# Identifying the Role of Transcription Factor RFX3 in 9P Deletion Syndrome

by

Lilly Edwards

B.S. Computer Science and Molecular Biology, MIT, 2024

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE AND MOLECULAR BIOLOGY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2025



| | |
|---|---|
| Authored by: | Lilly Edwards<br>Department of Electrical Engineering and Computer Science<br>January 24, 2025 |
| Certified by: | Timothy Yu<br>Associate Professor of Pediatrics, Harvard Medical School, Thesis Supervisor |
| Certified by: | Eunjung Alice Lee<br>Associate Professor of Pediatrics, Harvard Medical School, Thesis Supervisor |
| Certified by: | Manolis Kellis<br>Professor of Computer Science, MIT, Thesis Supervisor |
| Accepted by: | Katrina LaCurts<br>Chair<br>Master of Engineering Thesis Committee |

# Identifying the Role of Transcription Factor RFX3 in 9P Deletion Syndrome

by

Lilly Edwards

Submitted to the Department of Electrical Engineering and Computer Science
on January 24, 2025 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN COMPUTER SCIENCE AND MOLECULAR BIOLOGY

## ABSTRACT

9p deletion (9p-) syndrome is primarily characterized by intellectual disability, developmental delays, and autism. This project investigated how much of the neuronal phenotypes of 9p- syndrome could be attributed to RFX3, a transcription factor and autism risk gene. Bulk RNA-seq data of iPSC-derived neurons from patients with 9p- syndrome and CRISPR-engineered cell lines was analyzed using Principal Component Analysis, Differential Gene Expression analysis, and Functional Enrichment analysis. The findings indicate that RFX3 plays a significant role but is not the sole driver of the neuronal phenotypes. SMARCA2, a gene linked to intellectual disability and part of the SWI/SNF complex, was identified as a direct target of RFX3 in the commonly deleted region of chromosome 9p. Notably, the combined deletion of RFX3 and SMARCA2 led to greater dysregulation of SMARCA2 expression and SWI/SNF complex components than the deletion of either gene alone. These findings highlight the potential synergistic effects of RFX3 and SMARCA2 in 9p- syndrome and suggest their combined disruption may underlie the neuronal phenotypes observed.

Thesis supervisor: Timothy Yu
Title: Associate Professor of Pediatrics, Harvard Medical School

Thesis supervisor: Eunjung Alice Lee
Title: Associate Professor of Pediatrics, Harvard Medical School

Thesis supervisor: Manolis Kellis
Title: Professor of Computer Science, MIT

# Acknowledgments

I would like to thank Dr. Timothy Yu, Dr. Alice Lee, and Dr. Aikaterini Chatzipli for their guidance and advice as I researched this project.

I would also like to thank my family- Robyn, Jason, and Paige Edwards- for their support and encouragement this past year!

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

9p deletion (9p-) syndrome was first identified in 1973 by Alfi et al [1]. In 1963, Alfi et al described the syndrome's phenotypes based on six patients, which included intellectual disability, trigonocephaly, and distinctive facial features [2]. Trigonocephaly is when a patient has a premature fusion of a joint between the bones in their forehead, resulting in a triangularly shaped forehead [3] As the number of patients with the syndrome has grown, the set of phenotypes associated with 9p- syndrome has grown and solidified. A patient may have any subset of these symptoms, and some are apparent at diagnosis while others manifest over time.

9p deletion (9p-) syndrome is caused by a deletion of DNA on the short arm of chromosome 9, hence 9p. There are a total of 553 genes on this arm, and 207 of them are protein-coding [4]. Diagnosis requires identification of a deletion in the chromosome using either karyotyping, chromosomal microarray, or exome/genome sequencing [5]. The deletions often include multiple genes, and result in varying phenotypes [6]. It is also an ultra-rare disorder, resulting in limited data for analysis and informing medical care. While much work has been done to characterize the symptoms of the syndrome and potential genes involved, there remains the significant challenge of mapping which gene(s) are responsible for which symptoms. There is also no treatment for this disorder.

A study by Starosta et al. (2024) recruited 48 individuals with 9p- syndrome and extracted phenotypic data, providing insight into symptom prevalence and advice on management. From dedicated imaging and a physical exam at diagnosis, craniofacial abnormalities such as trigonocephaly (43.9%) and a cleft palate (18.9%) can be identified as well as dysmorphic features (89.3%) [5]. The patient can be referred for plastic surgery or neurosurgery depending on symtpom severity. Congenital heart disease (43.7%) can be identified from an echocardiogram at diagnosis, and patients can be referred to pediatric cardiology depending on the severity of their specific heart defect. Genitial defects (76.2% frequency in males and 22.2% frequency in females) and congenital abnormalities of the kidneys and urinary tract (21.2%) would also be apparent at birth. During the neonatal period, infants may develop neonatal hypoglycemia (36.3%) which requires stringent blood glucose monitoring. Other symptoms may be apparent at diagnosis or appear during childhood, and require routine monitoring. Developmental symptoms would become apparent as the child grows older, and managed through ancillary therapy services and a neuropsychology specialist. Such symptoms include global developmental delays (such as delays in speaking or crawling) [6] and

intellectual disability (97.9%) hypotonia (86.9%), autism spectrum disorder (ASD) (53.4%) [5]. Neurological issues including seizures (29.1%) would require working with a neurologist and taking anti-convulscent medications. Gastrointestinal issues such as chronic constipation (65.9%) and gastroesophoageal reflux disease (52.1%), and abdominal wall defects (34.1%) would require a strict diet and medication. Vision problems (54.3%), hearing loss (19.1%), and sleep apnea (36.1%) are also common but manageable symptoms. Later in childhood, psychiatric disorders (67.4%) such as Attention-Deficit/Hyperactivity Disorder (ADHD), mood disorders, and tic disorders would become apparent and result in a referral to psychiatry as necessary. Also, atopic diseases (58.9%) such as allergic rhinitis, eczema, and asthma may develop [5].

There has been some progress tying genes in this region to 9p- syndrome phenotypes. The region of common deletions has been narrowed to specific bands of the chromosome. Chromosome bands are regions of a chromosome that stain a certain way during testing, and are used to denote where a gene is in a chromosome. Bands are numbered so that they increase as you move away from the centromere (where the short and long arms of the chromosome meet) their numbering increases[7]. 38.1% of deletions occur in the terminal band, 9p24, 17.8% of deletions occur in the next band, 9p23, and 34.6% occur in the next band, 9p22 [6]. In this commonly deleted region of bands 9p22, 9p23, and 9p24, there are 185 genes including 63 protein-coding genes.

Within this region, some genes have been tied to specific phenotypes. Deletion of FREM1 has been identified as the driving deletion for trigonocephaly in the syndrome. There are cases of trigonocephaly even when FREM1 is not included in the deletion, but the other genes linked to this phenotype are lower-risk. Deletion of DMRT1, DMRT2, and DMRT3 is associated with female or ambiguous genitalia in people with an XY pair of sex chromosomes [5]. It has also been found that six genes (CDC37L1, NFIB, PTPRD, RFX3, SMARCA2, UHRF2) with pLI $>$ 0.9 in this commonly deleted region are enriched for deletion in individuals with neurodevelopmental disorders[6]. A pLI score for a gene is a measure of how likely a loss-of-function mutation in that gene will have phenotypic consequences; the score ranges between 0 and 1 [8] [6]. These six genes are strong candidates for contributing to the neurodevelopmental symptoms associated with 9p deletion syndrome– global developmental delays, intellectual disability, and autism– but it remains to be seen to what degree each gene contributes to these symptoms.

In this pilot analysis we chose to focus on RFX3, located in the 9p24 band. A clinical study by Harris et al. (2021) of 14 patients with RFX3 haploinsufficiency linked this genetic abnormality with autism, intellectual disability of varying severity, global developmental delays, and ADHD. These neurodevelopmental disorders and ADHD are all common symptoms of 9p- syndrome. Supporting this link is the finding that RFX3 expression is enriched in the human fetal cortex, and that its binding motifs are enriched in regulatory regions for autism risk genes [9].

These findings are further supported by RFX3's known role as a key transcriptional regulator. The RFX family of transcription factors are highly evolutionarily conserved and are present in all animals, which indicates their critical role in cellular function. Gene targets of RFX transcription factors have a DNA sequence called an X-box motif where the transcription factors can bind, and these motifs are also evolutionarily conserved. RFX transcription factors have a known role activating the expression of ciliary genes [10] [11].

When both copies of RFX3 were deleted in mice, embryonic ciliary gene expression is disrupted leading to damaged brain development [12]. This result was reproduced in human induced pluripotent stem cells (iPSCs) that were differentiated into neurons with both copies of RFX3 deleted [13].

However, when only one copy of RFX3 is deleted a distinct set of genes are dysregulated. In iPSC derived neurons haploinsufficient for RFX3, there was down regulation of genes involved in synapse assembly, chemical synaptic transmission, and glutamate receptor signaling. Additionally, genes downregulated in neurons with either one or two copies of RFX3 deleted were enriched for autism risk genes. Combined with additional profiling, this confirmed that RFX3 functions as a transcriptional activator of synaptic genes. RFX3 binding sites near both ciliary genes and synaptic genes were evolutionarily conserved, indicating its significance in brain development and function [13]. However, the relative contribution of RFX3 in the context of other gene deletions, particularly 9p deletion has not been studied.

Because one copy of RFX3 is commonly deleted in 9p- syndrome and RFX3 haploinsufficiency causes similar symptoms to those of 9p- syndrome, we sought to understand how RFX3 contributes to the neurologic phenotypes of this syndrome. To address these questions, we utilized neuronal cultures derived from both patient and PGP1 human induced pluripotent stem cells (iPSCs). The patient-derived iPSC lines (9P126, 9P117, and 9P141) exhibit varying sizes of chromosomal 9p deletions. Two of the lines include RFX3 in the deletion and one does not, allowing me to identify transcriptional changes in 9p- syndrome when RFX3 is present and absent (Table 1). In parallel, we grew isogenic control lines from the PGP1 background, with heterozygous (HET) and homozygous (KO) loss-of-function variants of RFX3 created using CRISPR-Cas9 editing to introduce frameshift mutations upstream of the DNA-binding domain. These lines were previously generated by Lai et al. (2023) [13] and provide a controlled system to compare the patient transcriptional profiles with the specific impact of RFX3 haploinsufficiency. Neuronal cultures were differentiated using Ngn2 induction, and RNA was harvested at day 14, a stage when cells exhibit transcriptomes characteristic of neural progenitor cells, allowing us to capture early neurodevelopmental changes.

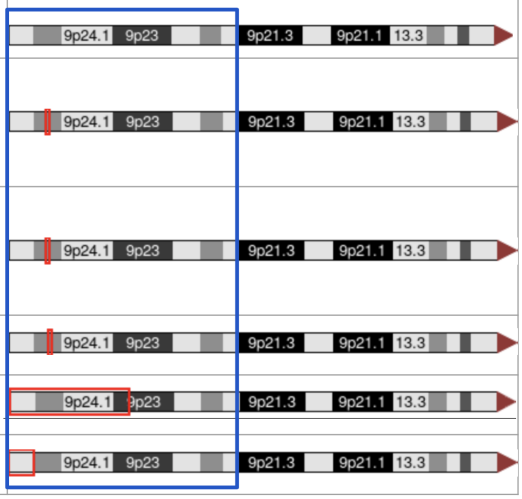
I hypothesized that RFX3 would play one of three roles: that it would be the main gene of consequence, that it worked in conjunction with other genes on chromosome 9p, or that it did not contribute to neuronal changes associated with this syndrome. I present evidence that RFX3 significantly contributes to transcriptional dysregulation associated with 9p- syndrome. I also provide evidence that this dysregulation can be tied to the autism-related features of 9p deletion syndrome. I also explore interactions between RFX3 and other deleted genes in the interval to understand potential synergistic deletions in this syndrome.

# Chapter 2

# Results

## 2.1 Identifying Evidence of the Transcriptional Signature of RFX3 in 9P Deletion Syndrome

To explore the role of RFX3 in 9p- syndrome, we obtained iPSC lines generated by a collaborator at Washington University from three patients with 9p- syndrome (9P126, 9P117, and 9P141) [5]. Each patient had varying sizes of deletions, described in Table 1a. For comparison, we generated iPSC lines in our laboratory bearing RFX3 mutations. These lines are all generated from the PGP1 iPSC line, and are therefore isogenic except for their mutations in RFX3. CRISPR-Cas9 was used to engineer heterozygous (HET) and homozygous (KO) loss-of-function variants of RFX3 in a control (WT) PGP1 iPSC line. These loss-of-function variants were generated using frameshift mutations in exon 5, located upstream of the DNA-binding domain [13]. Specifically, the RFX3 HET line harbored a heterozygous c.313insN mutation, while the RFX3 KO line contained a homozygous c.335insN mutation. Ngn2 induction was used to generate neuronal cultures of these six lines [14]. Gene expression patterns were profiled at day 14 (neural progenitor cells) to identify differentially expressed genes (FDR-adjusted $p<0.05$). The location of the deletions for each cell line is shown in Table 1a, and the RFX3 copy number of each cell line is shown in Table 1b. The boxed area in the Deleted Region column depicts the region with about 90% of deletions in 9p- syndrome [6].

| Cell Line | Alteration Type | Region (BP) | Deleted region |
|---|---|---|---|
| Wildtype | NA | NA |  |
| RFX3 Het | Heterozygous CRISPR-Cas9 engineered frameshift c.313insN | RFX3 Exon 5 | |
| RFX3 KO | Homozygous CRISPR-Cas9 engineered frameshift c.335insN | RFX3 Exon 5 | |
| 9P126 | Heterozygous Deletion | 3,443,917-3,726,606 | |
| 9P117 | Heterozygous Deletion | 1-10,327,830 | |
| 9P141 | Heterozygous Deletion | 1-2,123,411 | |



| Cell Line | Genetic Background | RFX3 copy Number | Number of Protein-coding Genes Deleted | Number of Genes Deleted |
|---|---|---|---|---|
| Wildtype | PGP1 | 2 | 0 | 0 |
| RFX3 Het | PGP1 | 1 | 1 | 1 |
| RFX3 KO | PGP1 | 0 | 1 | 1 |
| 9P126 | Patient 126 | 1 | 1 | 1 |
| 9P117 | Patient 117 | 1 | 40 | 112 |
| 9P141 | Patient 141 | 2 | 9 | 26 |

Table 1: (a) Table of the regions deleted or mutated in each cell line (b) Table of the number of disrupted genes in each cell line

To understand the similarities and differences of the transcriptional profiles of the six cell lines analyzed, I performed Principal Component Analysis (PCA) and visualized the first two principal components using dimensionality reduction (Fig. 1). The first two principal components explain the majority of the variance in the data. We found that the cell lines were grouped primarily according to their genetic background (i.e., whether they were generated

from the PGP1 genetic background or not). Other combinations of principal components are unable to create a meaningful clustering (Sup. Fig. 9). The clustering pattern observed in Fig. 1 has two interpretations: 1) that the individual genetic backgrounds of patient-derived cell lines and PGP1 cell lines was overwhelming any signal caused by RFX3 haploinsufficiency or 2) that the transcriptional dysregulation in the 9p- lines is not strongly linked to the loss of RFX3. If the transcriptional dysregulation in patient lines was strongly linked to RFX3 loss, we would expect to see 9P126, which is the patient sample with only RFX3 deleted, cluster near RFX3 HET. In this plot, they are very far apart.

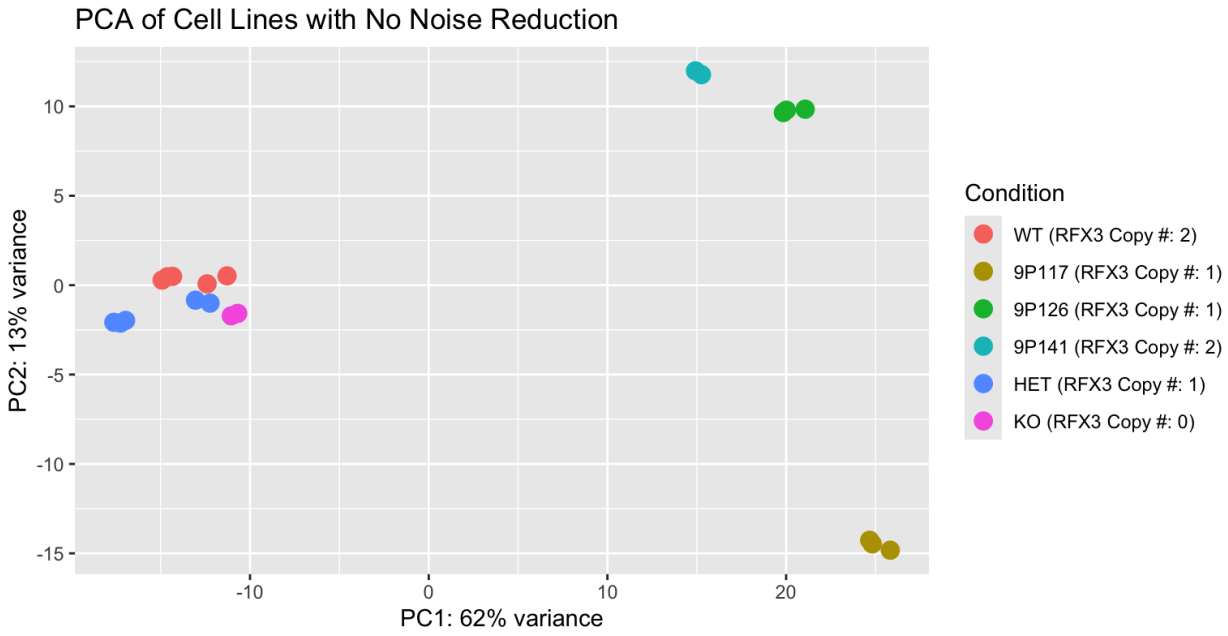
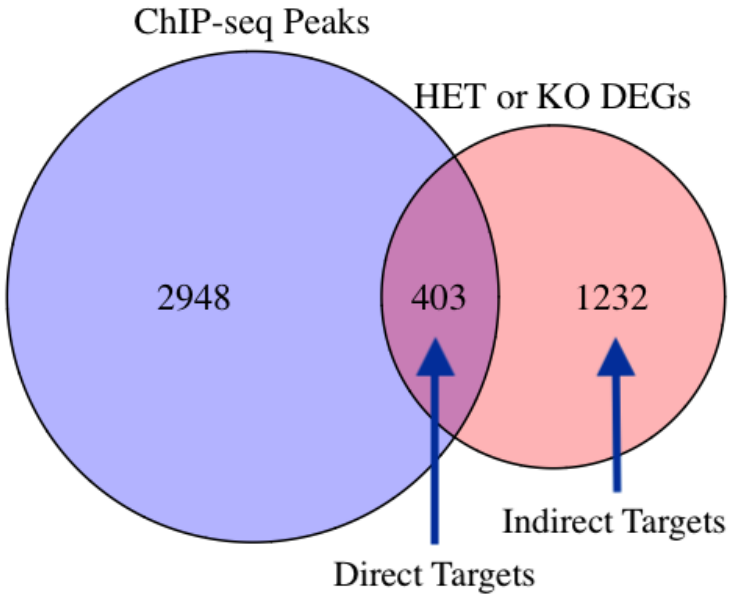



Figure 1: Principal Component analysis of all cell lines with all genes

To examine whether neurons from 9p- patients that include RFX3 in the deletion have transcriptomes consistent with RFX3 disruption, first I identified genes whose transcription is directly modified by RFX3. RFX3 is a transcription factor that is known to activate its target genes. Cut&Run data previously generated and analyzed by our group identified 4,024 RFX3 binding peaks present in WT neurons and absent in RFX3 KO neurons [13]. These binding sites were annotated to genomic regions using ChIPseeker, revealing that RFX3 binds within promoters, introns, distal intergenic regions, exons, and 3' UTRs. RFX3 binding sites occurred within 3 million bp of the transcription start site. From this data, I identified 3,351 genes associated with RFX3 binding, with sites located upstream, downstream, or within the genes themselves, suggesting potential roles in both transcriptional regulation and regulatory interactions. From previously analyzed bulk RNA-seq data, I identified genes that were significantly down-regulated when one or two copies of RFX3 were deleted [13]. The 403 genes included in both groups are "direct targets" of RFX3 (Fig. 2a).

Conducting PCA using only these 403 direct target genes revealed that 9p- patient cell lines exhibit RFX3 transcriptional signatures that correlate with their RFX3 copy number (Fig. 2). PC1 can be interpreted as differentiating cell lines based on their genetic background because all of the cell lines generated from PGP1 cells are on the left side of the y-axis and all of the cell lines generated from 9p- patients are on the right side. PC2 can be interpreted as differentiating cell lines based on RFX3-related transcriptional dysregulation by looking at the PGP1 cell lines. The WT samples with 2 copies of RFX3 cluster below the x-axis, HET samples with one copy of RFX3 cluster around the x-axis and KO samples with 0 copies of RFX3 cluster above the x-axis. The distribution of the patient cell lines along the y-axis (Fig. 2b) indicates that there is transcriptional disruption caused by RFX3 haploinsufficiency in the patient cell lines. The patient cell lines with one copy of RFX3 (9P117 AND 9P126) tend to cluster with RFX3 HET along the y-axis, and the patient cell line with both copies of RFX3 (9P141) tend to cluster along the y-axis as WT. This indicates that in Fig. 1, there were transcriptional changes caused by RFX3 haploinsufficiency that were overwhelmed by the differences in genetic background between the patient and PGP1 cell lines.



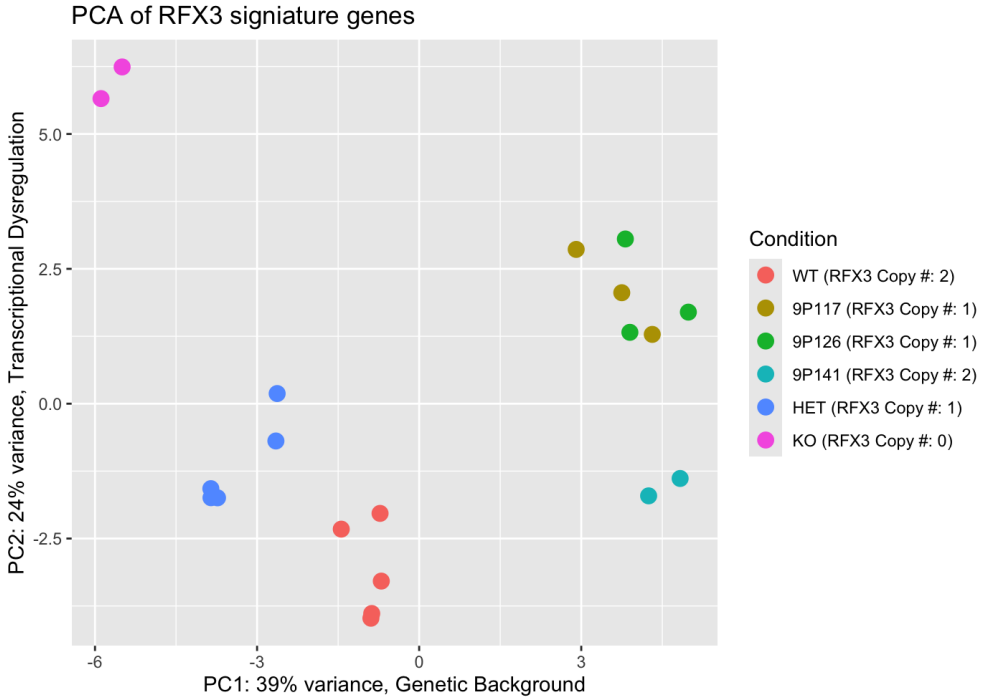




Figure 2: (a) Venn Diagram Depicting RFX3 Direct Target Selection and (b) PCA using only the 403 RFX3 Direct Target Genes.

## 2.2 RFX3 Significantly Contributes to Global Transcriptomic Disruption

Given that 9p- samples appear to cluster based on their RFX3 copy number when using direct targets of RFX3, we hypothesized that the lack of clustering based on RFX3 copy number using all genes (Fig. 1) is due to excessive "noise" from genetic background in the data. This "noise" may also come from genes whose expression is highly variable even in the absence of any genetic perturbation. Standard removal of gene with fewer than 10 transcripts across all samples resulted in 29,761 expressed genes. Noisy gene filtering identified 1,923 genes which were removed (4.58% of all genes). The noisy gene filtering process had 3 components: removing genes with high variability due to 1) differences in genetic backgrounds (9P126 vs PGP1 RFX3 HET), 2) technical processing differences, and 3) sex-specific effects. To identify genes that vary due to genetic background, 9P126 was compared with PGP1 RFX3 HET because both cell lines are haploinsufficient for RFX3 but have different genetic backgrounds. To identify genes that vary due to technical processing differences, WT samples grown and harvested by two different people were compared to each other to calculate DEGs, and RFX3 HET samples grown and harvested by two different people were compared to each other to calculate DEGs. To identify which genes had high variability for the first two components, DESeq2 was used and genes were removed based on their significance and log2 Fold Changes (FDR-adjusted p-value $< 0.05$). Log2 Fold Change (Log2FC) quantifies changes in gene expression between two conditions as the base-2 logarithm of the expression ratio, where positive values indicate upregulation, negative values indicate downregulation, and 0 indicates no change. The magnitude of the Log2FC needed to be over 3 for differences in genetic backgrounds and over 1 for technical processing differences. The threshold of 3 for differences due to genetic background was the highest threshold where the variance explained by PC1 fell below 50%. The threshold of 1 was used for technical processing differences so that WT and RFX3 HET samples processed by different people were still clustered together. To control for sex-specific variability, genes on the Y chromosome and XIST were removed. XIST is a gene only expressed in females to inactivate one copy of the X chromosome [15]. Other genes on the X chromosome were retained because the X chromosome contains direct targets of RFX3 and its genes are expressed in both males and females. Controlling for sex-related variability in this manner was necessary as the WT line is male, but 9P117 is female. These pre-processing steps significantly reduced PC1 variance from 62% to 48%, facilitating clearer clustering and improved data analysis. This filtered set of 28,052 genes was used for all subsequent analysis. Noisy Gene filtering reduced the number of direct RFX3 target genes from 403 to 356, and these 356 genes were used for all subsequent analysis.

PCA results show that the data cleaning steps were effective at reducing genetic background "noise" in the data, allowing for identification of the effects of the genetic deletions in each cell line. The PGP1 cell lines (WT, HET, and KO) now cluster separately based on their distinct genetic perturbations. 9P126, the patient cell line with only RFX3 deleted (green), clusters close to PGP1 RFX3 HET (blue) as expected for two cell lines that both have one copy of RFX3 deleted.

PCA results also showed that the 9P- cell lines cluster somewhat based on RFX3 copy number. PC1 on the x-axis still represents transcriptional differences due to genetic back-

ground between PGP1 cell lines and 9P- patient cell lines. PC2 on the y-axis still represents transcriptional dysregulation similar to the PC2 of PCA using RFX3 direct target genes only. The fact that PC2, which explains the second-most variance in the data, still stratifies samples based on RFX3 copy number indicates that RFX3 is a significant factor leading to transcriptional dysregulation in both direct targets of RFX3 and other genes downstream of those targets.

Figure 3 presents some interesting relationships between the cell lines for further exploration. 9P117 (yellow, patient line with a large deletion including RFX3) clusters closer to RFX3 KO (pink) than 9P126 (green, RFX3 only deletion) does. This suggests that 9P117 has transcriptional dysregulation that is somewhat similar to that seen when two copies of RFX3 are deleted. This could suggest that another gene deleted in 9P117 may act synergistically with RFX3. This potential synergy could exacerbate the effects of RFX3 haploinsufficiency, making the transcriptional profile of 9P117 more similar to RFX3 KO. Furthermore, the observation that 9P117 clusters far from 9P141, despite both lines sharing 79 deleted genes, highlights the significant impact of having two intact copies of RFX3 in 9P141. These findings emphasize the role of RFX3 in transcriptional regulation and suggest that the loss of additional genes in 9P117 amplifies transcriptional disruptions, potentially contributing to more severe phenotypes.

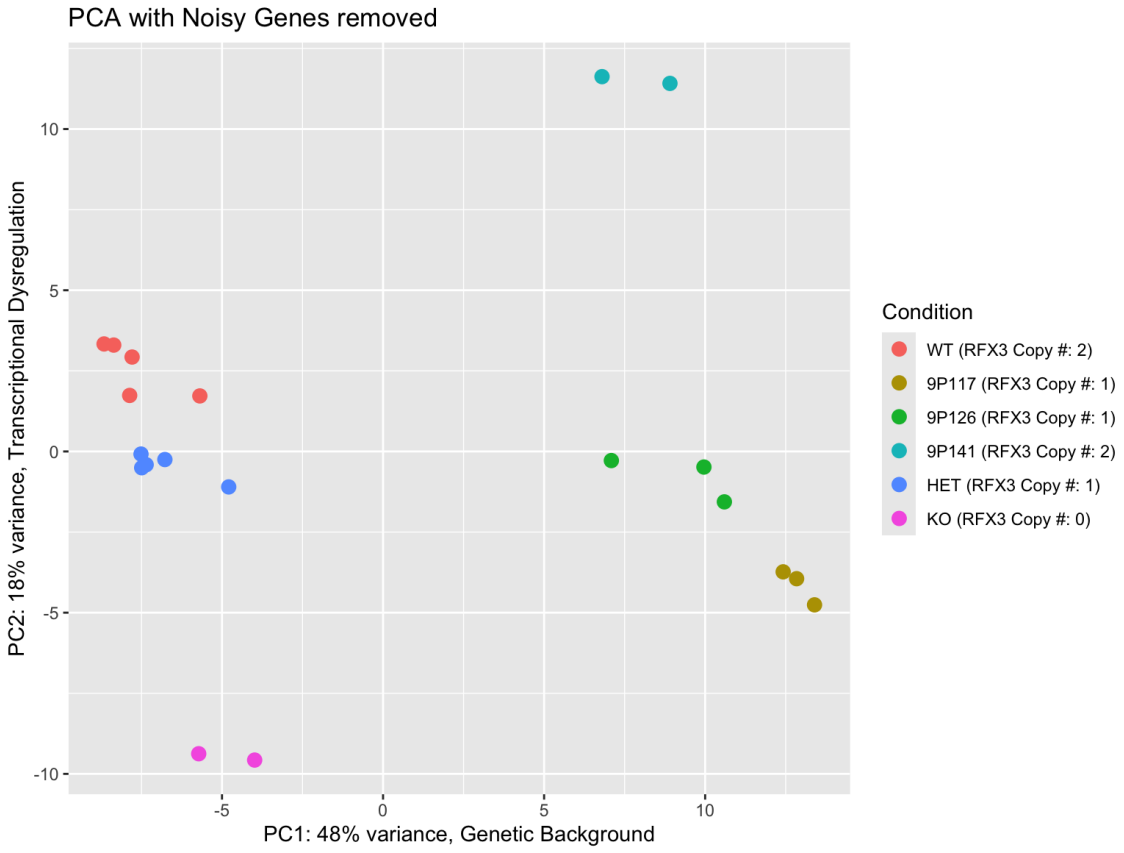
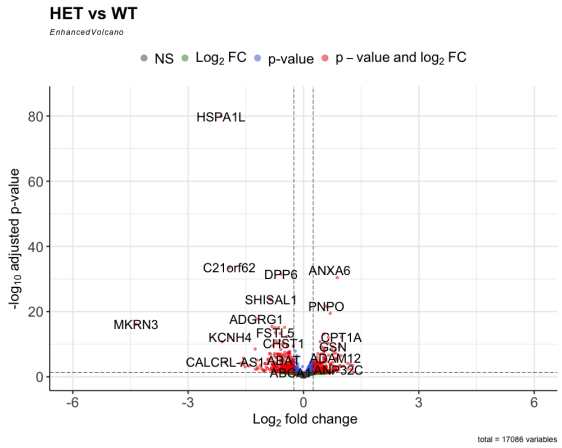


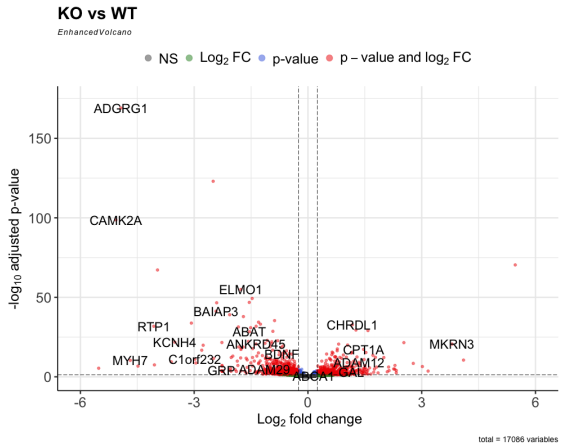

Figure 3: Principal Component Analysis of all cell lines with noisy genes removed

The volcano plots depict the total number of significant Differentially Expressed Genes

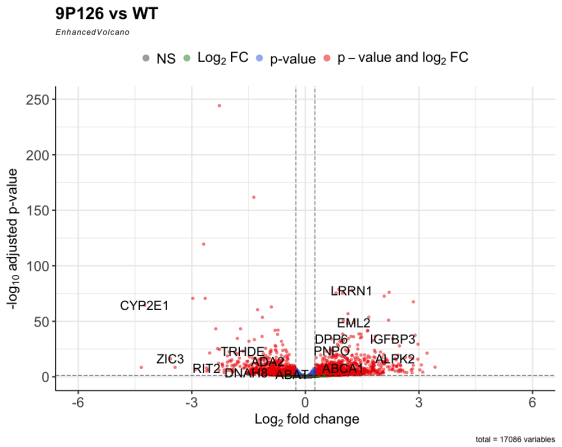
(DEGs) for each cell line relative to RFX3 WT. DEGs were calculated after noisy genes were removed. A cutoff of Log2FC $> 0.25$ or Log2FC $< -0.25$ was used to retain all genes with potential biological significance. A base mean expression of over 30 was used to ensure genes were sufficiently expressed to calculate meaningful Log2FCs and p-values. There were 617 DEGs for RFX3 HET, 1,273 DEGs for RFX3 KO, 4,024 DEGs for 9P117, 3,103 DEGs for 9P126, and 1,963 DEGs for 9P141 (FDR-adjusted p-value $< 0.05$). To confirm these DEGs represent true decreased expression, I determined if protein-coding genes deleted in a cell line appeared in that line's down regulated DEGs. 9P117 has 40 protein-coding genes deleted, and 29 of them are DEGs. Of the deleted genes that are not down-regulated, only one of these genes (DOCK8) was eliminated in noisy gene filtering, and this was an appropriate removal as the gene is not expressed in neurons [16]. Three expressed fewer than ten transcripts across all twenty samples. The remaining seven did not meet the requirements for significant differential expression although five of them were downregulated. 9P141 has 9 protein-coding genes deleted, and 9 of them are significantly downregulated. Of the deleted genes that are not down-regulated in 9p141, DOCK8 was still eliminated during noisy gene filtering and DMRT1 was too poorly expressed to calculate its differential gene expression. RFX3 is significantly downregulated in 9P126, and that is the only gene deleted in that line. Exploring how many deleted genes are down-regulated confirms that the data-preprocessing steps effectively removed genes not biologically relevant for this analysis. Further confirming this point, CBWD1 is a gene deleted in 9P117 and 9P141 known to be expressed in the human brain [17]. CBWD1 appears as a highly significant and strongly downregulated gene in the volcano plots for these cell lines.



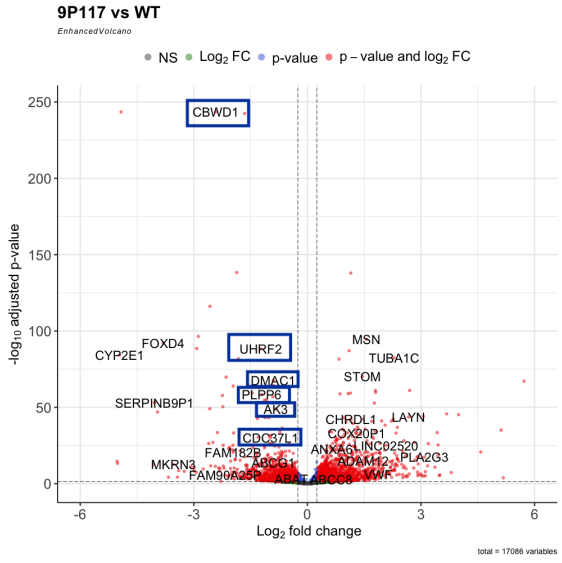

(a) HET vs WT



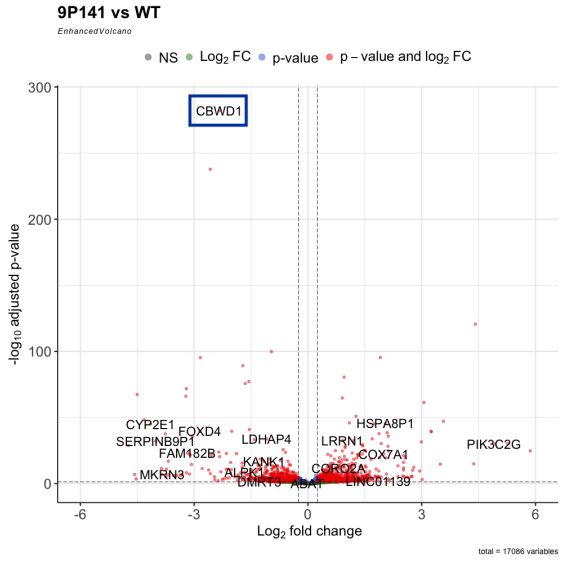

(b) KO vs WT




(c) 9P126 vs WT




(d) 9P117 vs WT



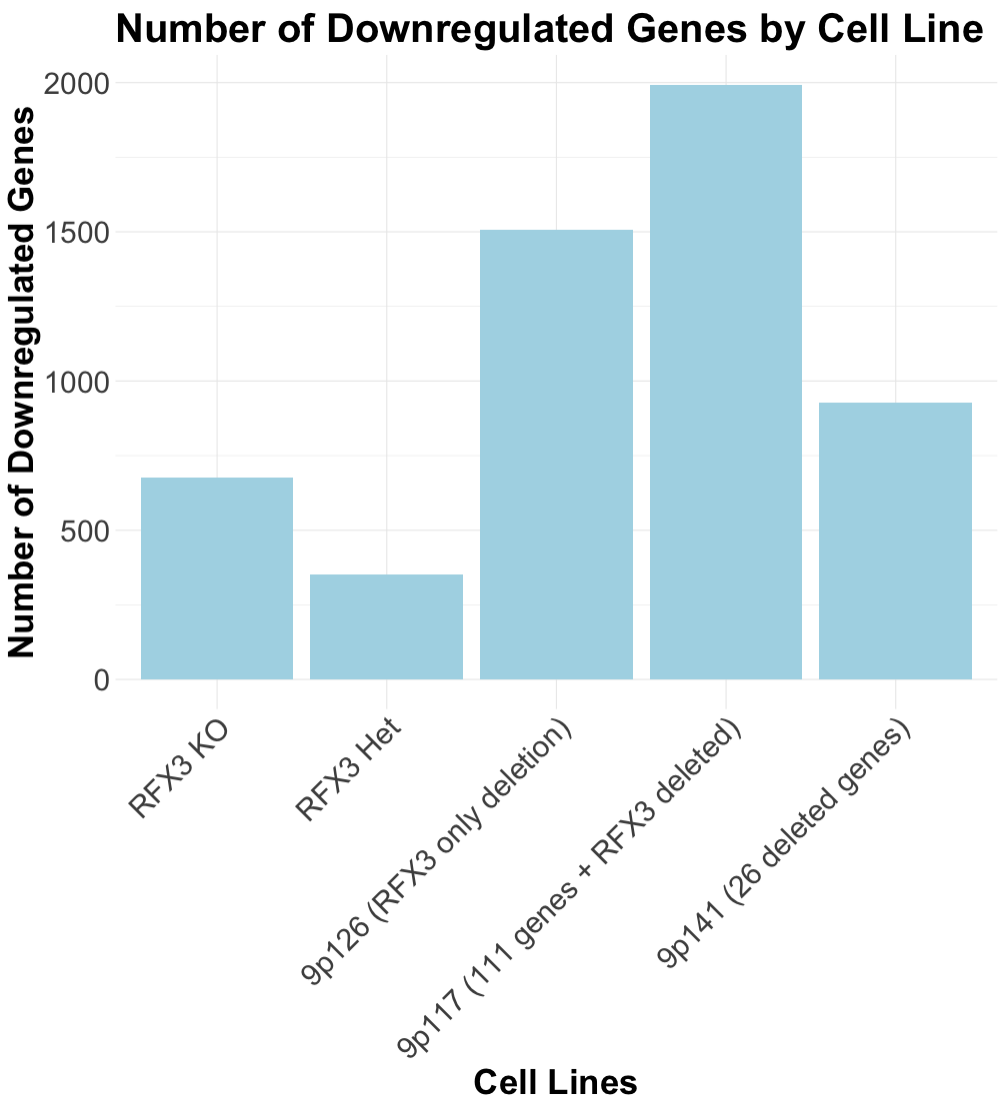

(e) 9P141 vs WT

Figure 4: Volcano plots for all 5 cell lines. Adjusted p-value $< 0.05$ and Log2FC $> 0.25$ or Log2FC $<$ -0.25. RFX3 HET DEGs are labeled in each plot. Boxed genes are deleted in that cell line, but are not RFX3 HET DEGs.

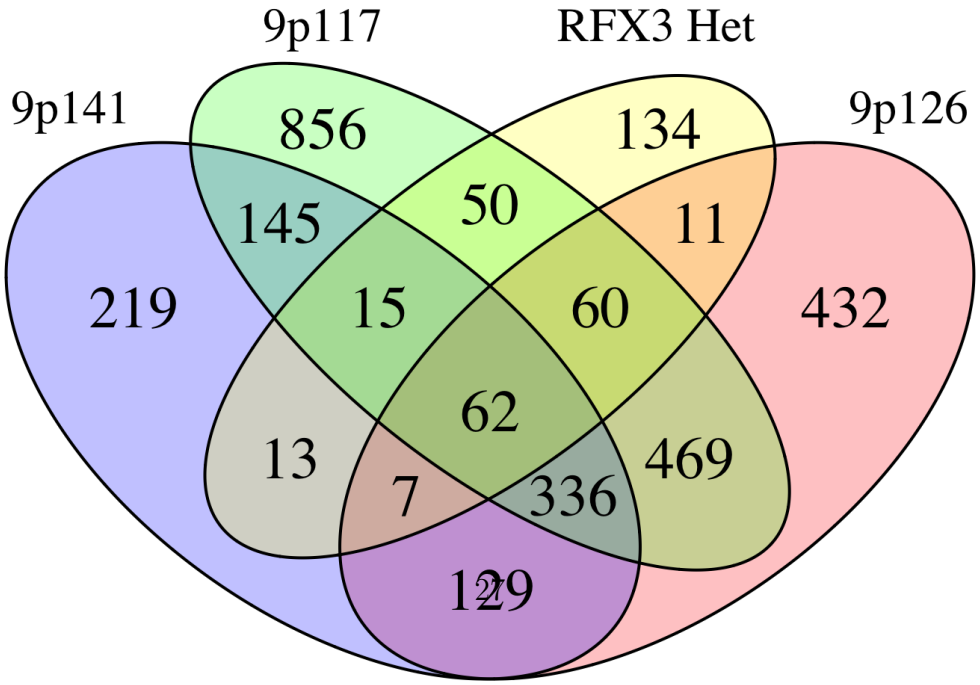
To explore the differentially expressed genes in each cell line, I chose to focus on downregulated genes. From Fig. 5a, it is apparent that 9P117 vs WT (112 gene deletion including RFX3) has the largest number of down-regulated genes followed by 9P126 vs WT (RFX3 only deletion). This result indicates that RFX3 deletion in a patient background causes increased genetic down-regulation and potentially more suppression of important biological pathways in the brain.

To complement this analysis, I examined the overlap and uniqueness of down-regulated differentially expressed genes (DEGs) across the conditions (Fig. 5b). 9P117 and 9P126 share the most down-regulated Differentially Expressed Genes (DEGs) out of any group with 469 shared DEGs. Because 9P117 and 9P126 share only one deleted gene, RFX3, its deletion contributes substantially to the shared transcriptional dysregulation observed in 9P117 and 9P126. In contrast, 9P117 and 9P141 share only 120 down-regulated DEGs, despite sharing 79 deleted genes. This further indicates that RFX3 is driving much of the transcriptional disruption seen in 9P117, because 9P141 retains 2 copies of RFX3. These findings underscore the significant role of RFX3 in regulating global transcription levels compared to other genes deleted in 9P141.

There are 122 significantly down-regulated genes shared among lines haploinsufficient for RFX3. Interestingly, 62 of these genes are also down-regulated in 9P141, which has two copies of RFX3. This implies that the deletion of one copy of some gene or genes in 9P141 has transcriptomic effects that overlap with those of RFX3 haploinsufficiency.




(a) Bar plot of the number of genes down-regulated in each cell line



Next I sought to identify which of the 356 direct target genes of RFX3 are significantly affected by RFX3 haploinsufficiency. I selected the 50 direct targets of RFX3 with the largest $\log_2$ Fold Change (Log2FC < -0.25) for RFX3 HET for analysis (FDR-adjusted p-value < 0.05; Fig. 6). The number 50 was selected as a cutoff for display purposes. I chose these genes because as direct targets they have a clear biological connection to RFX3 expression. Additionally, based on their Log2FC they are strongly affected by the deletion of one copy of RFX3 and therefore more likely to be biologically impactful. I identified genes that were significantly downregulated in all three cell lines with one copy of RFX3– RFX3 HET, 9P126, and 9P117 (Log2FC < -0.25; FDR-adjusted p-value < 0.05). I next looked for genes that were also not differentially expressed in 9P141, because 9P141 has two copies of RFX3 and therefore should not have differential expression of RFX3 direct targets. Out of the 50 direct target genes, thirteen are downregulated DEGs in 9P117, 9P126, and RFX3 HET, but not in 9P141 (Fig. 6). Across all 356 direct target genes, 25 are downregulated DEGs in 9P117, 9P126, and RFX3 HET, and not in 9P141 (Supplementary Fig. 10).

Hierarchical clustering revealed that for these 50 genes, RFX3 haploinsufficient lines—9P117, 9P126, and RFX3 HET—clustered together, indicating that they share the most similar expression patterns (Fig. 6). This result supports the conclusion that RFX3 haploinsufficiency consistently and reproducibly influences gene expression.

Notably, four of the thirteen genes—ABAT, CAMK2A, CELF6, and RIT2—are listed in the SFARI Gene database as Autism Risk Genes (Table 2). These four genes are also the only autism-risk genes within the broader set of 25 direct target genes (Supplementary Fig. 10). To assess whether autism-risk genes are overrepresented among the direct targets of RFX3, a hypergeometric test was performed. Out of 28,052 total genes, 945 were classified as autism-risk genes. Out of the 25 were direct targets of RFX3 that were down regulated in haploinsufficient lines and not downregulated in 9P141, 4 were autism-risk genes. The hypergeometric test yielded a statistically significant p-value (p=0.00920), indicating that autism-risk genes are significantly enriched among the direct targets of RFX3 down regulated in haploinsufficient lines and not downregulated in 9P141. The SFARI Gene database is a database of 1,231 genes implicated in autism, along with the supporting literature and a score of how strongly the genes are linked to autism [18]. Of these genes, 945 are scored as being clearly implicated in Autism or strong candidates as Autism-risk genes, and the remaining genes have only suggestive evidence. These 945 genes are the genes analyzed in this study and referred to as Autism-risk genes. The downregulation of these genes ties RFX3 haploinsufficiency to a mechanism for the autism phenotype in 9P- syndrome.

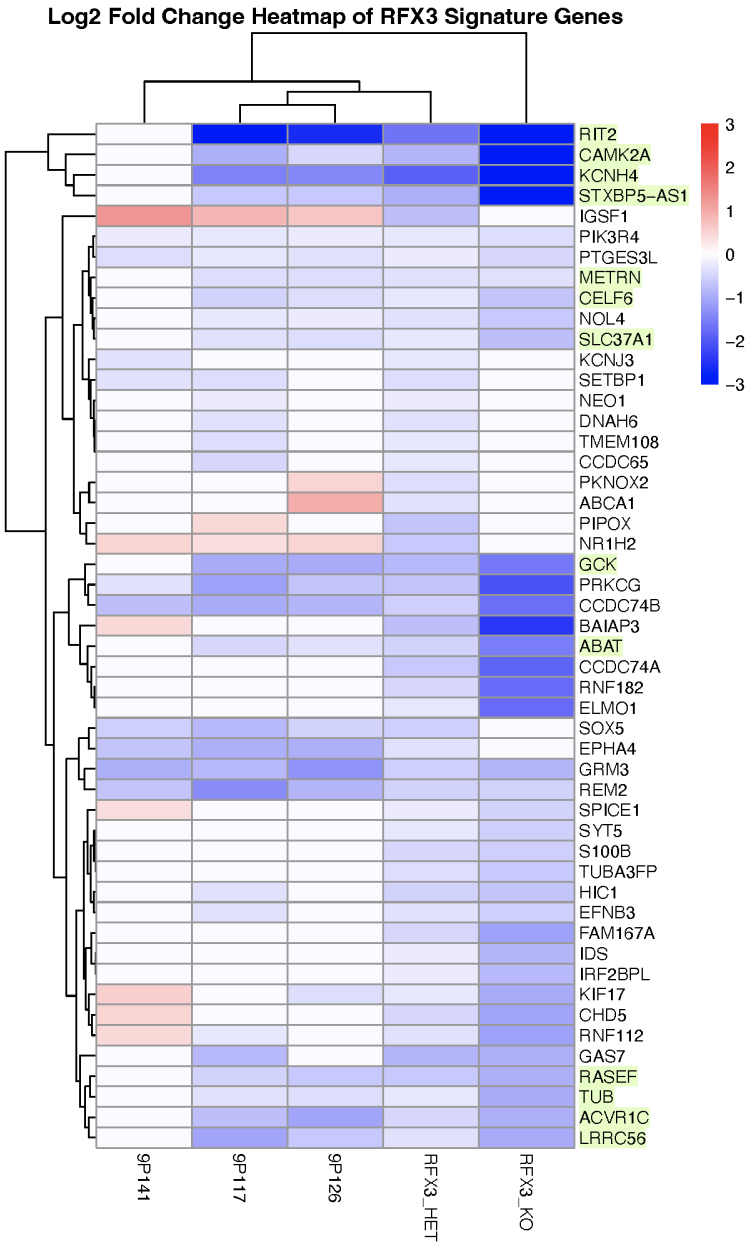



Figure 6: Log2 Fold Change heatmap of all five cell lines. Highlighted genes are downregulated DEGs in 9P117, 9P126, and RFX3 HET, and not down regulated DEGs in 9P141

CAMK2A is a kinase in neuronal synapses involved in modulating synaptic strength, which is essential for learning and memory. It is known that mice with both copies of CAMK2A disrupted display learning deficits and impaired synaptic activity. People with both heterozygous mutations develop intellectual disability, and people with homozygous missense mutations have intellectual disability, global developmental delays, and seizures [19] [20]. ABAT encodes an enzyme that helps break down the inhibitory neurotransmitter GABA. GABA is an important chemical in the brain, and when its synthesis or breakdown is disrupted can cause different neurological diseases. Genome studies have further implicated ABAT in autism, identifying genetic variants associated with the disorder [21]. A child

born with two heterozygous ABAT mutations developed global developmental delays along with other symptoms [22]. CELF6 encodes an RNA binding protein that regulates RNA transcription [23]. Knocking out this gene in mice results in the mice developing autism-like phenotypes [24]. RIT2 encodes an enzyme that regulates intracellular signalling in neurons. A meta-analysis found that a polymorphism in the gene was significantly associated with autism [25].

Table 2: Gene Expression Metrics Across Cell Lines for Selected Autism Risk Genes

| Gene | 9P117 | | 9P126 | | RFX3 HET | | RFX3 KO | |
|---|---|---|---|---|---|---|---|---|
| | LFC | padj | LFC | padj | LFC | padj | LFC | padj |
| ABAT | -0.420 | $7.59 \times 10^{-4}$ | -0.346 | $7.47 \times 10^{-3}$ | -0.509 | $8.08 \times 10^{-6}$ | -1.543 | $4.14 \times 10^{-29}$ |
| CAMK2A | -0.931 | $3.75 \times 10^{-6}$ | -0.474 | $1.85 \times 10^{-2}$ | -0.861 | $4.44 \times 10^{-6}$ | -5.063 | $1.89 \times 10^{-99}$ |
| CELF6 | -0.497 | $4.50 \times 10^{-10}$ | -0.382 | $3.19 \times 10^{-6}$ | -0.263 | $6.73 \times 10^{-4}$ | -0.680 | $1.28 \times 10^{-13}$ |
| RIT2 | -2.994 | $2.50 \times 10^{-11}$ | -2.610 | $9.88 \times 10^{-9}$ | -1.628 | $6.22 \times 10^{-5}$ | -3.603 | $3.82 \times 10^{-10}$ |

## 2.3 Functional Enrichment Analysis Ties DEGs to Neuronal Changes

To understand the biological pathways disrupted as a result of the down-regulated genes, I performed functional enrichment analysis on the genes down-regulated in each cell line, with noisy genes excluded (Fig. 7). The set of pathways used were Gene Ontology Biological Processes, which include sets of genes involved in specific cellular processes [26]. 9P141 and 9P126 had no enriched pathways that their down-regulated genes were significantly likely to affect even when raising the Log2FC and base-mean expression cutoffs. RFX3 HET had 84 significant pathways, most of them related to synaptic function (Fig. 7a). RFX3 KO had 36 affected pathways, most of them related to ciliary processes or synaptic function (Fig. 7b). 9P117 had 15 enriched pathways, most of them related to ciliary processes or synaptic function (Fig. 7c). Five of these pathways were also enriched in RFX3 HET: regulation of trans-synaptic signaling, modulation of chemical synaptic transmission, regulation of synapse assembly, cilium or flagellum-dependent cell motility, and cilium-dependent cell motility.

Functional enrichment analysis of GO Biological Processes revealed that 9P117 had the strongest biologically relevant disruption, which was quite similar to that of RFX3 deletion (Fig. 7). In contrast, 9P141 and 9P126 may have had too many noisy genes for functional enrichment analysis to detect significantly enriched pathways. This discrepancy suggests that 9P117 had a stronger signal, which could plausibly be explained by the deletion of a gene on chromosome 9p that works synergistically with RFX3. The simultaneous loss of both genes may have exacerbated their effects, resulting in more significantly downregulated genes in the disrupted pathways.

These findings confirm that RFX3 haploinsufficiency results has the anticipated effect of altering synaptic and ciliary functions in patient neurons. It emphasizes that RFX3

haploinsufficiency not only causes transcriptional dysregulation, but that the transcriptional dysregulation results in altered biologically significant processes.

To identify pathways consistently dysregulated in response to the loss of one copy of RFX3, I examined the overlap of downregulated genes across three cell lines: 9P117, 9P126, and RFX3 HET. All noisy genes were excluded from analysis. In the RFX3 haploinsufficient lines (9P117, 9P126, and RFX3 HET), 122 genes were consistently down-regulated, and functional enrichment analysis identified 25 significantly enriched pathways shown in Fig. 8. These pathways included critical processes such as regulation of trans-synaptic signaling, neuron migration, synapse organization, and regulation of membrane potential. These findings underscore the pivotal role of RFX3 in regulating neurodevelopmental and synaptic functions.

Based on these results, we can infer that genetic dysregulation in 9P126 does have phenotypic consequences even though it did not have any enriched gene sets when looking at its DEGs from 9P126 vs WT. By ensuring all the genes were also downregulated in other haploinsufficient cell lines as well as 9P126, the impact of RFX3 became much more apparent. This indicates that the other DEGs in 9P126 are additional "noise" potentially due to genetic background differences that was not removed during noisy gene elimination.

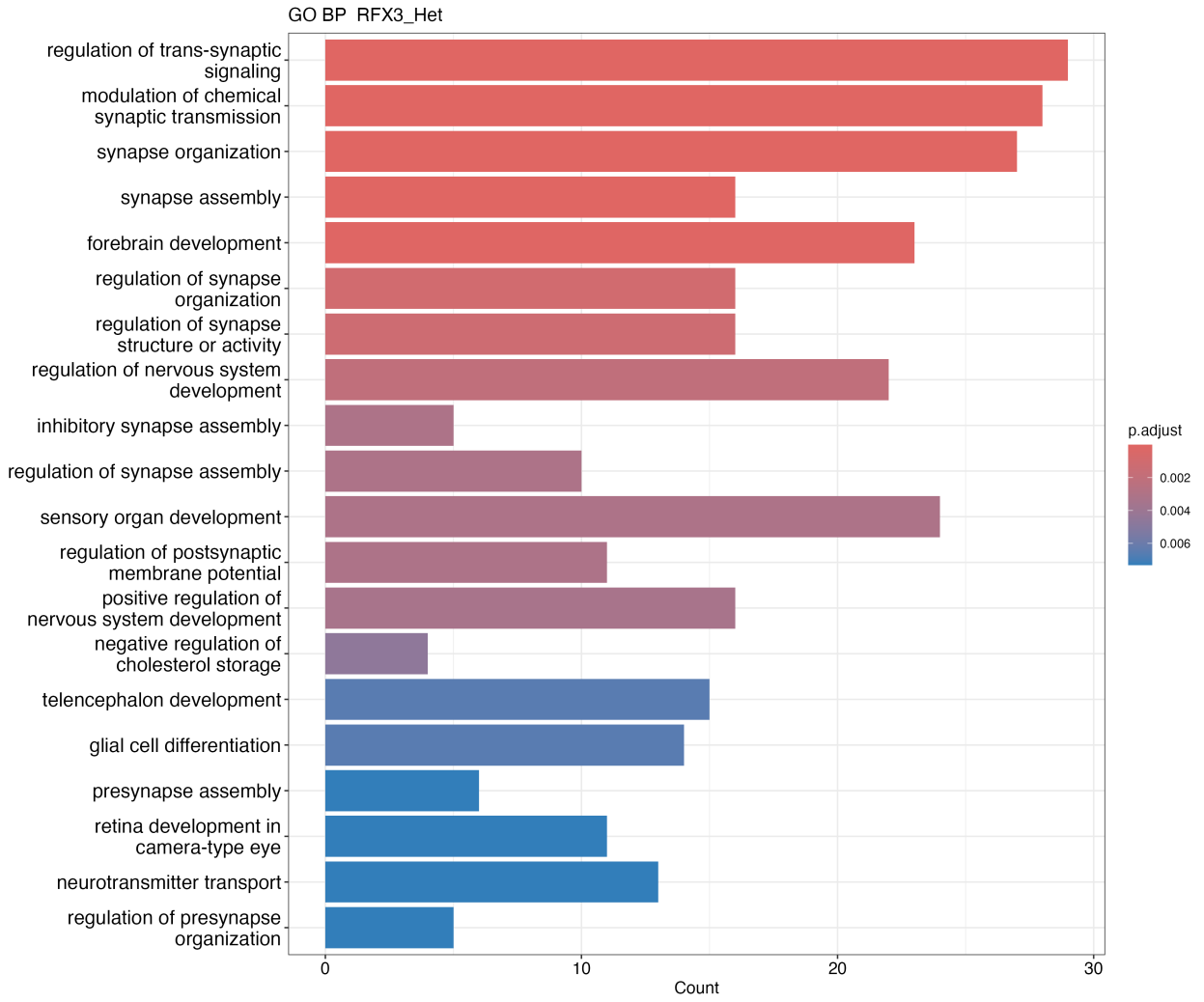



Figure 7: a) 20 most significant GO Biological Processes Enriched in Genes Significantly Down-regulated in RFX3 HET vs WT (Ranked by FDR-Adjusted p-value; FDR-Adjusted p-value $< 0.05$).

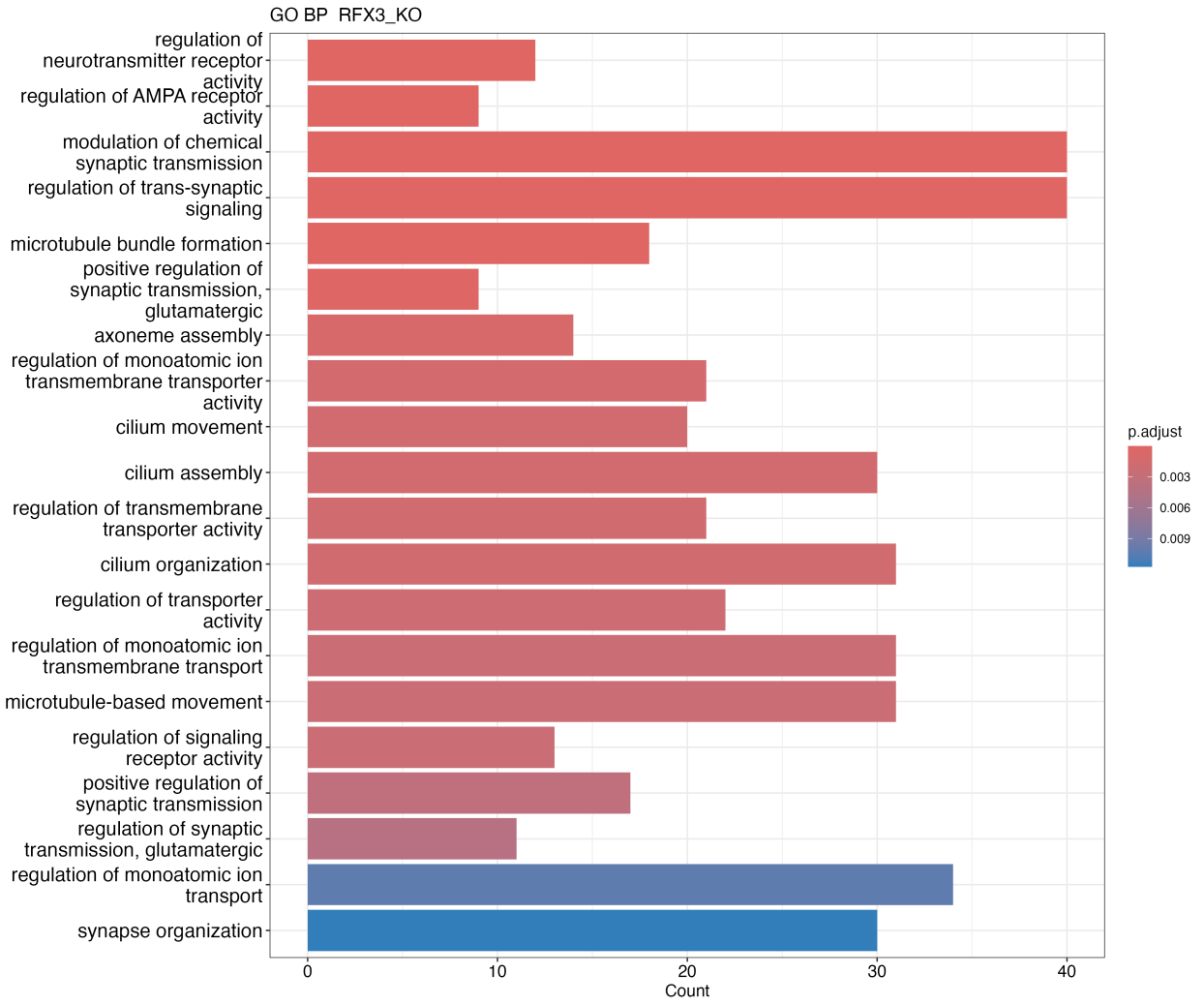



Figure 7: b) 20 most significant GO Biological Processes Enriched in Genes Significantly Down-regulated in RFX3 KO vs WT (Ranked by FDR-Adjusted p-value; FDR-Adjusted p-value $< 0.05$).

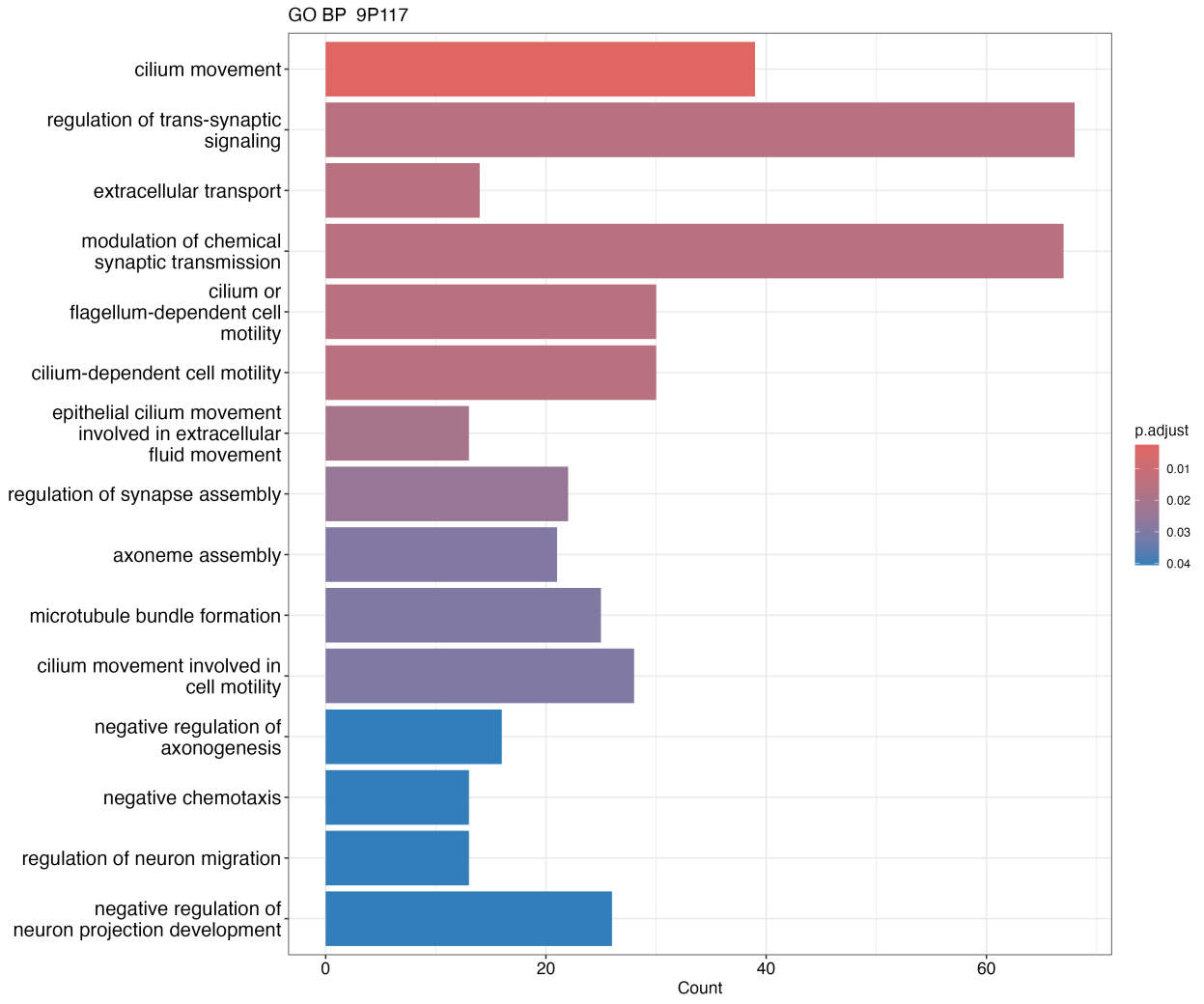



Figure 7: c) GO Biological Processes Enriched in Genes Down-regulated in 9P117 vs WT (FDR-Adjusted p-value $< 0.05$).

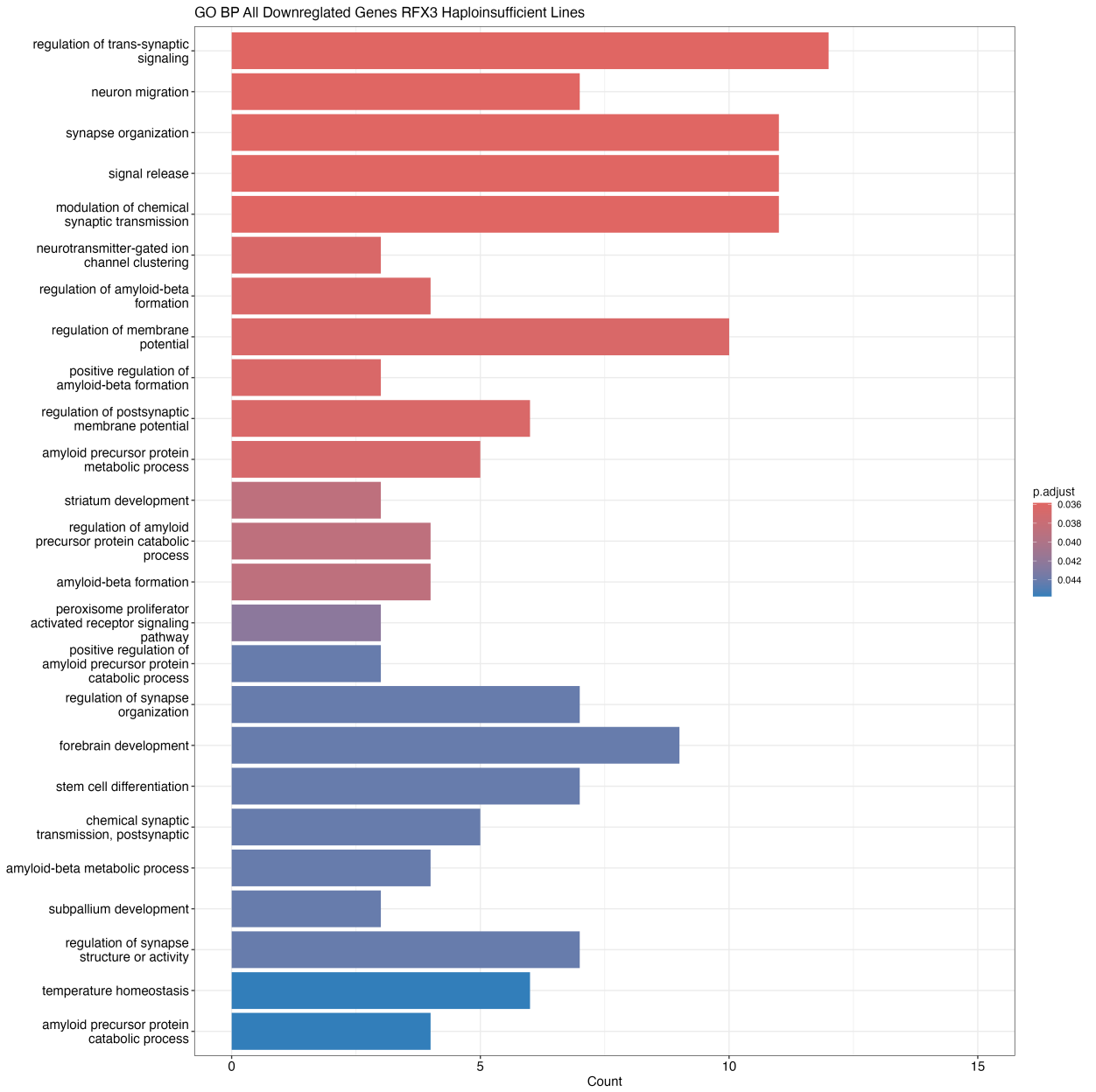



Figure 8: GO Biological Processes Enriched in Genes Significantly Down-regulated in RFX3 Haploinsufficient Lines (9P117, 9P126, RFX3 HET) (FDR-Adjusted p-value $< 0.05$)

## 2.4 Down-regulation of genes in 9P- Syndrome is tied to Autism

Genes significantly downregulated in 9P117, RFX3 HET, and 9P141 were enriched for ASD risk genes, defined by SFARI (n=945) and a recent large exome sequencing study (ASD1021, n=105) [27] (FDR-adjusted p-value <0.05; Table 3). In 9P126, autism was a nominally significant gene set (pval $< 0.05$, padj $> 0.05$). I also explored enrichment of gene sets for

rare genetic intellectual disability (ID) (ORPHA:183757, n=1,953), ADHD (n=25) [28] [29] [30], and Schizophrenia (n = 163) [31]. None of these three gene sets were enriched in any of the cell lines. 9P117 had the highest number of autism-related genes in its down-regulated DEGs, indicating there may be a synergistic deletion exacerbating the effects of RFX3 haploinsufficiency.

| Cell Line | Significant Gene Set | Number of Genes | p-value | FDR-adjusted p-value |
|---|---|---|---|---|
| 9P117 | Autism | 123 | 0.0000528 | 0.000211 |
| 9P126 | Autism | 73 | 0.0419 | 0.167 |
| RFX3 HET | Autism | 36 | 0.00152 | 0.00611 |
| 9P141 | Autism | 55 | 0.000158 | 0.000632 |

Table 3: Neurological Phenotype Enrichment Analysis Results. Enrichment analysis was performed using the ClusterProfiler package, which applies a hypergeometric test to assess the overrepresentation of genes in the Autism Gene Set. Number of Genes refers to the count of genes in the Autism Gene Set.

To understand the contribution of RFX3 to this enrichment of Autism risk genes, I identified which genes in the commonly deleted region of chromosome 9p [6] are direct or indirect targets of RFX3. Direct targets were previously defined as RFX3 or KO DEGs with an RFX3 ChIP-seq peak. Indirect targets are genes that are DEGs for RFX3 HET or RFX3 KO but do not have a ChIP-seq peak (RFX3 does not bind to the gene's regulatory sequence). Because the 403 direct targets (Fig. 2 a) were identified using the large bulk RNA-seq dataset previously generated by a collaborator, I included RFX3 HET and KO DEGs from this analysis to identify indirect targets of RFX3. Noisy gene filtering reduced the number of direct target genes to 356, and this set was used for analysis. There is one direct target of RFX3, SMARCA2, in the commonly deleted region of Chromosome 9 and it is an autism-risk gene. There are 10 genes that are indirect targets of RFX3 in the commonly deleted region, and 2 of them are linked to autism. There are 177 other genes in the region, and 6 of them are linked to autism. However, DOCK8 is expressed in microglia, not neurons, so its lack of expression in this RNA sequencing data from neurons should not indicate its relationship to RFX3 [16]. As previously mentioned, These findings suggest that RFX3 plays a pivotal role in regulating autism risk genes within the commonly deleted region of chromosome 9p. Of the eight autism risk genes in this region, three—SMARCA2 and two indirect targets—are influenced by RFX3 through direct or indirect mechanisms. It does also indicate that RFX3 likely works with the other autism risk genes that are not influenced by its expression.

Additionally, of the five genes other than RFX3 highlighted in the Sams et al. 2021 paper for enrichment in individuals with neurodevelopmental delays (CDC37L1, NFIB, PTPRD, SMARCA2, UHRF2), two of them (NFIB and SMARCA2) are linked with RFX3. No connection was found between RFX3 and CDC37L1, PTPRD, or UHRF The fact that half of the six genes (including RFX3) identified are linked to RFX3 further supports that RFX3 is significantly contributing to transcriptional disruption leading to neurodevelopmental delays.

| Relationship to RFX3 | Number of Genes | Number of Autism-Risk Genes |
|---|---|---|
| Direct targets in Deleted Region | 1 | 1 (SMARCA2) |
| Indirect targets in Deleted Region | 10 | 2 (NFIB, SLC24A2) |
| Unrelated to RFX3 in Deleted Region | 173 | 6 (KANK1, KDM4C, MLANA, PTPRD, SLC1A1, DOCK8) |

Table 4: Genes in the commonly deleted region and their relationship to RFX3 and Autism.

## 2.5 SMARCA2 is a Potential Synergistic Contributor with RFX3

To understand the functional significance of SMARCA2 being a direct target of RFX3, I first identified the log2 Fold Change in each cell line when SMARCA2 has an FDR-adjusted p-value $< 0.05$. SMARCA2 is one component of the 29 protein complex called the SWI/SNF complex [32]. I identified other components of the complex that were significant DEGs in each cell line to understand how down-regulation of SMARCA2 affects other components of the complex. For the RFX3 HET and KO lines I used the large bulk RNA-seq dataset previously generated by a collaborator [13]. Because they analyzed 6 samples for RFX3 KO, their results are more reliable than my data which is based on only 2 RFX3 KO samples. RFX3 het lacked significant dysregulation of SMARCA2. RFX3 KO resulted in downregulation of only SMARCA2. 9P141 has one copy of SMARCA2 deleted and retains both copies of RFX3 while 9P126 has one copy of RFX3 deleted and retains both copies of SMARCA2. Both have similar log2 fold changes that are larger than RFX3 KO, and a similar number of dysregulated SWI/SNF Complex Genes. 9P117 has one copy of RFX3 and one copy of SMARCA2 deleted, and has a much larger log2 fold change than the other cell lines, and more dysregulated SWIComplex Genes. Therefore, SMARCA2 has stronger downregulation when its transcriptional activator, RFX3, is also deleted and more SWI/SNF complex genes are dysregulated.

| Cell Line | SMARCA2 Log2FC | Number of Dysregulated SWI/SNF Complex Genes |
|---|---|---|
| RFX3 HET | NA | NA |
| RFX3 KO | -0.297 | 1 |
| 9P141 | -0.467 | 2 |
| 9P126 | -0.493 | 3 |
| 9P117 | -1.121 | 5 |

Table 5: SMARCA2 and SWI/SNF Complex Gene Dysregulation in Each Cell line

# Chapter 3

# Discussion

To investigate RFX3's contribution to 9p- syndrome, I profiled genome-wide transcriptomic changes resulting from RFX3 haploinsufficiency in human neurons and examined their connection to autism-related pathways and phenotypes. I used bulk RNA-sequencing data from iPSC-derived neurons from three patients with 9p- syndrome and CRISPR-engineered cell lines with zero, one, or two copies of RFX3. 9P126 had one copy of RFX3 deleted, while 9P117 had 112 genes heterozygously deleted on chromosome 9 including RFX3. 9P141 had 40 genes deleted, but retained both copies of RFX3. The 9P126 line allowed me to isolate changes caused by RFX3 deletion alone in a patient, 9P117 provided insight into changes driven by RFX3 deletion in the context of additional gene deletions, and 9P141 served as a "control" to identify changes attributed to other gene deletions on chromosome 9p. A multi-omic analysis identified 403 direct targets of RFX3. Using PCA, we demonstrated that transcriptional profiles cluster based on RFX3 copy number when using only the 403 direct targets of RFX3, supporting its contribution to transcriptomic changes. Filtering out genes with high variability due to genetic background, technical processing, or sex-specific effects reduced variance and clarified clustering patterns in PCA (Figure 3), allowing us to isolate the impact of RFX3 deletion. The clustering of RFX3 haploinsufficient cell lines (9P117, 9P126, and RFX3 HET) near each other and away from lines with two intact copies of RFX3 supports its central role in transcriptional regulation. However 9P117 and 9P126 did not form a single cluster, indicating that RFX3 is not the sole driver of transcriptional dysregulation in 9p- syndrome.

Differential Gene Expression Analysis revealed 9P117 and 9P126, which both have heterozygous RFX3 deletions, share 469 downregulated genes. 9P117 and 9P141 share 26 deleted genes and do not share RFX3 deletion, have only 120 commonly downregulated genes. This suggests that RFX3 haploinsufficiency drives significant transcriptional dysregulation, as evidenced by the large overlap of downregulated genes in 9P117 and 9P126 compared to the minimal overlap between 9P117 and 9P141, which lacks RFX3 deletion. I also found 122 genes significantly downregulated in all three haploinsufficient lines—9P117, 9P126, and RFX3 HET. Functional Enrichment Analysis of these genes identified pathways related to synaptic and ciliary processes. This result aligns with the known biological pathways affected by RFX3 haploinsufficiency, which were previously identified in CRISPR-modified isogenic cell lines [13]. Their enrichment in iPSC-derived neurons from patients with 9p- syndrome confirms that RFX3 haploinsufficiency has the anticipated effects on transcriptional regulation in 9p-

syndrome. Importantly, these findings validate that the known mechanistic link between RFX3 disruption and autism—via synaptic disruption—is also present in patient cells.

Twenty-five direct targets of RFX3 were consistently significantly downregulated across all RFX3 haploinsufficient lines and not downregulated in 9P141. Four autism-risk genes (ABAT, CAMK2A, CELF6, and RIT2) from the SFARI database were included in these 25 genes. This overlap was statistically significant ($p = 0.00887$), linking RFX3 haploinsufficiency in 9p- patient cell lines to autism phenotypes in 9p- syndrome. Functional Enrichment Analyses revealed that the 9P117 cell line, which has heterozygous deletions of RFX3 and seven other autism-risk genes (SMARCA2, DOCK8, KANK1, KDM4C, MLANA, PTPRD, SLC1A1), showed stronger enrichment for autism-related transcriptional changes than 9P141, where only three autism-risk genes (SMARCA2, DOCK8, KANK1) are deleted. This could suggest that RFX3 haploinsufficiency exacerbates autism-related dysregulation, but its specific contribution cannot be disentangled from the effects of other deleted genes.

The interaction between SMARCA2 and RFX3 is particularly intriguing. This study identified SMARCA2 as a direct target of RFX3. SMARCA2, a component of the SWI/SNF chromatin remodeling complex, has been linked to Coffin-Siris syndrome, Nicolaides-Baraitser syndrome, and autism. Previous studies indicate that SMARCA2 loss can destabilize the SWI/SNF complex, leading to the downregulation of other subunits [33]. Interestingly, both RFX3 and SMARCA2 mutations are enriched in individuals with neurodevelopmental disorders. They also both have a $pLI > 0.9$, indicating that mutations in these genes often have phenotypic consequences [6]. In this study, 9P117 neurons (with one copy of RFX3 and SMARCA2 deleted) showed decreased expression of SMARCA2 and other SWI/SNF subunits beyond what was observed when just SMARCA2 was deleted in 9P141, or just RFX3 was deleted in 9P126. This result suggests potential synergy between RFX3 and SMARCA2 deletions. Future experiments should compare transcriptomic changes in isogenic neurons haploinsufficient for RFX3, SMARCA2, and both genes to clarify their individual and combined roles in neurodevelopmental phenotypes.

9p- syndrome is a rare chromosomal deletion syndrome characterized by intellectual disabilities, developmental delays, and autism. In some chromosomal deletion syndromes, driver genes within the deleted region can contribute disproportionately to the phenotype. For instance, TBX1 is a major driving gene observed in 22q11.2 deletion syndrome (DiGeorge syndrome). DiGeorge Syndrome is similar to 9p- syndrome in that there is a deletion of 30- 40 gene, and TBX1 is similar to RFX3 in that they're both developmental transcription factors [34]. Conversely, deletions like 16p11.2, one of the most well-known autism-associated chromosomal deletions, lack a single driver gene, with evidence suggesting that phenotypes arise from a cumulative effect of multiple gene deletions [35] [36]. This study positions 9p- syndrome as neither having a single driver, nor having all genes equally contribute: it suggests that RFX3 plays a major role in autism-related dysregulation in 9p- syndrome but that neurodevelopmental phenotypes are also influenced by other deleted genes in the region.

## 3.1 Limitations of the Study

This pilot study faced inherent challenges, particularly the use of non-isogenic patient-derived cell lines. Genetic background variation introduced significant interpersonal differences

unrelated to the pathobiology of 9p- syndrome. While efforts were made to mitigate these effects, such as excluding Y chromosome genes, XIST, genes with low expression, and genes with high variance unrelated to genetic perturbations, residual variation likely influenced the results. Despite this variation, the impact of RFX3 haploinsufficiency in 9p- syndrome was apparent. Future research should occur in an isogenic background, or a more genetically diverse WT baseline should be used such as a female WT line. These modifications would improve the robustness of differential gene expression analyses.

The limited number of cell lines also constrained this study's ability to generalize findings. Incorporating more patient-derived cell lines into transcriptomic analyses could refine our understanding of RFX3's role and its interaction with other genes. Additionally, this study used data exclusively from day 14 neurons. Exploring data from other times in development and other brain cell types would shed light on the sets of genes RFX3 affects throughout neurodevelopment.

## 3.2 Clinical Implications

The identification of RFX3 as a significant contributor to autism-related transcriptional dysregulation in 9p- syndrome provides a foundation for exploring targeted therapies. RNA therapies, such as antisense oligonucleotides (ASOs), have been successfully used to ameliorate phenotypes in other cases of gene knockouts, such as the SMN1 gene in Spinal Muscular Atrophy [37]. A similar approach could be employed to overexpress RFX3 in 9p- patient iPSC-derived neurons, testing whether RFX3 restoration is sufficient to recover autism-related transcriptional dysregulation. This is an especially promising avenue as a ASO targeting RFX3 has already been proven to increase expression of RFX3 and its direct target, CAMK2A [13]. Success in these efforts could pave the way for RNA-based treatments to alleviate the neurodevelopmental symptoms of 9p- syndrome, improving outcomes for affected individuals. I am currently planning to use an ASO to overexpress RFX3 in 9P126 and 9P117, and compare my results to ASO overexpression of RFX3 in RFX3 HET. Even if an ASO is not an effective treatment, identifying that RFX3 plays a significant role in 9p- syndrome has clinical implications. As the phenotypic consequences of deleting other genes in the commonly deleted region of chromosome 9p are identified, doctors will be able to ascertain a child's risk for neurodevelopmental disorders. Future research should investigate whether RFX3 deletion serves as a predictive factor for autism development or quantify the extent to which RFX3 haploinsufficiency contributes to autism risk in this population. To facilitate this research, the severity of autism and the specific autism-related symptoms should be noted for each patient with 9p- syndrome.

# Chapter 4

# Methods

## 4.1 Identification of Genes in Deleted Regions

To identify the genes located on chromosome 9p and determine which were deleted in each patient, I utilized the UCSC Genome Browser Table Browser tool. Using the Genes and Gene Predictions group and the HGNC track, I extracted a list of annotated genes within the 9p region, and deleted regions for each patient.

## 4.2 CUT&RUN-sequencing

I utilized CUT&RUN-sequencing data from Lai et al. (2023) which identified RFX3 binding sites in neurons [13]. Annotated binding peaks were used to identify genes to which RFX3 bound. Data generation and processing steps were performed as described in the cited study.

## 4.3 Maintenance and culture of iPSCs

9p- patient iPSC samples were obtained from Tychell Turner's lab at Washington University and Parental PGP1-SV1 iPSC clones and RFX3 CRISPR-Cas9 edited clones were obtained from Synthego. All iPSC lines were mycoplasma negative, karyotypically normal, and expressed pluripotency markers OCT4, SOX2, NANOG, SSEA4, TRA-1-60. All lines were cultured following the protocol described in Lai et al. (2023) [13].

## 4.4 Differentiation of iPSCs to neurons

Human iPSC-derived neurons were generated from iPSCs transduced with the lentiviruses Tet-O-Ngn2-Puro and FUW-M2rtTA following the protocol described by Jenny Lai [13], based on the method originally published by Zhang et al. [14]. Neurons were maintained in conditioned Sudhof Neuronal Growth Medium with weekly media changes, following identical steps and using identical materials outlined in the cited protocol.

## 4.5 Bulk RNA-sequencing

I utilized differential gene expression analysis results frome 6 samples of RFX3 HET and RFX3 KO iPSC-derived neurons from Lai et al. (2023) [13]. This data was used to identify direct transcriptional targets of RFX3. RFX3 KO results were also used in Table 5. Data generation and processing steps were performed as described in the cited work.

Total RNA was harvested from samples using the PureLink RNA Mini kit (Life Technologies Catalog # 12183018A). Libraries were prepared with the KAPA mRNA prep. Sequencing was performed on Illumina NovaSeq, 2x150bp configuration at approximately 100X depth per sample.

For our analysis of the smaller dataset, reads were aligned to hg38 and bam files were manipulated using STAR v2.7.11a [38] and Samtools=1.21 [39]. Trimming of the reads and quality control were performed using fastp/0.23.4 [40]. Processing the reads with featureCounts (-p -s 2) to obtain a gene-level counts matrix [41]. Genes with fewer than 10 counts across all samples were excluded from the analysis. Differential gene expression analysis was performed using DESeq2 [42] with WT set as the reference condition and results coefficient set as 'condition_HET_vs_WT', 'condition_KO_vs_WT', 'condition_9P117_vs_WT', 'condition_9P126_vs_WT', 'condition_9P141_vs_WT', and lfcShrink(type="apeglm").

Differentially expressed genes with FDR-adjusted p-value $<0.05$, $|\text{Log2FC}| > 0.25$, and basemean gene expression $> 30$ were considered significant and used in downstream analysis.

## 4.6 Noisy Gene Filtering

To reduce noise in the dataset, genes were filtered based on high variability beween genetic backgrounds and experimental conditions. Differential gene expression analysis was performed using DESeq2 [42] with with HET and WT_JL set as the reference conditions, and contrasts were defined for the following coefficients: '9P126 vs HET_JL', 'HET_KP vs HET_JL', and 'WT_KP vs WT_JL'. 9P126 and HET are samples with loss of function mutations in only RFX3, but they are from different genetic backgrounds. 9P126 is from the patient background, and HET_JL is from the PGP1 background. HET_JL and HET_KP are identical heterozygous samples processed by different individuals. WT_JL and WT_KP are identical wildtype samples processed by different individuals. In the coefficient names, *_JL and *_KP indicates who processed the samples. Including these comparisons allowed for the identification of DEGs that may be subject to variability due to differences in sample processing rather than true biological differences.

Noise attributable to genetic background was addressed by comparing 9p126 samples with RFX3 HET samples, eliminating DEGs with log2 fold changes (Log2FC) $<$ -3 or $>$ 3 and FDR-adjusted p-value $< 0.05$. Similarly, to mitigate experimental variability, DEGs with Log2FC $<$ -1 or $>$ 1 and FDR-adjusted p-value $< 0.05$ were removed in comparisons between HET_JL and HET_KP, as well as WT_JL and WT_KP. These filtering steps helped refine the dataset, reducing variability and improving its biological interpretability.

Sex-specific variability was minimized by excluding genes on the Y chromosome and XIST, a gene only expressed in females [15]. Additionally, genes with fewer than 10 counts across all samples were removed.

## 4.7 Gene Ontology Enrichment Analysis

Gene Ontology (GO) functional enrichment analysis was performed on the significantly downregulated genes in each cell type using clusterProfiler [43]. GO biological processes with FDR-adjusted p-value $<0.05$ were considered significant.

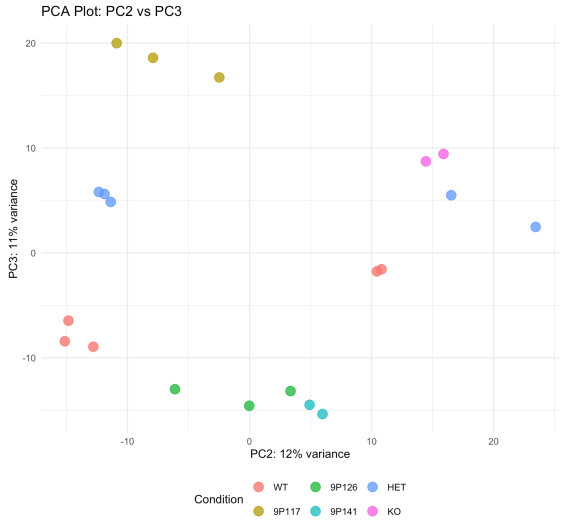
## 4.8 Data Visualization

PCA was performed using DESeq2's plotPCA function with Condition as the grouping variable. Variance-stabilizing transformation (vst, blind = FALSE) was applied to normalize the data for visualization [42]. Plots were generated using ggplot2 [44].

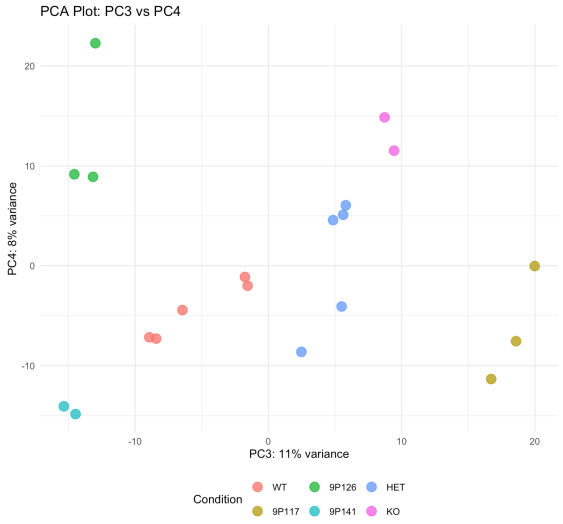
Volcano plots were generated using the EnhancedVolcano R package (version 1.24.0) [45] to visualize differentially expressed genes. The plots highlight genes with an absolute log2 fold change greater than 0.25 and an adjusted p-value $< 0.05$.

Heatmaps of Log2FCs were generated using the pheatmap R package (version 1.0.12) for visualization of differentially expressed genes. Hierarchical clustering, implemented within pheatmap, was applied to both rows (genes) and columns (conditions) to group them based on similarity. Default clustering methods and distance metrics were used [46].

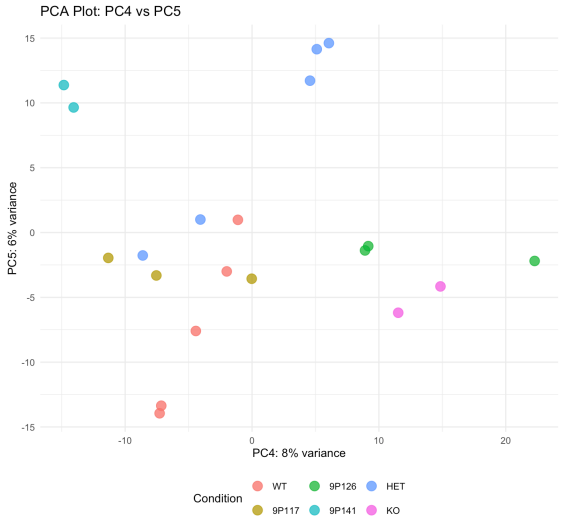
## 4.9 Supplementary Figures




(a) PCA plot of principal components 2 vs 3. There is lack of significant clustering based on known commonly deleted genes.




(b) (PCA plot of principal components 3 vs 4. There is lack of significant clustering based on known commonly deleted genes.




(c) PCA plot of principal components 4 vs 5. There is lack of significant clustering based on known commonly deleted genes.

Figure 9: Subfigures (a), (b), and (c) show the PCA plot for genes prior to noisy gene removal.

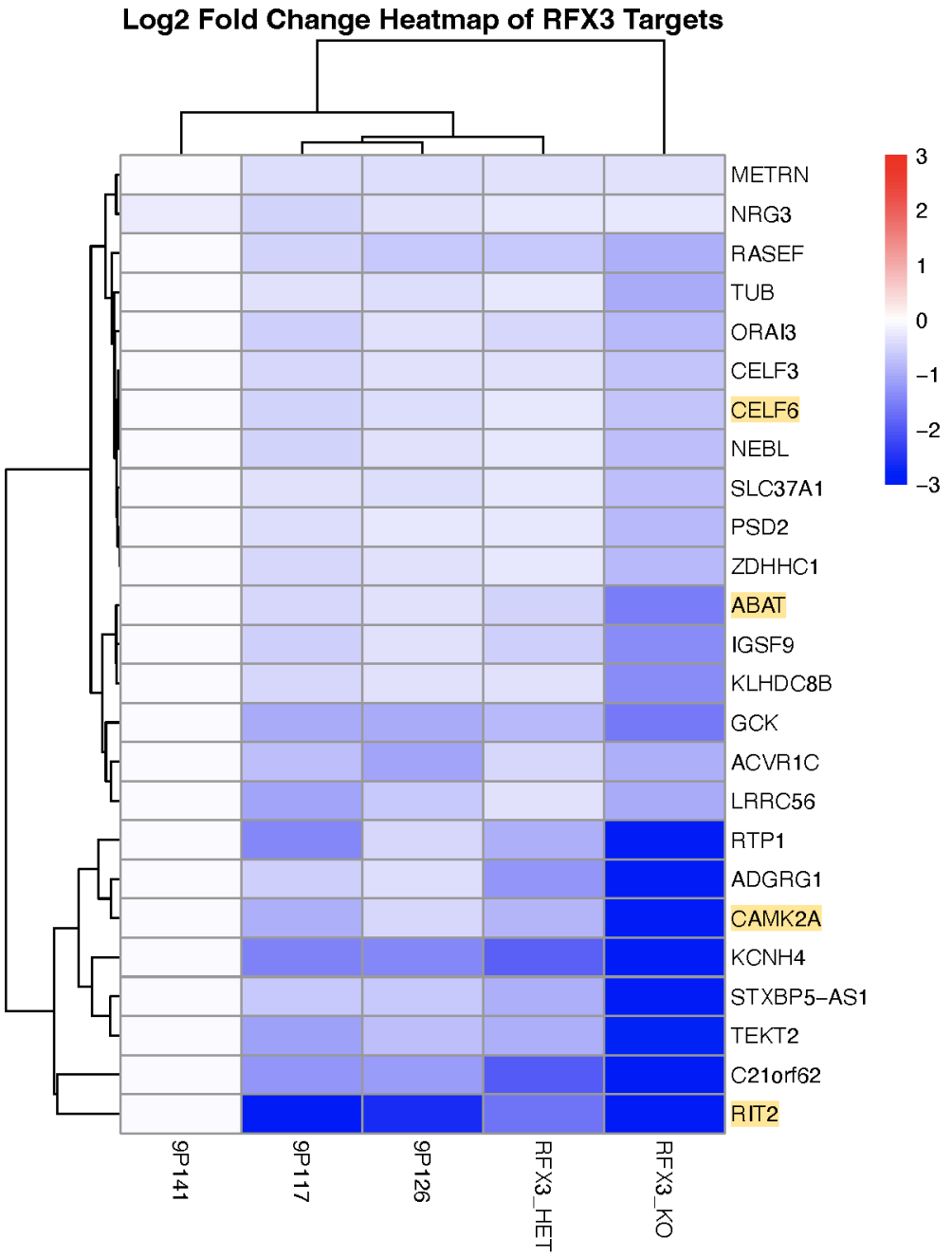



Figure 10: Heatmap of 25 direct target genes significantly downregulated in lines haploinsufficient for RFX3 (9P117, 9P126, RFX3 HET) and not in 9P141. Log2FC $<$ -0.25, FDR-adjusted p-value $<$ 0.05, baseMean $>$ 30. Highlighted Genes have a score of 1 or 2 in the SFARI database

# References


[1] O. Alfi, G. N. Donnell, B. F. Crandall, A. Derencsenyi, and R. Menon. "Deletion of the short arm of chromosome no. 9 (46,9p-): a new deletion syndrome". In: *Annales de Génétique* 16.1 (Mar. 1973), pp. 17–22.

[2] O. S. Alfi, G. N. Donnell, P. W. Allderdice, and A. Derencsenyi. "The 9p- syndrome". In: *Annales de Génétique* 19.1 (Mar. 1976), pp. 11–16.

[3] N. O. for Rare Disorders. *Chromosome 9, Partial Monosomy 9p - Symptoms, Causes, Treatment | NORD*. Accessed: 2025-01-14. 2009. URL: https://rarediseases.org/rare-diseases/chromosome-9-partial-monosomy-9p/ (visited on 01/14/2025).

[4] H. G. N. C. (HGNC). *Gene Symbols*. Accessed: 2025-01-13. 2023. URL: https://www.genenames.org/.

[5] R. T. Starosta et al. "Using a New Analytic Approach for Genotyping and Phenotyping Chromosome 9p Deletion Syndrome". In: *European Journal of Human Genetics* 32.9 (2024), pp. 1095–1105. DOI: 10.1038/s41431-024-01667-y.

[6] E. I. Sams et al. "From Karyotypes to Precision Genomics in 9p Deletion and Duplication Syndromes". In: *Human Genetics and Genomics Advances* 3.1 (Dec. 2021), p. 100081. DOI: 10.1016/j.xhgg.2021.100081. URL: https://doi.org/10.1016/j.xhgg.2021.100081.

[7] MedlinePlus. *How do geneticists indicate the location of a gene?* Accessed January 14, 2025. Mar. 2021. URL: https://medlineplus.gov/genetics/understanding/howgeneswork/genelocation/ (visited on 01/14/2025).

[8] D. Alyousfi, D. Baralle, and A. Collins. "Gene-specific metrics to facilitate identification of disease genes for molecular diagnosis in patient genomes: a systematic review". In: *Briefings in Functional Genomics* 18.1 (Jan. 2019), pp. 23–29. DOI: 10.1093/bfgp/ely033. URL: https://doi.org/10.1093/bfgp/ely033.

[9] H. K. Harris, T. Nakayama, J. Lai, B. Zhao, N. Argyrou, C. S. Gubbels, A. Soucy, C. A. Genetti, V. Suslovitch, L. H. Rodan, et al. "Disruption of RFX family transcription factors causes autism, attention-deficit/hyperactivity disorder, intellectual disability, and dysregulated behavior". In: *Genetics in Medicine* 23 (2021), pp. 1028–1040. DOI: 10.1038/s41436-021-01114-z.

[10] B. P. Piasecki, J. Burghoorn, and P. Swoboda. "Regulatory Factor X (RFX)-mediated transcriptional rewiring of ciliary genes in animals". In: *Proceedings of the National Academy of Sciences of the United States of America* 107.29 (2010), pp. 12969–12974. DOI: 10.1073/pnas.0914241107.

[11] L. El Zein, A. Ait-Lounis, L. Morlé, J. Thomas, B. Chhin, N. Spassky, W. Reith, and B. Durand. "RFX3 governs growth and beating efficiency of motile cilia in mouse and controls the expression of genes involved in human ciliopathies". In: *Journal of Cell Science* 122.17 (2009), pp. 3180–3189. DOI: 10.1242/jcs.048348.

[12] C. Benadiba et al. "The Ciliogenic Transcription Factor RFX3 Regulates Early Midline Distribution of Guidepost Neurons Required for Corpus Callosum Development". In: *PLoS Genetics* 8.3 (2012), e1002606. DOI: 10.1371/journal.pgen.1002606. URL: https://doi.org/10.1371/journal.pgen.1002606.

[13] J. Lai. "Functional genomic dissection of human brain development, degeneration, and dysfunction in pediatric neurogenetic disorders". Doctoral dissertation. Harvard University Graduate School of Arts and Sciences, 2023.

[14] Y. Zhang, C. Pak, Y. Han, H. Ahlenius, Z. Zhang, S. Chanda, S. Marro, C. Patzke, C. Acuna, J. Covy, et al. "Rapid single-step induction of functional neurons from human pluripotent stem cells". In: *Neuron* 78 (2013), pp. 785–798. DOI: 10.1016/j.neuron.2013.05.029.

[15] A. Loda and E. Heard. "Xist RNA in action: Past, present, and future". In: *PLoS Genetics* 15.9 (2019), e1008333. DOI: 10.1371/journal.pgen.1008333.

[16] K. Namekata, X. Guo, A. Kimura, N. Arai, C. Harada, and T. Harada. "DOCK8 is expressed in microglia, and it regulates microglial activity during neurodegeneration in murine disease models". In: *The Journal of Biological Chemistry* 294.36 (Sept. 2019), pp. 13421–13433. ISSN: 0021-9258. DOI: 10.1074/jbc.RA119.007645. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6737224/ (visited on 01/16/2025).

[17] The Human Protein Atlas. *Brain tissue expression of CBWD1*. Accessed: 2025-01-24. n.d. URL: https://v22.proteinatlas.org/ENSG00000172785-CBWD1/brain.

[18] S. H. G. Module. *SFARI*. 2024. URL: https://www.sfari.org/resource/sfari-gene/.

[19] P. H. Chia, F. Zhong, B. Chen, B. Lee, K. Jin, Y.-C. Yang, Q. Zeng, W. Han, E. L.-Y. Goh, W. J. Chng, et al. "A homozygous loss-of-function CAMK2A mutation causes growth delay, frequent seizures and severe intellectual disability". In: *eLife* 7 (2018), e32451. DOI: 10.7554/eLife.32451. URL: https://elifesciences.org/articles/32451.

[20] S. Küry, G. M. van Woerden, T. Besnard, M. Proietti Onori, X. Latypova, H. Lamothe, O. Alibeu, A. Afenjar, S. Naudion, K. Siquier-Pernet, et al. "De Novo Mutations in Protein Kinase Genes CAMK2A and CAMK2B Cause Intellectual Disability". In: *American Journal of Human Genetics* 101.5 (2017), pp. 768–788. DOI: 10.1016/j.ajhg.2017.10.003. URL: https://www.cell.com/ajhg/fulltext/S0002-9297(17)30449-4.

[21] G. Barnby, A. Abbott, N. Sykes, A. Morris, D. E. Weeks, R. Mott, J. Lamb, A. J. Bailey, A. P. Monaco, and I. M. G. S. of Autism Consortium. "Candidate-gene screening and association analysis at the autism-susceptibility locus on chromosome 16p: evidence of association at GRIN2A and ABAT". In: *American Journal of Human Genetics* 76.6 (June 2005), pp. 950–966. DOI: 10.1086/430454. URL: https://www.cell.com/ajhg/fulltext/S0002-9297(07)61404-6.

[22] A. Besse, B. Wu, G. Bruni, H. K. Vanyai, N. H. Elcioglu, C. M. B. De Gusmão, A. Mühl, H. Hu, G. Arno, E. Stevens, et al. "Personalized medicine approach confirms a milder case of ABAT deficiency". In: *Molecular Brain* 9 (2016), p. 93. DOI: 10.1186/s13041-016-0266-0. URL: https://molecularbrain.biomedcentral.com/articles/10.1186/s13041-016-0266-0.

[23] G. Liu, K. Li, X. Qin, M. Wu, X. Ren, M. Wang, Y. Chen, R. Zeng, Y. Xue, and W. Yuan. "RNA-binding protein CELF6 is cell cycle regulated and controls cancer cell proliferation by stabilizing p21". In: *Cell Death & Disease* 10 (2019), pp. 1–14. DOI: 10.1038/s41419-019-1854-4. URL: https://www.nature.com/articles/s41419-019-1854-4.

[24] J. D. Dougherty, S. E. Maloney, D. F. Wozniak, M. A. Rieger, L. Sonnenblick, G. Coppola, N. G. Mahieu, J. Zhang, J. Cai, G. J. Patti, et al. "The Disruption of Celf6, a Gene Identified by Translational Profiling of Serotonergic Neurons, Results in Autism-Related Behaviors". In: *The Journal of Neuroscience* 33 (2013), pp. 2732–2753. DOI: 10.1523/JNEUROSCI.4636-12.2013. URL: https://www.jneurosci.org/content/33/7/2732.

[25] J. Wang, S. Wei, J. Zhang, and H. Wang. "Association between RIT2 rs16976358 Polymorphism and Autism Spectrum Disorder in Asian Populations: A Meta-analysis". In: *Biomed Research International* 2023 (Feb. 2023), p. 8886927. DOI: 10.1155/2023/8886927. URL: https://www.hindawi.com/journals/bmri/2023/8886927/.

[26] *GEO Accession Viewer*. Accessed: Nov. 14, 2024. Nov. 2024. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3633223.

[27] F. K. Satterstrom, J. A. Kosmicki, J. Wang, et al. "Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism". In: *Cell* 180.3 (2020), 568–584.e23. DOI: 10.1016/j.cell.2019.12.036.

[28] D. Demontis, F. Lescai, A. Børglum, S. Glerup, S. D. Østergaard, O. Mors, Q. Li, et al. "Whole-Exome Sequencing Reveals Increased Burden of Rare Functional and Disruptive Variants in Candidate Risk Genes in Individuals With Persistent Attention-Deficit/Hyperactivity Disorder". In: *Journal of the American Academy of Child & Adolescent Psychiatry* 55 (2016), pp. 521–523. DOI: 10.1016/j.jaac.2016.03.009.

[29] D. S. Kim, A. A. Burt, J. E. Ranchalis, B. Wilmot, J. D. Smith, K. E. Patterson, B. P. Coe, et al. "Sequencing of sporadic Attention-Deficit Hyperactivity Disorder (ADHD) identifies novel and potentially pathogenic de novo variants and excludes overlap with genes associated with autism spectrum disorder". In: *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* 174 (2017), pp. 381–389. DOI: 10.1002/ajmg.b.32527.

[30] T. Zayats, K. K. Jacobsen, R. Kleppe, C. P. Jacob, S. Kittel-Schneider, M. Ribasés, J. A. Ramos-Quiroga, et al. "Exome chip analyses in adult attention deficit hyperactivity disorder". In: *Translational Psychiatry* 6 (2016), e923. DOI: 10.1038/tp.2016.196.

[31] D. Liu, D. Meyer, B. Fennessy, C. Feng, E. Cheng, J. S. Johnson, Y. J. Park, et al. "Schizophrenia risk conferred by rare protein-truncating variants is conserved across diverse human populations". In: *Nature Genetics* 55 (2023), pp. 369–376. DOI: 10.1038/s41588-023-01305-1.

[32] P. Mittal and C. W. M. Roberts. "The SWI/SNF complex in cancer — biology, biomarkers and therapy". In: *Nature Reviews Clinical Oncology* 17 (2020), pp. 435–448. DOI: 10.1038/s41571-020-0357-3.

[33] F. Gao et al. "Heterozygous Mutations in SMARCA2 Reprogram the Enhancer Landscape by Global Retargeting of SMARCA4". In: *Molecular Cell* 75.5 (Sept. 2019), 891–904.e7. DOI: 10.1016/j.molcel.2019.06.024. URL: https://doi.org/10.1016/j.molcel.2019.06.024.

[34] S. Gao, X. Li, and B. A. Amendt. "Understanding the Role of Tbx1 as a Candidate Gene for 22q11.2 Deletion Syndrome". In: *Current Allergy and Asthma Reports* 13 (2013), 10.1007/s11882-013-0384–6. DOI: 10.1007/s11882-013-0384-6.

[35] R. Fetit, D. J. Price, S. M. Lawrie, and M. Johnstone. "Understanding the clinical manifestations of 16p11.2 deletion syndrome: a series of developmental case reports in children". In: *Psychiatric Genetics* 30 (2020), pp. 136–140. DOI: 10.1097/YPG.0000000000000278.

[36] N. Vos et al. "Evaluation of 100 Dutch cases with 16p11.2 deletion and duplication syndromes; from clinical manifestations towards personalized treatment options". In: *European Journal of Human Genetics* 32 (2024), pp. 1387–1401. DOI: 10.1038/s41431-024-01572-0.

[37] Muscular Dystrophy Association. *Causes/Inheritance - Spinal Muscular Atrophy (SMA)*. https://www.mda.org/disease/spinal-muscular-atrophy/causes-inheritance. Accessed: 2025-01-24. 2015.

[38] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21. DOI: 10.1093/bioinformatics/bts635.

[39] P. Danecek et al. "Twelve years of SAMtools and BCFtools". In: *Gigascience* 10.2 (Feb. 2021), giab008. DOI: 10.1093/gigascience/giab008.

[40] S. Chen. "Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp". In: *iMeta* 2 (2023), e107. DOI: 10.1002/imt2.107. URL: https://doi.org/10.1002/imt2.107.

[41] Y. Liao, G. K. Smyth, and W. Shi. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". In: *Bioinformatics* 30.7 (2014), pp. 923–930. DOI: 10.1093/bioinformatics/btt656.

[42] M. I. Love, W. Huber, and S. Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15 (2014), p. 550. DOI: 10.1186/s13059-014-0550-8.

[43] G. Yu, L.-G. Wang, Y. Han, and Q.-Y. He. "clusterProfiler: an R package for comparing biological themes among gene clusters". In: *OMICS: A Journal of Integrative Biology* 16.5 (2012), pp. 284–287. DOI: 10.1089/omi.2011.0118.

[44] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org.

[45] K. Blighe, S. Rana, and M. Lewis. *EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling*. R package version 1.24.0. 2024. URL: https://github.com/kevinblighe/EnhancedVolcano.

[46] R. Kolde. *pheatmap: Pretty Heatmaps*. R package version 1.0.12. 2018. URL: https://github.com/raivokolde/pheatmap.