decodes gene targets and mechanisms of the non-coding genome* (rejected after being under review at Cell, currently with scientific editor at Cell Press Transfer)

2. Lison Lemoine, Sarah Hoelzl, **Tim P. Hasenbein**, Elisabeth Graf, Daniel Andergassen *Using long-read sequencing to shed light on complex allele-specific loci* (in review process, Scientific Reports)

9.3 Conference presentations

Tim P. Hasenbein, Sarah Hoelzl, Stefan Engelhardt, and Daniel Andergassen *Allele-specific genomics decodes gene targets and mechanisms of the non-coding genome,* TRR267 Retreat, Würzburg, 2022 – **Poster**

Tim P. Hasenbein, Sarah Hoelzl, Stefan Engelhardt, and Daniel Andergassen *Allele-specific genomics decodes gene targets and mechanisms of the non-coding genome*, Munich Heart Alliance Summer Meeting, Bernried am Starnberger See, 2022 – **Poster**

Tim P. Hasenbein, Sarah Hoelzl, Stefan Engelhardt, and Daniel Andergassen *Allele-specific genomics decodes gene targets and mechanisms of the non-coding genome*, Munich Heart Alliance Winter Meeting, München, 2023 – **Poster**

Tim P. Hasenbein*, Sarah Hoelzl*, Zachary D. Smith, Chiara Gerhardinger, Marion O. C. Gonner, Antonio Aguilar-Pimentel, Oana V. Amarie, Lore Becker, Julia Calzada-Wack, Nathalia R. V. Dragano, Patricia da Silva-Buttkus, Lillian Garrett, Sabine M. Hölter, Markus Kraiger, Manuela A. Östereicher, Birgit Rathkolb, Adrián Sanz-Moreno, Nadine Spielmann, Wolfgang Wurst, Valerie Gailus-Durner, Helmut Fuchs, Martin Hrabě de Angelis, Alexander Meissner, Stefan Engelhardt, John L. Rinn† and Daniel Andergassen† *X-linked deletion of Crossfirre, Firre, and Dxz4 in vivo uncovers diverse phenotypes and combinatorial effects on autosomes*, EMBO Workshop X-chromosome inactivation: new insights on its 60th anniversary, Berlin 2023 – **Talk**

Tim P. Hasenbein, Sarah Hoelzl, Stefan Engelhardt, and Daniel Andergassen *Allele-specific genomics decodes gene targets and mechanisms of the non-coding genome*, Munich Heart Alliance Summer Meeting, München, 2023 – **Poster**

Tim P. Hasenbein*, Sarah Hoelzl*, Zachary D. Smith, Chiara Gerhardinger, Marion O. C. Gonner, Antonio Aguilar-Pimentel, Oana V. Amarie, Lore Becker, Julia Calzada-Wack, Nathalia R. V. Dragano, Patricia da Silva-Buttkus, Lillian Garrett, Sabine M. Hölter, Markus Kraiger, Manuela A. Östereicher, Birgit Rathkolb, Adrián Sanz-Moreno, Nadine Spielmann, Wolfgang Wurst, Valerie Gailus-Durner, Helmut Fuchs, Martin Hrabě de Angelis, Alexander Meissner, Stefan Engelhardt, John L. Rinn† and Daniel Andergassen† *X-linked deletion of Crossfirre, Firre, and Dxz4 in vivo uncovers diverse phenotypes and combinatorial effects on autosomes*, EMBO | EMBL Symposium: The non-coding genome, Heidelberg, 2023 – **Poster**

(Award)

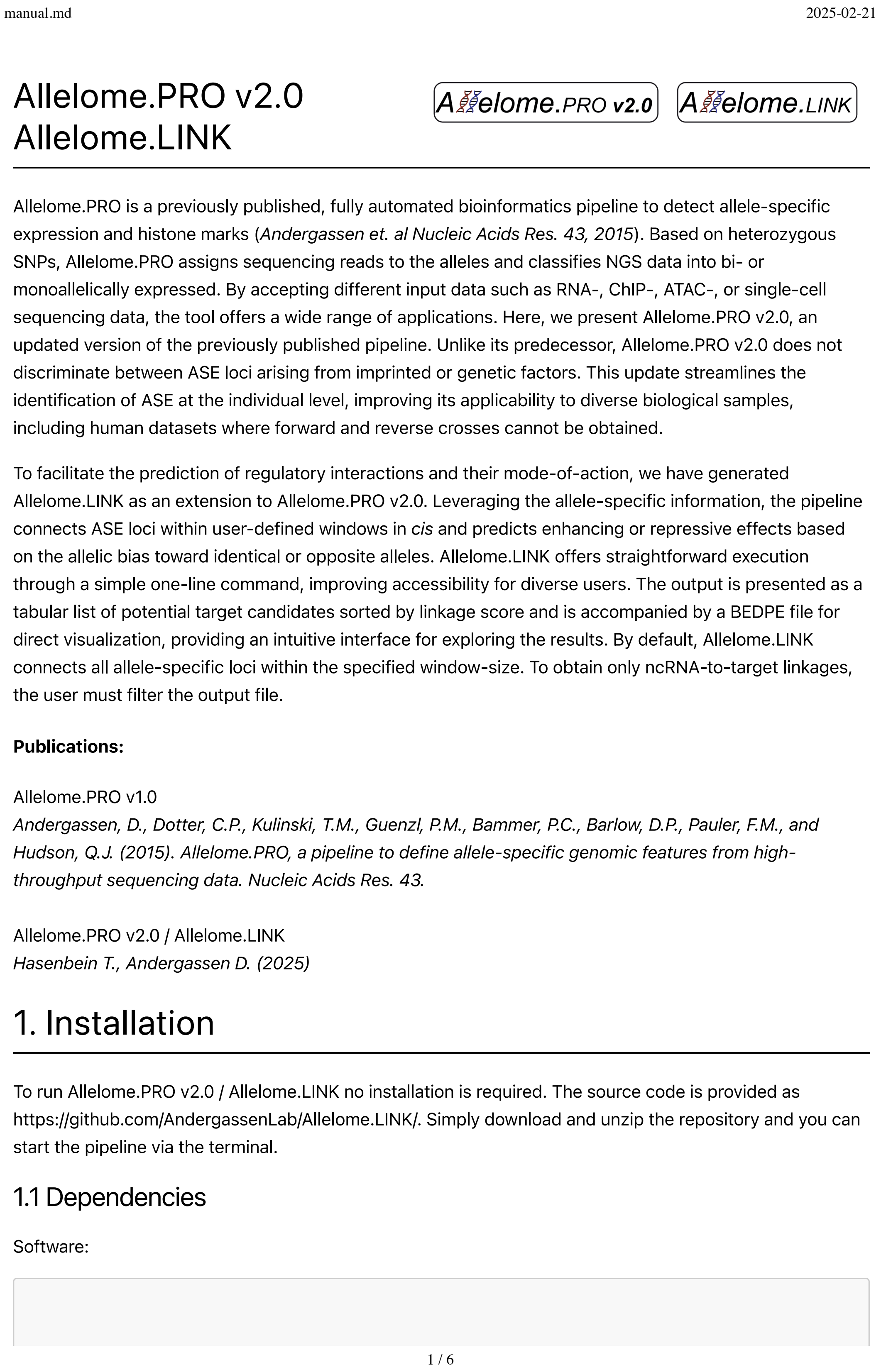
Tim P. Hasenbein, Sarah Hoelzl, Stefan Engelhardt, and Daniel Andergassen *Allele-specific genomics decodes gene targets and mechanisms of the non-coding genome*, Munich Heart Alliance Winter Meeting, München, 2024 – **Poster**

Tim P. Hasenbein*, Sarah Hoelzl*, Zachary D. Smith, Chiara Gerhardinger, Marion O. C. Gonner, Antonio Aguilar-Pimentel, Oana V. Amarie, Lore Becker, Julia Calzada-Wack, Nathalia R. V. Dragano, Patricia da Silva-Buttkus, Lillian Garrett, Sabine M. Hölter, Markus Kraiger, Manuela A. Östereicher, Birgit Rathkolb, Adrián Sanz-Moreno, Nadine Spielmann, Wolfgang Wurst, Valerie Gailus-Durner, Helmut Fuchs, Martin Hrabě de Angelis, Alexander Meissner, Stefan Engelhardt, John L. Rinn† and Daniel Andergassen† *X-linked deletion of Crossfirre, Firre, and Dxz4 in vivo uncovers diverse phenotypes and combinatorial effects on autosomes*, 9th German Pharm-Tox Summit, München, 2024 – **Poster**

Tim P. Hasenbein*, Sarah Hoelzl*, Zachary D. Smith, Chiara Gerhardinger, Marion O. C. Gonner, Antonio Aguilar-Pimentel, Oana V. Amarie, Lore Becker, Julia Calzada-Wack, Nathalia R. V. Dragano, Patricia da Silva-Buttkus, Lillian Garrett, Sabine M. Hölter, Markus Kraiger, Manuela A. Östereicher, Birgit Rathkolb, Adrián Sanz-Moreno, Nadine Spielmann, Wolfgang Wurst, Valerie Gailus-Durner, Helmut Fuchs, Martin Hrabě de Angelis, Alexander Meissner, Stefan Engelhardt, John L. Rinn† and Daniel Andergassen† *X-linked deletion of Crossfirre, Firre, and Dxz4 in vivo uncovers diverse phenotypes and combinatorial effects on autosomes*, Keystone Symposia: Non-Coding RNA Biology: New Roles and Diversity, Keystone, Colorado 2025 – **Talk**

10 Appendix

10.1 Manual of the Allelome.PRO v2.0 and Allelome.LINK pipeline



Allelome.PRO v2.0 Allelome.LINK

Allelome.PRO is a previously published, fully automated bioinformatics pipeline to detect allele-specific expression and histone marks (*Andergassen et. al Nucleic Acids Res. 43, 2015*). Based on heterozygous SNPs, Allelome.PRO assigns sequencing reads to the alleles and classifies NGS data into bi- or monoallelically expressed. By accepting different input data such as RNA-, ChIP-, ATAC-, or single-cell sequencing data, the tool offers a wide range of applications. Here, we present Allelome.PRO v2.0, an updated version of the previously published pipeline. Unlike its predecessor, Allelome.PRO v2.0 does not discriminate between ASE loci arising from imprinted or genetic factors. This update streamlines the identification of ASE at the individual level, improving its applicability to diverse biological samples, including human datasets where forward and reverse crosses cannot be obtained.

To facilitate the prediction of regulatory interactions and their mode-of-action, we have generated Allelome.LINK as an extension to Allelome.PRO v2.0. Leveraging the allele-specific information, the pipeline connects ASE loci within user-defined windows in *cis* and predicts enhancing or repressive effects based on the allelic bias toward identical or opposite alleles. Allelome.LINK offers straightforward execution through a simple one-line command, improving accessibility for diverse users. The output is presented as a tabular list of potential target candidates sorted by linkage score and is accompanied by a BEDPE file for direct visualization, providing an intuitive interface for exploring the results. By default, Allelome.LINK connects all allele-specific loci within the specified window-size. To obtain only ncRNA-to-target linkages, the user must filter the output file.

Publications:

Allelome.PRO v1.0
Andergassen, D., Dotter, C.P., Kulinski, T.M., Guenzl, P.M., Bammer, P.C., Barlow, D.P., Pauler, F.M., and Hudson, Q.J. (2015). Allelome.PRO, a pipeline to define allele-specific genomic features from high-throughput sequencing data. Nucleic Acids Res. 43.

Allelome.PRO v2.0 / Allelome.LINK
Hasenbein T., Andergassen D. (2025)

1. Installation

To run Allelome.PRO v2.0 / Allelome.LINK no installation is required. The source code is provided as https://github.com/AndergassenLab/Allelome.LINK/. Simply download and unzip the repository and you can start the pipeline via the terminal.

1.1 Dependencies

Software:

```
bedtools (≥ version 2.20.1)
SAMtools (≥ version 0.1.19)
R (≥ version 3.1.0)
Perl (≥ version 5.20.0)
fetchChromSizes (≥ version 377)
bedToBigBed (≥ version 377)
```

R packages:

```
plyr; gtools; optparse
# If not installed, Allelome.PRO v2.0 / Allelome.LINK will try to install
them in the default R library path
```

2. Run Allelome.PRO v2.0

To run Allelome.PRO v2.0, you need to prepare your input files as described below (2.1). Once you have your sample BAM, the annotation and SNP file, you can start the pipeline from the command line by typing:

```
bash Allelome.PROv2.0.sh -i <input_bam> -a <annotation_file> -s <SNP_file>
-o <output_directory> [options]
```

Please make sure that Allelome.PROv2.0.sh is in your PATH, or specify the absolute path of the script. For all input files, we suggest giving full paths and locations. Besides the required input flags, you have additional options, as shown below in table 2.

2.1 Input files

Three different files are required as input for Allelome.PRO v2.0:

- Aligned sample file (BAM format)
- Annotation file (BED6 format)
- SNP file (BED6 format)

2.1.1 Sample file:

The sample file is a BAM file that stores the aligned NGS-reads from either RNA-, ChIP-, ATAC-, or scRNA-sequencing in binary format. The file can be sorted by coordinates or unsorted. If unsorted, Allelome.PRO v2.0 can sort the BAM file for you.

2.1.2 Annotation file:

The annotation file is a six-column text file containing the position information of your loci of interest in BED6 format (see UCSC format description for more details). The helperscript createAnnotation.sh can be used to create your own annotation file.

Table 1 Overview of the BED6 annotation file format.

Column	Description
1	Chromosome(eg. chr1)
2	Start Position
3	End Position
4	Name (e.g. gene name)
5	Score (not used here)
6	Strand (e.g. +,- or . for not defined)

2.1.3 SNP file:

The SNP file is also a six-column text file in BED6 format. However, the difference to the annotation file is that the *name* column consists of two letters representing the two SNPs present at a given locus. The SNP file is used to determine which SNP is located on which allele. Therefore, the order of the two-letter SNPs is important since the first base indicates the variant at allele 1, while the second base is present at allele 2. The way the SNP file is created determines which allele is "1" and "2". SNP positions must be based on the same reference genome to which the BAM file was aligned to (e.g. mm10). A source for mouse SNP data is the FTP site of the Sanger Institute (e.g. https://ftp.ebi.ac.uk/pub/databases/mousegenomes/REL-1505-SNPs_Indels/). If Allelome.PRO v2.0 is used with hybrid F1 mice samples, the helperscript createSNPfile.sh can be used to generate the SNP file via:

```
sh create_SNPfile_v5.sh mgp.v5.merged.snps_all.dbSNP142.vcf
```

If Allelome.PRO v2.0 is used with eg. human data, SNPs must be called and phased in advance to generate the SNP file.

2.2 Input flags for Allelome.PRO.v2.0:

Table 2. Overview of the Allelome.PRO v2.0 options.

Required	Description
-i	* Input sample file (bam format)
-a	* Annotation file (BED6 format)
-s	* SNP file (BED6 format)
-o	* Output directory
Optional:	

Required	Description
-z	* Specify if bam file is sorted (1 sorted (default); 0 unsorted)
-r	* Min. number of reads to cover a SNP to be included (default 1)
-t	* Min. number of total reads to cover a locus to be included (default 20)
Misc:	
-h	* Display this help message

2.3 Output files

Allelome.PRO v2.0 generates a results directory with different files, including the allelic bias of your sequencing reads, as well as files for debugging and visualization. The main output files are the **locus_table.txt** file with information on the allelic status of each locus in the annotation and the **.bed** file for visualizing your results in a genome browser. Please find detailed information about all output files in Table 3.

Table 3. Overview of the Allelome.PRO v2.0 output files.

Name	Description
locus_table.txt	Information about the allelic ratio of all informative loci
read_count_per_SNP.txt	Information about the number of reads covering the individual SNPs
.log	Log file with information about your Allelome.PRO v2.0 run
./BED_files:	
.bed	BED6-file containing all loci, color-coded according to their allelic ratio
.bb	Big bed file of .bed
/debug	Folder containing files created during the run for debugging
_SNP.pileup	
.pileup	
trimmed_s.bam	
/debug/annotation	Folder containing files created during the run for debugging
annotation_overlapping_snps.txt	
annotation_sorted.bed	
snps_overlapping_annotation_forjoin.txt	
snps_overlapping_annotation_list.txt	

Name	Description
snps_overlapping_annotation.bed	

3. Run Allelome.LINK

To run Allelome.LINK the *locus_table.txt* file from the Allelome.PRO v2.0 run is required. To start Allelome.LINK, type in your command line:

```
Rscript Allelome.LINK.R –i <input_locus_table.txt> –o <output_directory>
[options]
```

Again, please make sure that Allelome.LINK is in your PATH, or give the full path to the file. Allelome.LINK will link loci within a given window size in *cis*. You can specify your genomic window with the --window-size or -w flag, which is the number of base pairs up- and downstream of the locus (see Table 4). The default value is 100kb in each direction from a given locus. Further, you can specify the number of total reads that must cover a gene to be included for the analysis (--total-reads/-r; default: 20), as well as the --allelelic-bias/-b flag to define the cutoff that distinguishes biallelic from allele-specific expression. The default value is 0.7, which means that at least 70% of the reads from a gene must be expressed from one allele to consider it as imbalanced. Please find all options described in Table 4.

3.1 Input files

- Locus table (.txt format) as given by Allelome.PRO. To run Allelome.LINK the *locus_table.txt* file from the Allelome.PRO v2.0 run is needed.

3.2 Input flags for Allelome.LINK

Table 4 Overview of the Allelome.LINK options.

Required	Description
-input -i	* Input locus_table.txt (as given by Allelome.PRO v2.0)
Optional:	
--name -n	* Sample name (default date and time)
--window-size -w	* Window range to draw links (in kb; default ±100)
--total-reads -r	* Total read cutoff to consider allele-specific genes (default 20)
--allelic-bias -b	* Cutoff to define genes with an allelic bias (default 0.7)

Required	Description
--duplicates -d	* Remove mirrored duplicates (default TRUE)
--output -o	* Output directory (default ./)
Misc:	
--help -h	* Display this help message

3.3 Output files

Allelome.LINK generates a results directory including different output files. The main output files are **links_table.txt** with information about the linked loci and a **.bedpe** file to visualize the links in a genome browser. Please find the detailed information about all output files in Table 5.

Table 5. Overview of the Allelome.LINK output files.

Name	Description
_links_table.txt	Information about the predicted regulatory association between nearby loci
_links_full_table.txt	As _links_table.txt but including the read and allele-specific information
.log	Log file with information about your Allelome.LINK run
/BED_files	
.bed	BED6-file containing all loci passed, color-coded according to their allelic bias
.bedpe	BEDPE file for link representation, color-coded according to mechanism
_repressing.bedpe	BEDPE file for IGV browser visualization, repressive links only
_enhancing.bedpe	BEDPE file for IGV browser visualization, enhancing links only

10.2 Abbreviations: Phenotyping screens

Table 1 Abbreviations: Phenotyping screens

Abbreviation	Name
TEWL	Transepidermal water loss
Calc IFN gamma	Calculated IFN gamma
Calc IL5	Calculated IL5
Calc IL6	Calculated IL6
Calc TNF alpha	Calculated TNF alpha
DisTTot	Distance traveled - Total
NRTot	Number of rears - Total
PcDisCenTot	Percent center distance - Total
PcTiCenTot	Percent time spent in the center - Total
CenPermTi	Center - Permanence time
BN	Startle amplitude - Background Noise
ST110	Acoustic Startle Response at 110 dB
PcPPI_PP67	Percentage Prepulse inhibition - PP 67 dB
PcPPI_PP69	Percentage Prepulse inhibition - PP 69 dB
PcPPI_PP73	Percentage Prepulse inhibition - PP 73 dB
PcPPI_PP81	Percentage Prepulse inhibition - PP 81 dB
PcPPI_Global	Percentage Prepulse inhibition - Global
Conc TNF	Concentration TNF
Conc Insulin	Concentration Insulin
Conc FGF 21	Concentration FGF21
Conc Leptin	Concentration Leptin
LVPWD	Left ventricular posterior wall width in diastole
LVIDd	Left ventricular internal dimension in diastole
LVIDs	Left ventricular internal dimension in systole
IVSd	Interventricular septum in diastole
IVSs	Interventricular septum in systole
LVPWs	Left ventricular posterior wall in systole
Heart rate	Heart Rate
PR	Duration of the PR interval

QRS	Duration of the QRS interval
QTcB	Duration of the QTC interval
EJ fraction	Ejection fraction
Fract shortening	Fractional shortening
QTcM	Heart rate-corrected QT interval (QTcM)
Ca	Calcium concentration in plasma
Chol	Cholesterol concentration in plasma
Fe	Iron concentration in plasma
Gluc	Glucose concentration in plasma
K	Potassium concentration in plasma
LDH	Lactat-dehydrogenase activity in plasma
Phos	Inorganic phosphate concentration in plasma
TP	Total protein concentration in plasma
Trig	Triglyceride concentration in plasma
Urea	Urea concentration in plasma
AP	Alkaline phosphatase activity in plasma
Cl	Chloride concentration in plasma
Na	Sodium concentration in plasma
RBC	Red blood cell count in whole blood
WBC	Total white blood cell count in whole blood
HGB	Hemoglobin concentration in whole blood
HCT	Hematocrit - percentage of cellular components on whole blood
MCV	Mean corpuscular volume
MCH	Mean corpuscular hemoglobin content of erythrocytes
MCHC	Mean corpuscular hemoglobin concentration of erythrocytes
GPT	Alanine aminotransferase (Glutamat pyruvat transaminase) activity in plasma
GOT	Aspartate aminotransferase (Glutamat oxalacetat transaminase) activity in plasma
MPV	Mean platelets volume
PLYM	Percentage of lymphocytes in total white blood cells
PMO	Percentage of monocytes in total white blood cells
RDW	Distribution index of red blood cells

UIBC	Unsaturated iron binding capacity in plasma
Albumin	Albumin concentration in plasma
Ekrea	Creatinine concentration in plasma measured enzymatically
alpha Amylase CNPG3	alpha-Amylase (CNPG3)
Glucose conc	Glucose concentration
Lactat AU400	Lactat concentration in plasma (AU400)
PDW	Calculated distribution width of platelets
PLCR	Platelet large cell ratio
Bw X Ray	Body weight at x-ray analysis
BMC whole mouse	Bone mineral content whole mouse (excluding skull)
BMD whole mouse	Bone mineral density whole mouse (excluding skull)
Lean mass	Lean mass whole mouse (excluding skull)
Fat mass	Fat mass whole mouse (excluding skull)
Axial length l	Axial length left eye
Le fnmv	Left fundus number of main vessels
Le rethi	Left retinal thickness
Fat mass NMR	Fat mass determination at nuclear magnetic resonance

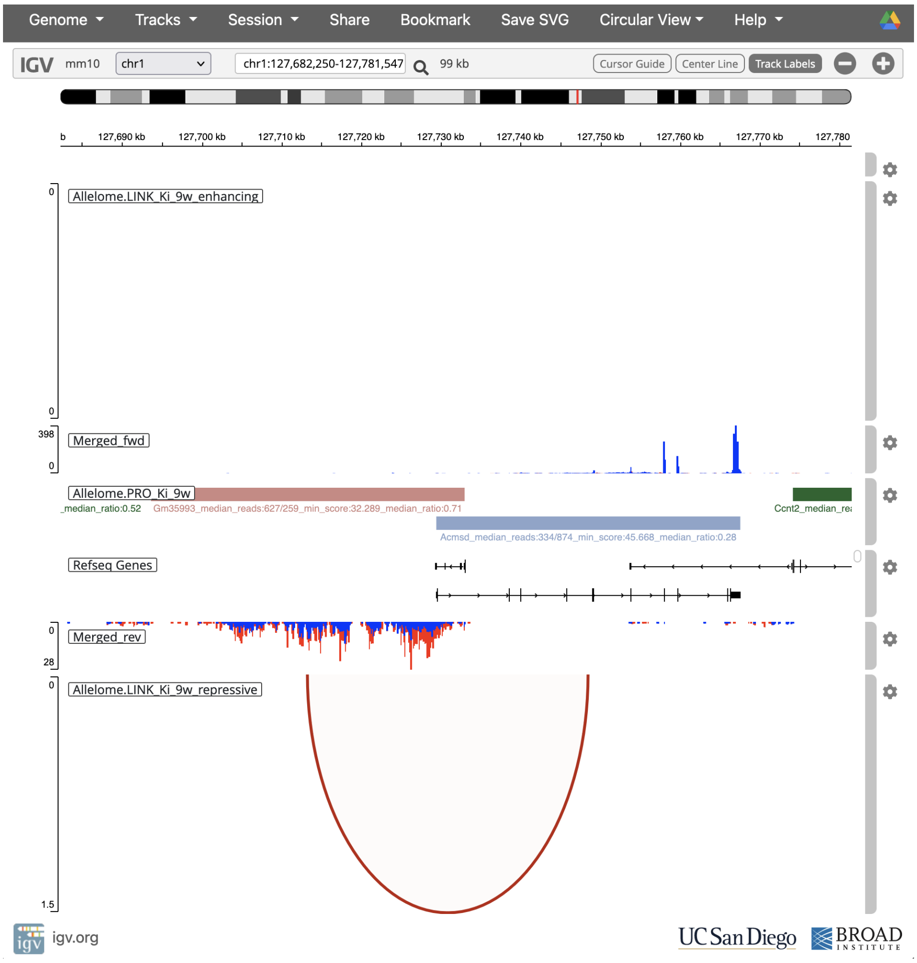
10.3 Resource file: ncRNA-to-target linkages across mice and human samples

Explanation file for Resource: ncRNA-to-target linkages

Allelome.PRO v2.0 Allelome.LINK

To interactively access the resource generated by *Hasenbein et al.*, please click the URL link for the tissue of interest at the resource section of the corresponding GitHub page (https://github.com/AndergassenLab/Allelome.LINK). Candidate predictions are displayed via the Integrative Genomics Viewer (IGV)[1]. A description of the individual tracks can be found in the example section for mice and human.

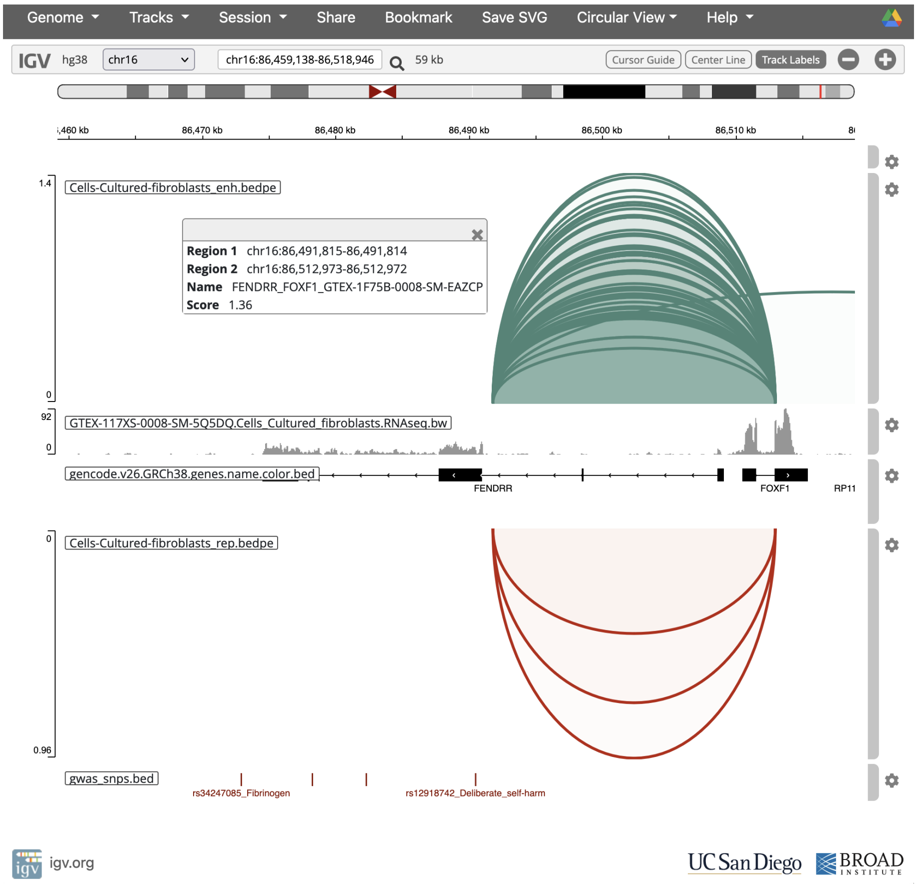
1. Mice example: Kidney



Track	Description

Track	Description
Allelome.LINK_Ki_9w_enhancing	Green arcs show enhancing linkages, where a ncRNA was predicted to have an enhancing function on the linked protein-coding target gene. Clicking on the linkage will name both interaction loci and display the linkage score.
Merged_fwd	RNA-seq track as a representative example. The track was obtained from a single replicate of the three replicates on which the linkage is based. Reads are shown for the forward strand only and are split towards the maternal (red) and paternal (blue) allele.
Allelome.PRO_Ki_9w	Allelome.PRO v2.0 output showing the locus, the median number of reads, the minimum allelic score and the median allelic ratio. The color denotes the allelic bias (red: maternal, green: biallelic, blue: paternal).
Refseq Genes	RefSeq gene annotation
Merged_rev	RNA-seq track as a representative example. The track was obtained from a single replicate of the three replicates on which the linkage is based. Reads are shown for the reverse strand only and are split towards the maternal (red) and paternal (blue) allele.
Allelome.LINK_Ki_9w_repressive	Red arcs show repressive linkages, where a ncRNA was predicted to have a repressive function on the linked protein-coding target gene. Clicking on the linkage will name both interaction loci and display the linkage score.

2. Human GTEx example: Cells - cultured fibroblasts



Region 1 chr16:86,491,815-86,491,814
Region 2 chr16:86,512,973-86,512,972
Name FENDRR_FOXF1_GTEX-1F75B-0008-SM-EAZCP
Score 1.36

Track	Description
Cells-Cultured-fibroblasts_enh	Green arcs show enhancing linkages, where a ncRNA was predicted to have an enhancing function on the linked protein-coding target gene. Clicking on the linkage will name both interaction loci and display the linkage score, along with the GTEx sample where the linkage was found. Different arcs represent different samples.
RNAseq.bw	RNA-seq track as a representative example. The track was obtained from a single sample of the tissue on which the linkages are based on.
annotation.bed	GENCODE v26 GRCh38 annotation
Cells-Cultured-fibroblasts_rep	Red arcs show repressive linkages, where a ncRNA was predicted to have a repressive function on the linked protein-coding target gene. Clicking on the linkage will name both interaction loci and display the linkage score, along with the GTEx sample where the linkage was found. Different arcs represent different samples.
gwas_snps.bed	GWAS SNPs derived from the GWAS catalog [2]

[1]: *James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. Nature Biotechnology 29, 24–26 (2011). A public access*

version is also available: PMC3346182.

[2]: *Sollis, E. et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res 51, D977-d985 (2023).*

10.4 Abbreviations: GTEx tissues

Table 2 Abbreviations: GTEx tissues

Abbreviation	Name
ADPSBQ	Adipose - Subcutaneous
ADPVSC	Adipose - Visceral Omentum
ADRNLG	Adrenal Gland
ARTAORT	Artery - Aorta
ARTCRN	Artery - Coronary
ARTTBL	Artery - Tibial
BLDDER	Bladder
BREAST	Breast - Mammary Tissue
BRNACC	Brain - Anterior cingulate cortex BA24
BRNAMY	Brain - Amygdala
BRNCDT	Brain - Caudate basal ganglia
BRNCHA	Brain - Cerebellum
BRNCHB	Brain - Cerebellar Hemisphere
BRNCTXA	Brain - Cortex
BRNCTXB	Brain - Frontal Cortex BA9
BRNHPP	Brain - Hippocampus
BRNHPT	Brain - Hypothalamus
BRNNCC	Brain - Nucleus accumbens basal ganglia
BRNPTM	Brain - Putamen basal ganglia
BRNSNG	Brain - Substantia nigra
BRNSPC	Brain - Spinal cord cervical c1
CLNSGM	Colon - Sigmoid
CLNTRN	Colon - Transverse
CVSEND	Cervix - Endocervix
CVXECT	Cervix - Ectocervix
ESPGEJ	Esophagus - Gastroesophageal Junction
ESPMCS	Esophagus - Mucosa
ESPMSL	Esophagus - Muscularis
FIBRBLS	Cells - Cultured fibroblasts

FLLPNT	Fallopian Tube
HRTAA	Heart - Atrial Appendage
HRTLV	Heart - Left Ventricle
KDNCTX	Kidney - Cortex
KDNMDL	Kidney - Medulla
LCL	Cells - EBV-transformed lymphocytes
LIVER	Liver
LUNG	Lung
MSCLSK	Muscle - Skeletal
NERVET	Nerve - Tibial
OVARY	Ovary
PNCREAS	Pancreas
PRSTTE	Prostate
PTTARY	Pituitary
SKINNS	Skin - Not Sun Exposed Suprapubic
SKINS	Skin - Sun Exposed Lower leg
SLVRYG	Minor Salivary Gland
SNTTRM	Small Intestine - Terminal Ileum
SPLEEN	Spleen
STMACH	Stomach
TESTIS	Testis
THYROID	Thyroid
UTERUS	Uterus
VAGINA	Vagina
WHLBLD	Whole Blood

10.5 Confirmation of publication and licensing rights

49 Spadina Ave. Suite 200
Toronto ON M5V 2J1 Canada
www.biorender.com

Confirmation of Publication and Licensing Rights

January 7th, 2025

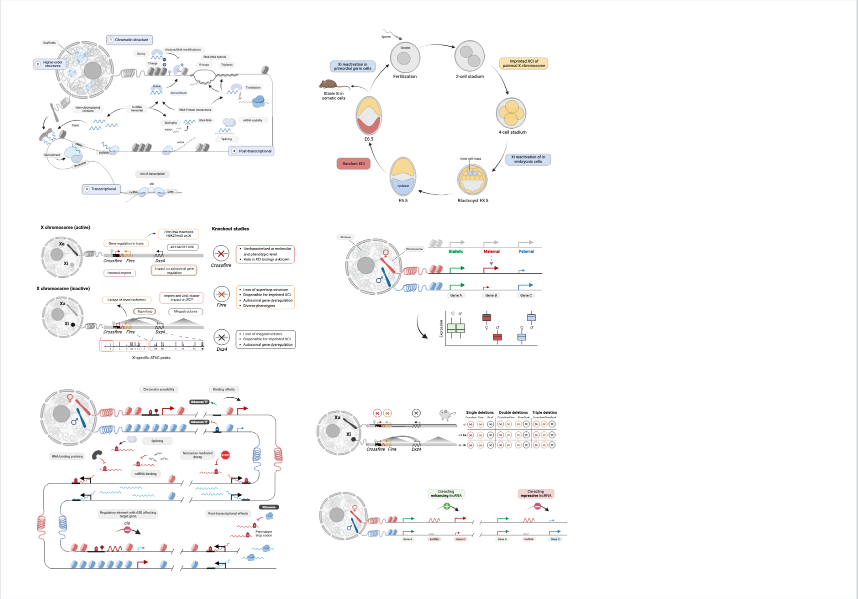
Subscription Type:	*Individual - Academic*
Agreement number:	*WD27RF8DST*
Publisher Name:	*Thesis*
Citation to Use:	*Created in BioRender. Andergassen, D. (2025) https://BioRender.com/b98a895*

To whom this may concern,

This document is to confirm that Daniel Andergassen has been granted a license to use the BioRender Content, including icons, templates, and other original artwork, appearing in the attached Completed Graphic pursuant to BioRender's Academic License Terms. This license permits BioRender Content to be sublicensed for use in publications (journals, textbooks, websites, etc.).

All rights and ownership of BioRender Content are reserved by BioRender. All Completed Graphics must be accompanied by the following citation: "Created in BioRender. Andergassen, D. (2025) https://BioRender.com/b98a895".

BioRender Content included in the Completed Graphic is not licensed for any commercial uses beyond use in a publication. For any commercial use of this figure, users may, if allowed, recreate it in BioRender under an Industry BioRender Plan.



For any questions regarding this document, or other questions about publishing with BioRender, please refer to our BioRender Publication Guide, or contact BioRender Support at support@biorender.com.