参赛队员姓名：杨珺萌

中学：上海星河湾双语学校

省份：上海市

国家/地区：中国

指导教师姓名：白永胜

论文题目：Identification and Genetic Signature Analysis of MiRNA-targeted Region SNVs in Intellectual Disability

本参赛团队声明所提交的论文是在指导老师指导下进行的研究工作和取得的研究成果。尽本团队所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果。若有不实之处，本人愿意承担一切相关责任。

参赛队员：杨珺萌　　指导老师：白永胜

2020年　9月　14　日

Identification and Genetic Signature Analysis of MiRNA-targeted Region SNVs in Intellectual Disability

Junmeng Yang [1]*

[1] Shanghai Starriver Bilingual School, Shanghai, China

* Corresponding author. Tel.: +86 13661703632; email: jasmine.yang2015@qq.com

Abstract: Intellectual Disability (ID) cases often involve multiple genes and loci, thus creating extreme genetic and phenotypic heterogeneity. MiRNAs are short RNA sequences that regulate ~60% post- transcriptional protein-coding genes' expression in mammals. Exonic mutations in ID risk genes, mostly located in the coding sequence (CDS) and 3′ untranslated region (UTR), are often causative of diseases due to their impact on micro RNA (miRNA) targeting.

Alongside the discovery of novel SNVs associated with miRNA targeting, genetic signatures of candidate ID genes further helps the prioritization of candidate genes. In this study, motif conservation and protein- protein interactions among candidate genes are used to uncover relationships among these genes, while GO functional analysis, disease association, and tissue distribution of expression, are used to reveal correlation among genes and symptoms and/or syndromes. Exon number, in addition to genomic pattern and biological function, is further used in prioritizing X-linked miRNA targeted and sex-biased genes.

Results of identification and prioritization nonsynonymous SNVs (nsSNVs) harboring in the 3' UTR or CDS regions in our candidate ID genes are reviewed. In a recently published study about 3'UTR SNVs in ID with myself as the first author [1], genes *GPR88* and *WNT7A* are prioritized with SNVs with miRNA-targeting associations exclusively located in 3'UTR. Among 114 pairs of ID-related accurate miRNA–mRNA targeting pairs, 6 genome locations among 183 SNVs, which include numerous novel variants, are further prioritized. In my collaboration with Anna Liu (Manuscript submitted to International Conference on Bioscience, Biochemistry and Bioinformatics (ICBBB 2021)), *GNAO1*, *ACTB*, and *PTPN11* are prioritized by GO term clusters, brain expression, and motifs conservations. 175 SNVs in the three candidate genes' CDS were obtained, and further motif analysis and genetic signature study helped to prioritize the SNVs.

As ~80% of Non-syndromic ID genes reside on the X-chromosome, contributing to the high male-female ratio in ID patients, we further investigated gender bias in ID by obtaining targeted genes of 13 X-linked and sex-biased miRNAs associated with NS-ID and studying the genetic signatures of 73 brain-expressed sex- biased miRNA-targeted genes. Our result further shows that while the proportion of brain-expressed male- biased genes targeted by X-linked miRNAs is higher, sex-biased ID-related genes are more prevalent, tend to have more exon numbers, and interact significantly more with miRNA-targeted female-biased genes. Numerous associations between our prioritized genes and ID-related pathways, symptoms, and syndromes further indicate the role miRNAs and SNVs have in targeting plays in the context of ID. We plan to automate the methods through developing a bioinformatics pipeline for identifying novel CDS and 3'UTR SNVs harbored in miRNA-targeted genes in the future.

Key words: Intellectual disability, miRNA-targeting, single nucleotide variants

1. Introduction

According to the Diagnostic and Statistical Manual of Mental Disorders (DSM–5), intellectual disability (ID), also called mental retardation (MR), is categorized as neurodevelopmental disorders

(NDDs) along with autism spectrum disorder (ASD), neurodevelopmental motor disorders, and others [1]. Indeed, ~70% of ASD patients having co-occurring ID symptoms: common ASDs, including Down Syndrome (DS), Rett Syndrome, and Fragile X Syndrome (FXS) all have severe ID as one of its symptoms.

Many common ID syndromes, including Rett Syndrome and Fragile X Syndrome, are monogenetic. However, ID and ASD cases often involve multiple risk genes and loci, creating extreme genetic and phenotypic heterogeneity. As current research interest moves on from identifying monogenic causes of ID to exploring complicated genetic mechanisms and studying the genetic signature of ID genes, the discovery of disease-causative de no*vo* variants and selection of new candidate genes are increasingly important.

ID and ASD, which together affect 3-5% of the population [2-4], often involves *de novo* mutations in both coding and noncoding (untranslated) regions (CDS and UTR) of ID-related genes. For instance, a whole-genome sequencing applied to 50 patients with severe ID and their unaffected parents showed 84 *de novo* single nucleotide variants (SNVs) affecting the coding region (CDS), revealing a not only reveals an enrichment of genes previously identified in ID-related disorders, but also an enrichment of loss-of-function mutations [5].

Genetic signatures of verified and candidate ID genes reveals interactions among genes as well as correlation among genes and symptoms. An example of phenotypic symptom is dendritic spines dysgenesis, which is a symptom of most common ID syndromes, including Downs Syndrome, FXS, Rett Syndrome, Angelman Syndrome, and Tuberous Sclerosis (TS) [6-9]. Disruption of core neurodevelopmental signaling pathways also reveals potential correlations among ID-related genes: the evolutionarily-conserved Wingless (Wnt) signaling pathway, for instance, involves multiple ID-causing genes and is associated with multiple ID syndromes, including the FXS [10]. The upregulation of the canonical Wnt/β-catenin pathway in ASD further indicates the impact of pathway abnormalities.

With ID as a common symptom of autism and autistic features frequently present in ID patients, investigating non-syndromic ID (NS-ID), cases where presence of ID as the sole clinical feature, thus are likelier to have close association with ID.

Sex biases in ID has also raised much research interest: a male-to female ratio of ~4:1 in autism and ~1.3:1 in ID have been previously reported [11, 12]. As more than 100 ID genes are identified on the X-chromosome, sex differences in the expression of the ~1,100 genes located solely on the X chromosome are likely to underlie many sex differences in the expression of diseases affected by these genes. In fact, among the ~40 genes proven to cause NS-ID, ~80% of them reside on the X-chromosome, contributing to the high male-to-female ratio in the NS-ID population [13]. In fact, X-linked ID (XLID) is estimated to account for 10-12% of ID, which means the abundance of ID genes on the X chromosome contribute to the ~1.4-fold excess of ID-affected males compared to females [14].

Though ID-causing X-linked microRNAs (miRNAs) are not well-studied, existing literatures have demonstrated the potential role miRNAs in ID. SNVs within the 3′UTR regions of susceptible genes may affect miRNA binding affinity, and will thus be causative to ID-related genes. For instance, *miR-137*, the only miRNA shown to be related to ID according to Human microRNA Disease Database (HMDD) [15] that plays an important role in neural development and neoplastic transformation, has mutations causative to Alzheimer's disease, Huntington Disease, Schizophrenia, etc. [16]. Its target genes include *GPR88* and *WNT7A*, both of which are involved in neural development, specifically, neural stem cell proliferation and differentiation [17]. Dendritic spine regulation, as mentioned, is also associated with regulation of 3 -UTR elements from *WNT7A* and *GPR88* by *miR-137 [17, 18]*. Furthermore, miR-130a's abnormal regulation of *MECP2* expression causes Rett Syndrome, in which a *MECP2* missense mutation causes much severer ID phenotype in males than in females [19]. Another example is the FXS, in which the expression of the 39 untranslated region (39UTR) of the *FXR1* gene is significantly increased in the absence of miRNAs [20]. With miRNA's critical role in regulating gene expression, a systematic analysis of ID-related miRNA could uncover part of the ID genetic landscape.

In this project, we selected candidate miRNA-targeted ID genes, identified and prioritized novel ID-related SNVs harbored in both 3'UTR and CDS regions of ID candidate genes, and further investigated the genomic pattern and biological function differences for miRNA-targeted sex-biased genes.

2. Method

2.1. miRNA-Gene Pairs Selection and input in dbMTS

Though there are multiple miRNA targeting databases, to obtain accurate ID-related miRNA–mRNA pairings, we collected known and predicted pairs of miRNA and target genes from published literatures [16, 21-26]. By manually screening for accurate miRNA–mRNA targeting pairs from related literatures, we obtained a total of 114 pairs.

We then used a customized python script pipeline to format the mRNA and miRNA candidate input pairs that the dbMTS database accepts with Ensembl ID downloaded from Ensembl Biomart (www.biomart.org) [27] and pre-miRNAs' corresponding mature miRNA numbers with hairpin.fa and mature.fa downloaded from miRbase (www.mirbase.org) [28].

2.2. Genetic Signature Study of Candidate ID Gene with 3′UTR SNVs

We found the SNVs identified in these three genes novel by checking the IDGenetics database (www. ccgenomics.cn/IDGenetics/), UCSC Genome Browser (genome.ucsc.edu), and dbSNP of NCBI (www. ncbi.nlm.nih.gov/SNP) [29-31].

In order to determine the functionality of the three genes with SNVs exclusively located in the 3'UTR, namely *GPR88, WNT7A,* and *CDK6*, we studied their Gene Ontology (GO) annotation, associated pathways, and expression distribution among different tissues.

The expression distribution data were obtained from the Genotype-Tissue Expression (GTEx) project (www.gtexportal.org) [32], which provides significant variant and gene associations based on permutations of genes among 53 human body tissues, 13 of which are located in the brain. By normalizing the data, we generated a heatmap using R to represent the significantly higher expression of three genes, namely *GPR88*, *WNT7A*, and *CDK6*, in brain tissues. The pathways each gene was involved in were obtained from the Kyoto Encyclopedia of Genes and Genomes Pathways (KEGG Pathways) (www.genome.jp/kegg/pathway.html), while the diseases that genes were causative to were obtained from KEGG Disease (www.genome.jp/kegg/disease) [33].

We chose mice and rats as representative species and obtained the *WNT7A* and *GPR88* 3′UTR sequences from Ensembl. We then checked the 3′UTR sequence as well as the location of the SNV using UCSC Genome Browser, and searched for motifs of a corresponding 3′UTR region within the *GPR88* gene from the Multiple Expression motifs for Motif Elicitation (MEME) Suite (meme-suite.org/db/motifs) [34].

To guarantee that the ratio of the MEME-identified motif's and its reverse-complement's appearances in the 3′UTR of transcripts is significant, we extracted 100-nucleotide long sequence with discovered SNVs in the middle, searched again for motifs conservation among representative species using MEME, and the Fisher exact test was employed to compare the result with identified motifs' appearance in 3'UTR sequences of random genes.

Above mentioned steps in selecting miRNA–mRNA target pairs and subsequent genetic signature study are shown in **Fig. 1**.

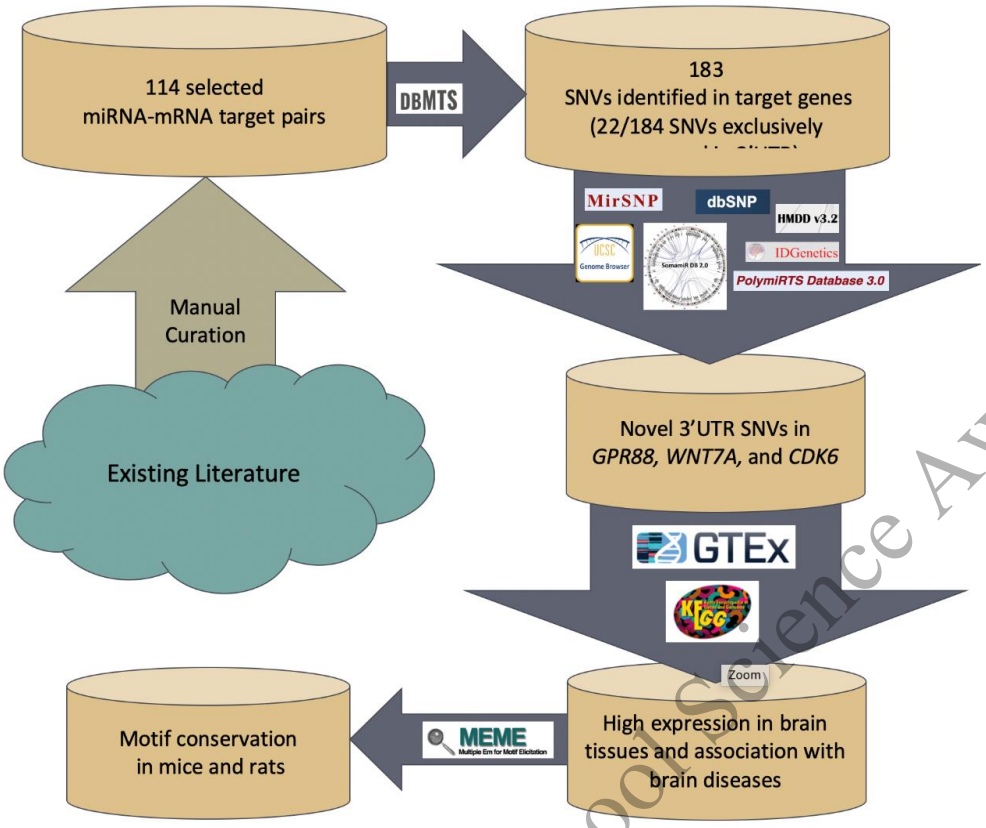


Fig. 1. Work flow of prioritizing 3'UTR SNVs in candidate ID-causing genes

2.3. Verification and Prioritization of Identified SNVs in CDS of candidate genes

In collaboration with Anna Liu (Manuscript submitted to International Conference on Bioscience, Biochemistry and Bioinformatics (ICBBB 2021), we obtained three genes (*ACTB*, *GNAO1*, and *PTPN11*) from motif identification result of genes with conserved domains and high expressions in brain tissues selected from published literatures. We then prioritized this list of genes harboring nsSNVs based on their damaging significance reported by dbNSFP and ClinVar (ncbi.nlm.nih.gov/) [35] phenotype (supporting functional consequence of variants). We used MEME to identify the motifs among the CDS sequences and inspected whether genes enriched with identified motifs containing non-synonymous SNVs contribute to ID symptoms. With the database for Nonsynonymous SNPs' Functional Predictions (dbNSFP) [36], we further obtained functionally annotated candidate SNVs exclusively located in the CDS region.

2.4. Sex-biased miRNA and Genes Collection and Curation

In a previous study [37], researchers have identified 13 brain-expressed X-chromosomal miRNAs in a cohort of 464 patients with non-syndromic XL-ID. By checking the 13 miRNAs against DIANA-TarBase-v8 [38], we obtained a total of 8,508 miRNA-mRNA targeting pairs, reflecting 4,911 unique genes targeted by 13 miRNAs.

Guo and her colleagues [39] have comprehensively searched gene expression data sets from the GEO database. Using the criteria including human tissues samples, control samples, and the existence of 'sex' information in the data sets, they reported 194 unique sex-biased genes expressed in brain tissues. By cross-checking the 4,911 genes obtained from TarBase, we obtained 73 sex-biased genes that are targeted by 13 brain-expressed X-chromosomal miRNAs.

2.5. Gene Set Enrichment, Transcript Diversity, and Protein-Protein Interacrtion Analysis for Sex-biased Genes

To analyze sex-biased genes, we employed Enrichr to perform a gene set enrichment analysis, including gene ontology (GO) and disease association. We split the 73 input genes into 52 male-biased genes and 21 female-biased genes to report analysis results separately. For male-biased genes, we only selected terms with an adjusted p-value lower than 0.01, and for female-biased genes, we selected terms with a p- value lower than 0.01, since the list of female-biased genes was a smaller size.

We have also analyzed the exon numbers for 73 sex- biased genes targeted by the 13 miRNAs to see if the sex- biased genes tend to have more isoforms than random sets of genes in the human genome. We used the total numbers of exons in each dataset to estimate the transcript diversity. The p-value for a set of genes' exons is estimated as the proportion of randomized gene sets that has a set whose total number of exons is equal to or greater than the one detected in the sex-biased 73 genes set.

We employed the STRING database to study the protein-protein interaction (PPI) pattern of the male-biased and female-biased genes. Specifically, the list of male-biased brain-expressed genes resulting from a cross-check with the TarBase result consists of 52 genes, while that of the female-biased consists of 21 genes. This allows us to elucidate the scenario between gender-biased sets of genes related to ID.

To test the significance of male and female biased genes, we also ran the STRING database for all sex-biased genes reported by Guo and her colleagues [26]. We then used the Fisher's Exact test to determine if the PPI of the targeted brain-expressed genes is significant compared to that of all sex-biased genes while excluding ambiguous ones.

Above mentioned steps in prioritizing X-linked miRNAs and Genes associated with ID are shown in **Fig. 2.**

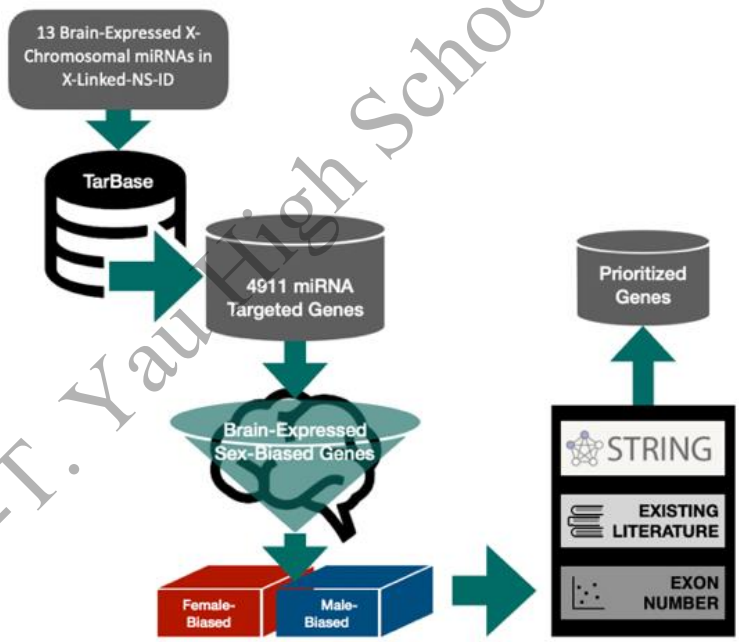


Fig. 2. Work flow of prioritizing X-linked miRNAs and Genes associated with ID

3. Results

3.1. miRNA-Gene Pairs Selection and input in dbMTS

By manually screening for predicted and/or experimentally validated pairs of miRNAs and each of their target mRNA, we obtained 18 pairs from the dataset provided in a previous study [36]. Similarly, we obtained 23 pairs from another study [35] as well as 45 genes experimentally validated to be targeted by miR-137 [24]. Genes including *SHROOM3*, *FCER1A*, *DLEC1*, and *PSAPL1* are shown to be targeted by miR-145; genes *ROBO2*, *CHAF1B*, *P2RY22*, *XRRA1*, *GPX8*, *P2RY2*, and *NSG1* are targeted by miR-183 [33]. Gene *ARMCX2*, *NEFH*, and *FMR1* are also shown to be targeted by *miR-222*

[37]. The gene *MECP2*, which is related to IDs like Rett Syndrome, can be targeted by *miR-302c*, *miR-483–5p*, *miR-130a*, and *miR-200a*. The pair of *miR-128* and *PHF6* [34] and the pair of *miR-155* and *AGTR1* [32] are similarly obtained.

Although dbMTS reports all SNVs for single reference positions, some cases could fall into different genomic classification regions. Variant could be specific for different isoforms. 22 of the SNVs were exclusively located in the 3′UTR of all possible transcripts. Those 22 highly confident 3'UTR SNVs are harbored within three genes: *GPR88*, *WNT7A*, and *CDK6*, as shown in Table 1.

Table 1. Variants located in 3′ UTR regions that alter miRNA targeting based on dbMTS

chr	pos	ref	alt	VEP_ensembl_Gene_Name
1	100541492	T	G	*GPR88*
3	13818428	G	T	*WNT7A*
3	13818646	T	A	*WNT7A*
3	13818646	T	C	*WNT7A*
3	13818646	T	G	*WNT7A*
3	13818647	A	C	*WNT7A*
3	13818647	A	G	*WNT7A*
3	13818647	A	T	*WNT7A*
3	13818648	T	A	*WNT7A*
3	13818648	T	C	*WNT7A*
3	13818648	T	G	*WNT7A*
3	13818649	T	A	*WNT7A*
3	13818649	T	C	*WNT7A*
3	13818649	T	G	*WNT7A*
3	13818650	G	A	*WNT7A*
3	13818650	G	C	*WNT7A*
3	13818650	G	T	*WNT7A*
7	92605125	T	A	*CDK6*
7	92607925	A	G	*CDK6*
7	92607993	A	C	*CDK6*
7	92607993	A	G	*CDK6*
7	92607993	A	T	*CDK6*

1 The first column, "chr", indicates the chromosome number where the SNV is found; the second column, "pos", refers to the position of each SNV on the chromosome. The third and fourth column, "ref" and "alt", show reference nucleotide and the mutant nucleotide respectively, and thereby presenting the single nucleotide variant. The last column, "VEP_ensembl_Gene_Name", indicates the gene name in which the SNV occurs.

3.2. Genetic Signature Study of Candidate ID Gene with 3′UTR SNVs

We obtained expression distribution data in 53 tissues provided by Genotype-Tissue Expression (GTEx). Normalized data is shown in **Fig. 3.**

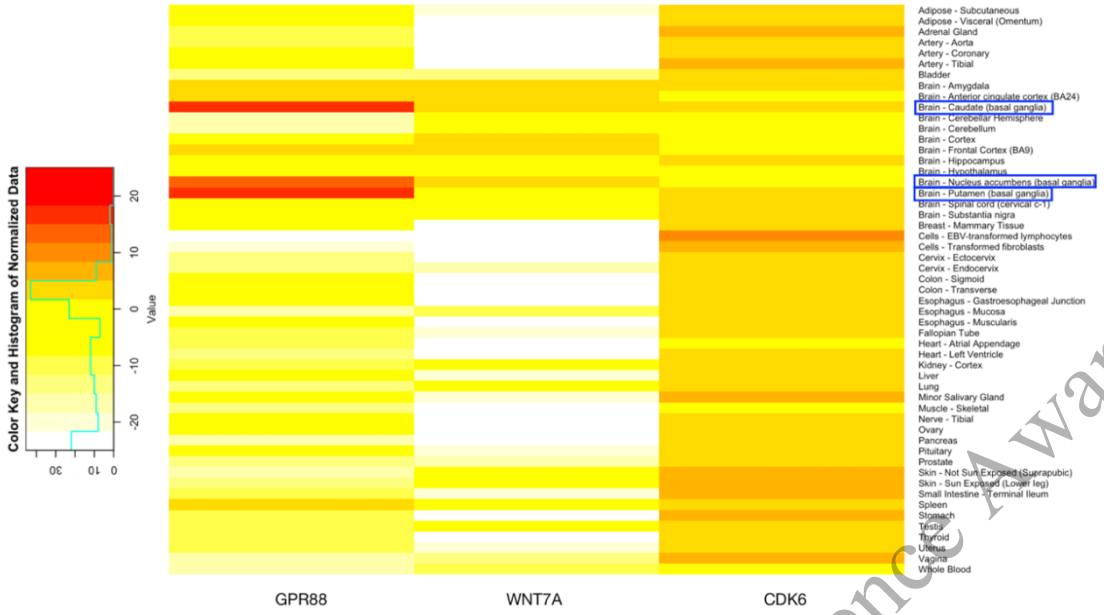


Fig. 3. Expression Data of the genes GPR88, WNT7A, and CDK6 obtained from GTEx.

GPR88 is over-expressed in tissues including Caudate, Nucleus accumbens, and Putamen, all of which are in the brain, while gene WNT7A is over-expressed in Amygdala, Anterior cingulate cortex, Caudate, Cortex, Frontal Cortex, and Nucelus accumbens of the brain.

GO analysis of GPR88 suggests that is involved the molecular function G-protein coupled activity as well as multicellular organism development. The gene is also shown to be the only related gene to the disease Chorea, childhood-onset, with psychomotor retardation, by KEGG. WNT7A is shown to be involved in all three GO functional enrichment, namely biological processes regarding the Wnt signaling pathway, molecular function in terms of Wnt-protein and Wnt-activated receptor binding and Wnt-protein binding, and cellular component in/on the cell and vesical membranes. Wnt signaling pathway will be discussed in detail in the following sections. On the other hand, CDK6 is suggested to interact with Cyclin-dependent kinase 4 inhibitor and Regulatory component of the cyclin. All biological process, molecular function, and cellular component of GO functional enrichment suggested by UniProt are related to the cell cycle, further indicating the fact that it is a ubiquitous gene and is more related cancer.

All biological process, molecular function, and cellular component of GO functional enrichment of all three genes, namely GPR88, WNT7A, and CDK6, are conserved in multiple species. MEME identified many motifs of the gene WNT7A in humans, mice, and rats. Among them, the motif that contains five out of the six SNVs in WNT7A that are identified by dbMTS is the third most significant. With an E-value of 2.1e-10, the motif ranges from 13818626 to 13818675 in the human genome, with the five neucleotides TATTG (in reverse-complement order since the template strand of WNT7A in humans is on the negative strand) in the middle. In addition, the human WNT7A gene has a p-value of 4.69e-29 to the motif. **Fig. 4(a)** shows the five neucleotides as boxed. MEME search result does not give a significant motif that includes the variant identified by dbMTS in GPR88 genes in humans, mice, and rats. The variant, however, is found to be conserved among humans and mice.

(a)

(b)

Fig. 4. Motif conservation of genes *GPR88* and *WNT7A*. Each graph obtained from MEME represents a motif discovered. **(a)** The motif in WNT7A with the boxed pos. 13818650 – 13818646 in reverse complement order. **(b)** The motif containing the SNV at pos. 100541492 is in GPR88, conserved among humans and mice.

3.3. Functional Annotation and Tissue Expression for Coding Regions

Among a total of 2,066 ID candidate genes from published literatures, gene expression in brain and gene clusters are first used in candidate gene selection. MEME is then employed to obtain the top three motif conservation among the prioritized nine genes. With phylogenetic information, resulting in three genes, namely *PTPN11*, *ACTB*, and *GNAO1*. A total of 175 SNVs located in the three genes are then obtained from dbSNFP.

Among the 175 SNVs, eight variant locations, including 10 SNVs, are located within the three most significant motifs. Though none of the SNVs reported by dbNSFP for gene *PTPN11* are located in the three most significant motifs, PTPN11, with other genes, does share the second-most significant motif with a p-value of 5.55e-18 reported by MEME. SNVs' locations in motifs are shown in **Fig. 5.**

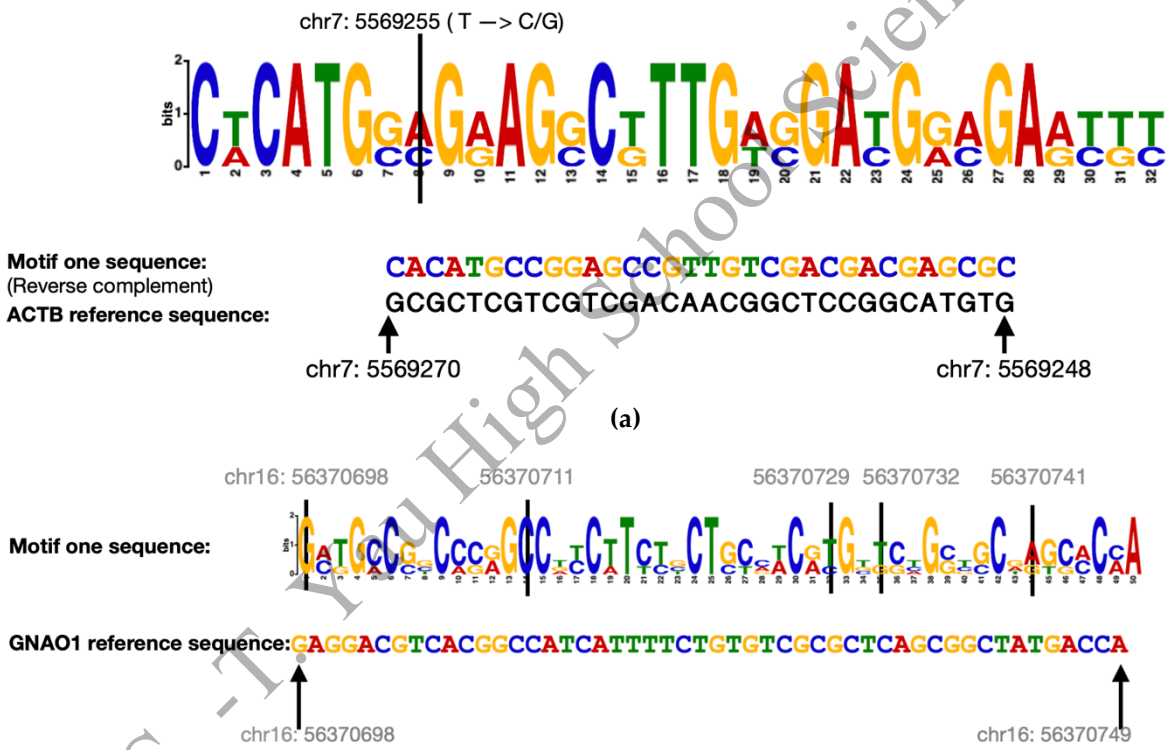


Figure 5. ACTB and GNAO1 SNV locations in prioritized motif sequence. **(a)** ACTB. **(b)** GNAO1

After observing damaging significance and phenotypic (disease) association reported by dbNSFP and ClinVar, we studied the clinical significance of the SNVs. **Table 2** summarizes our discovery and the ten SNVs.

All three variant locations in *ACTB* that are included in the motifs are pathogenic (causative to the Baraitser-winter syndrome). On location chr7:5569255, there reside three SNVs that are located in the most significant motif with a p-value of 2.48e-15, as shown in **Fig.6** and **Table 2**. Baraitser-winter syndrome is a multiple congenital anomaly syndrome characterized by intellectual disability, thus suggesting the high possibility for this variant to also be associated with ID. Similarly, the third SNV

in *ACTB* located at chr7:5569225 with an amino acid index of 22 reported a single variant A>T and another connection with Baraitser-winter syndrome.

The splice donor site variant (position chr16:56226265) in *GNAO1* is identified as two potential SNVs (G>R and G>W) inside the first exon sequence (chr16:56226148-chr:56226266), in which the second motif is located. Clinical significance on ClinVar reported both SNVs as pathogenic and associated with early infantile epileptic encephalopathy (EIEE1), a type of neurological disorder caused by brain malformation or genetic mutations caused by the lack of ARX proteins which disrupts normal brain development and leads to seizures and intellectual disability (reference). Our results verified the strong correlation between GNAO1 variant mutations and EIEE1 that leads to ID later in life.

Furthermore, a group of five locations in GNAO1, specifically chr16:56370698, chr16:56370711, chr16:56370729, chr16:56370732, and chr16:56370741 of chromosome 16, are all located in the third most significant motif. The third most significant motif, also shared by gene ACTB, has a p-value of 1.78e-18 in gene GNAO1. The fact that the five SNVs are located close to each other but not consecutive further suggests the potential influence through miRNA binding. Clinical reports on ClinVar also proved two of the five SNVs- chr16:56370732 and chr16:56370741-to be probably pathogenic and pathogenic respectively. Out of the remaining three variants, two were novel variants we discovered at chr16:56370698 and chr16:56370711, and although the last SNV at chr16:56370729 did not report any clinical significance, we believe all three SNVs are highly likely to be pathogenic and causative of intellectual disability and other related syndromes as were the other SNVs we identified. In the future, clinical experiments could seek to verify our results.

Table 2. 10 Non-synonymous Variants on 8 locations in *ACTB* and *GNAO1* obtained from dbNSFP

chr	pos (hg19)	ref	alt	aaref	aaalt	Gene	Location	Motif	Clinical Significance
7	5569225	C	T	A	T	ACTB	CDS	1	Pathogenic
7	5569255	T	C	N	D	ACTB	CDS	1,3	Pathogenic
7	5569255	T	G	N	H	ACTB	CDS	1,3	Pathogenic
16	56226265	G	C	G	R	GNAO1	Splice donor site	2	Likely pathogenic
16	56226265	G	T	G	W	GNAO1	Splice donor site	2	Likely pathogenic
16	56370698	G	A	E	K	GNAO1	Splice acceptor site	3	Novel variant
16	56370711	C	A	A	D	GNAO1	CDS	3	Novel variant
16	56370729	C	T	A	V	GNAO1	CDS	3	Clinical significance not reported
16	56370732	T	C	L	P	GNAO1	CDS	3	Likely pathogenic
16	56370741	A	G	Y	C	GNAO1	CDS	3	Likely pathogenic

3.4. Sex-biased miRNA and Gene Selection and Curation

Upon checking the 13 miRNAs against TarBase [25] for the sex-biased gene list [26], we obtained sex-biased miRNA-targeted brain-expressed gene.

Among the 13 miRNAs, *hsa-miR-105-5p* and *hsa-miR-223-5p* exclusively target male-biased genes. In addition, the majority of the 13 miRNAs target both male-biased and female-biased genes, except that *hsa-miR-504* does not target any of the 73 genes. In the list of female-biased genes, 9 of 21 genes are located on the X chromosome, and two X-linked miRNA families dominate of these. Specifically, hsa-miR-19b targets seven genes (NR3C2, PLXNA4, CAPRIN2, LUC7L, KDM6A, STS, and ZFX),

while hsa-miR-221/222 targets ten genes (NKTR, PPP1R2, CCSER1, TET2, KIF16B, DDX3X, KDM5C, SMC1A, TXLNG, and USP9X).

3.5. GO Analysis, Transcript Diversity, and Protein-Protein Interaction Analysis for Sex-biased Genes

The GO and ClinVar disease association for male-biased genes are shown in **Table 3**. ClinVar reports that genes *COL1A1*, *COL1A2*, and *SERPINH* are related to osteogenesis imperfecta (OI) and postmenopausal osteoporosis. The Gene Ontology Biological Process data of the 52 male-biased genes also shows collagen fibril organization (GO:0030199). As the mineral density increases, the tensile modulus of the network increases monotonically, well beyond that of pure collagen fibrils (Nair). Both the GO Biological Process data and ClinVar disease association data suggests a correlation between ID and osteogenesis and osteoporosis, which will be discussed in detail in the next section.

Table 3. GO Analysis and Disease Association of Male-biased Genes

Gene Name	Enrichr Analysis (Adjusted P-value)	
	Gene Ontology Biological Process	ClinVar Diseases Association
COL1A1	Collagen fibril organization -GO:0030199 (0.00469198)	osteogenesis imperfecta (3.57E-04); postmenopausal osteoporosis (0.00599324)
COL1A2	Collagen fibril organization -GO:0030199 (0.00469198)	osteogenesis imperfecta (3.57E-04); postmenopausal osteoporosis (0.00599324)
COL3A1	Collagen fibril organization -GO:0030199 (0.00469198)	N/A
SERPINH1	Collagen fibril organization -GO:0030199 (0.00469198)	osteogenesis imperfecta (3.57E-04);

As exhibited in **Table 3**, *CAPRIN2* and *DDX3X* are shown to be involved in the Wnt signaling pathway by GO Biological Pathways. In fact, another prototypic member of the Caprin (protein family is a novel *FMRP* cellular partner which interacts with *FMR1* at the level of the translation machinery [30].

CAPRIN2 is also shown to be involved in dendritic spine dysgeneses, which, as aforementioned, is a phenotypic signature for most of the common ID syndromes, including DS, FXS, Rett Syndrome, Angelman Syndrome, and Tuberous Sclerosis (TS). A previous study [7] has also identified *DDX3X* as a regulator of the Wnt-β-catenin network as a critical factor in canonical Wnt signaling.

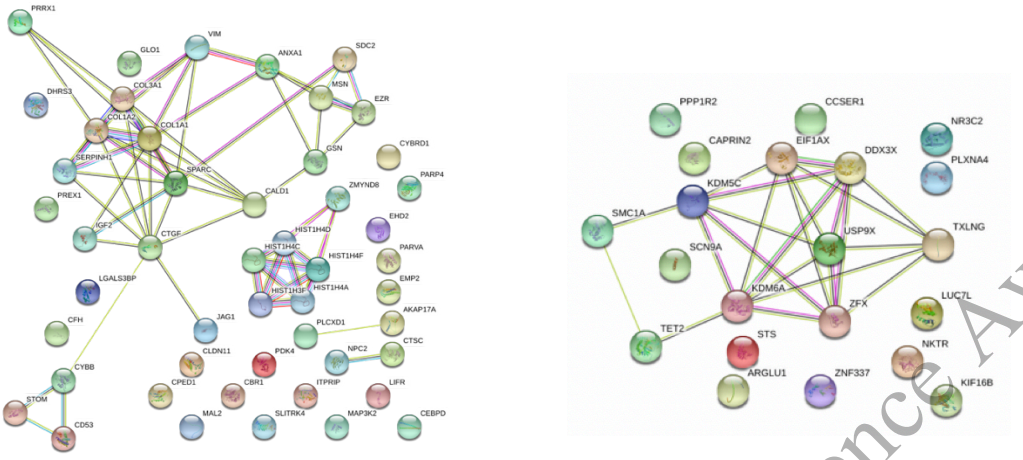
Table 4. GO Analysis of Female-biased Genes

Gene Name	Gene Ontology (Biological Process)	p-value
TET2;KDM6A	histone H3-K4 methylation (GO:0051568)	2.62E-04
CAPRIN2; PLXNA4	positive regulation of cell morphogenesis involved in differentiation (GO:0010770)	0.00167527
DDX3X; CAPRIN2	positive regulation of canonical Wnt signaling pathway (GO:0090263)	0.0061944
SMC1A	response to DNA damage checkpoint signaling (GO:0072423)	0.00628419
DDX3X; CAPRIN2	negative regulation of cell growth (GO:0030308)	0.00640853
DDX3X; CAPRIN2	negative regulation of growth (GO:0045926)	0.00695849
SMC1A	regulation of DNA endoreduplication (GO:0032875)	0.0073279
KDM5C	histone H3-K4 demethylation (GO:0034720)	0.0073279
DDX3X	protein localization to cytoplasmic stress granule (GO:1903608)	0.0073279
CAPRIN2	positive regulation of dendritic spine morphogenesis (GO:0061003)	0.0073279
DDX3X; KIF16B	regulation of macromolecule metabolic process (GO:0060255)	0.00800061
DDX3X; CAPRIN2	positive regulation of Wnt signaling pathway (GO:0030177)	0.0093639
PLXNA4	semaphorin-plexin signaling pathway involved in axon guidance (GO:1902287)	0.00941218
TET2	histone H3-K4 trimethylation (GO:0080182)	0.00941218

To check transcript variety, we counted the total number of exons of sex-biased genes by considering the isoform with the maximum number of exons for each gene in the list. We further ran 10,000 random gene sets selected from the human genome with the size of 73 genes for each set.

Statistical results show that 28 of the gene sets have a higher number of total exons than that of the sex-biased gene set (p-value = 0.0028).

The result of running the STRING database [40] for the 73 selected male-biased and female-biased gene lists is shown in **Fig. 6**.



(a) **(b)**

Fig. 6. Protein-Protein Interaction graph obtained from STRING Discussion
(a) Male-biased, brain-expressed genes targeted by the 13 miRNAs
(b) Female-biased, brain-expressed genes targeted by the 13 miRNAs

According to STRING, both PPI networks have significantly more interactions than random protein sets from the human genome, with a PPI enrichment p-value of 4.44e-16 for female-biased genes and <1.0e-16 for male-biased genes.

Out of 52 male-biased genes, 49 have corresponding proteins in the STRING database. We noticed that the majority of genes in two major clusters are from the same protein families (i.e. type 1 collagen and the histone H4 family.) *SERPINH1* is reported to interact heavily as shown in the combined score with other type 1 collagen family proteins (*COL3A1*: 0.959, *COL1A2*: 0.974, *COL1A1*: 0.973), causing osteogenesis imperfecta. In the female-biased gene list consisting of a total of nine genes on the X-chromosomes, eight of them have proteins interacting with each other within the same cluster. The *DDX3X* protein has the most interactions among this cluster.

We employed the STRING database again and found 3,351 edges among the 997 male-biased genes excluding the 52 brain-expressed and miRNA-targeted genes, and 6,176 edges among the 1,335 female-biased genes excluding the 21 brain-expressed and miRNA-targeted genes. The results show that miRNA-targeted brain-expressed genes is significant between genders (Fisher's Exact Test p-value < 0.00001).

4. Discussion

The basal ganglia, part of the subcortical structures of the brain, is involved in memory, emotion, pleasure, and hormone production. A research study [41] about the *ARX* gene suggests that it contributes to milder XLID. In fact, not only is the basal ganglia known to regulate sensorimotor processing, ARX patients also have a significantly decreased volume of brain structures including the nucleus of caudate, in which *GPR88* and *WNT7A* overexpresses; as shown in **Fig.2**.

GNAO1 encodes proteins that represent the alpha subunit of the G-protein signal-transducing complex. Mutations within GNAO1 have been found to cause neurodevelopmental disorder with involuntary movements (NEDIM) and EIEE1, both of which are severe neurologic disorders within the brain. Moreover, in another separate study (MS in preparation), PTPN11 and ACTB were both reported to be targeted by hsa-miR-221. This indicates that both genes likely play similar roles in ID etiology.

Previous study has also studied 12 brain-expressed XL pre-miRNAs and their corresponding 18 mature miRNAs [42], showing that eight miRNAs, including *miR-221-3p/222-3p, miR-504-5p.1, miR-505-3p.1, and miR-505-3p.2*, are essential regulators ID genes in a wholly connected network. In addition, the researched identified no sequence variations, indicating an intense selective pressure.

The two sex-biased genes shown to be involved in the Wnt signaling pathway, *DDX3X* and *CAPRIN1*, both reaches genome-wide significance for *de novo* variants as female-biased genes in data based on NDD patients [12]. Through comparing the number of SNVs in NDD-affected males and females, *DDX3X* is also identified as the only replicated genome-wide significant gene by Turner and his colleague [12]. Furthermore, ID-affected males carry less severe *DDX3X* mutations compared females. Thus, Wnt-β-catenin signaling depends heavily on *DDX3X*, which further demonstrating the importance of the Wnt pathway in proper neurodevelopment [10]. *DDX3Y*, the male-specific region of Y chromosome gene, significantly over-expresses neural differentiation and potentially plays a multifunctional role in neural cell development in a sexually dimorphic manner [43].

The 4,911 genes found to be targeted by TarBase consist nine out of ten of ASD-risk genes suggested by Fernandez et al. (with NAV1 and NAV2 being aliases of gene SCN2A and SHANK3 not found in the gene list). All ten genes have a neurobehavioral phenotype of ID, except for ANK with currently unknown phenotype [44]. Dendritic Spine Dysgenesis is consistently in the cortex and hippocampus of patients with Down Syndrome, the most common ID syndrome. In Angelman Syndrome, the E3 ubiquitin ligase Ube3A, which abnormal levels are also been found in Fragile X syndrome (FXS) and Tuberous Sclerosis (TS) controls spine formation [9]. Moreover, experiments on mice models with the knockdown of gene MeCP2 (causative to Rett Syndrome in humans) results in lowered dendritic spine density of all principal neurons of the hippocampus [8]. FXS, the most common inherited form of mental retardation, is further characterized by autistic behaviors, childhood seizures and abnormal dendritic spine [6].

5. Conclusion

Existing literatures have reported miRNAs and their target genes associated with ID, but genetic variants located in the 3′UTR region have not been elucidated for the target relationship creation and/or disruption [1]. In this study, we discovered novel variants in the 3′UTR and CDS of genes associated with intellectual disability and targeted by miRNAs. In studying the gender-bias in ID, we found that there are more brain-expressed male-biased genes than female-biased genes targeted by X-linked miRNAs, though X- linked ID-related genes are more prevalent in females. Our result further indicates that miRNA-targeted female- biased genes interact significantly higher than male-biased genes do.

Our study provides the bioinformatic process of prioritizing SNVs involved in miRNA–mRNA target relationship associated with diseases and informative aspects in studying sex-bias in specific diseases. We look forward to developing a bioinformatics pipeline for identifying novel 3′UTR SNVs harbored by miRNA-targeted genes and analyzing sex-bias information for other diseases. Future experimental validation can also be performed to validate if our identified novel 3′UTR and CDS variants for GPR88, WNT7A, ACTB, and GNAO1 serve as the miRNAs binding sites.

Conflict of Interest

The author declares no conflict of interest.

Acknowledgment

The author thanks Dr. Yongsheng Bai for aids in methodology, analysis, and supervision; Anna Chang Liu for aids in resources, methodology, and analysis; Isabella He for aids in methodology; Susan Huang for aids in resource; and Eric Li for aids in resource.

References

1. AmericanPsychiatricAssociation, *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed. 2013, Arlington, VA.
2. Braam, W., et al., *Low maternal melatonin level increases autism spectrum disorder risk in children.* Res Dev Disabil, 2018. **82**: p. 79-89.
3. Coll-Tané, M., et al., *Intellectual disability and autism spectrum disorders 'on the fly': insights from Drosophila.* Dis Model Mech, 2019. **12**(5).
4. Srivastava, A.K. and C.E. Schwartz, *Intellectual disability and autism spectrum disorders: causal genes and molecular mechanisms.* Neurosci Biobehav Rev, 2014. **46 Pt 2**: p. 161-74.
5. Gilissen, C., et al., *Genome sequencing identifies major causes of severe intellectual disability.* Nature, 2014. **511**(7509): p. 344-7.
6. Darnell, J.C., et al., *FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism.* Cell, 2011. **146**(2): p. 247-61.
7. Dindot, S.V., et al., *The Angelman syndrome ubiquitin ligase localizes to the synapse and nucleus, and maternal deficiency results in abnormal dendritic spine morphology.* Hum Mol Genet, 2008. **17**(1): p. 111-8.
8. Li, W. and L. Pozzo-Miller, *Beyond Widespread Mecp2 Deletions to Model Rett Syndrome: Conditional Spatio-Temporal Knockout, Single-Point Mutations and Transgenic Rescue Mice.* Autism Open Access, 2012. **2012**(Suppl 1): p. 5.
9. Phillips, M. and L. Pozzo-Miller, *Dendritic spine dysgenesis in autism related disorders.* Neurosci Lett, 2015. **601**: p. 30-40.
10. Kwan, V., B.K. Unda, and K.K. Singh, *Wnt signaling networks in autism spectrum disorder and intellectual disability.* J Neurodev Disord, 2016. **8**: p. 45.
11. Harripaul, R., et al., *The Use of Next-Generation Sequencing for Research and Diagnostics for Intellectual Disability.* Cold Spring Harb Perspect Med, 2017. **7**(3).
12. Turner, T.N., et al., *Sex-Based Analysis of De Novo Variants in Neurodevelopmental Disorders.* Am J Hum Genet, 2019. **105**(6): p. 1274-1285.
13. Migeon, B.R., *Why females are mosaics, X-chromosome inactivation, and sex differences in disease.* Gend Med, 2007. **4**(2): p. 97-105.
14. Garber, K., S. Warren, and J. Visootsak, *Fragile X Syndrome and X-Linked Intellectual Disability.* Emery and Rimoin's Principles and Practice of Medical Genetics, 2013: p. 1-27.
15. Huang, Z., et al., *HMDD v3.0: a database for experimentally supported human microRNA–disease associations.* Nucleic Acids Research, 2018. **47**(D1): p. D1013-D1017.
16. Mahmoudi, E. and M.J. Cairns, *MiR-137: an important player in neural development and neoplastic transformation.* Mol Psychiatry, 2017. **22**(1): p. 44-55.
17. Hollins, S. and I. Tuffrey-Wijne, *Improving hospital care for patients with intellectual disabilities.* Br J Hosp Med (Lond), 2014. **75**(6): p. 304-5.
18. Kozaki, K. and J. Inazawa, *Tumor-suppressive microRNA silenced by tumor-specific DNA hypermethylation in cancer cells.* Cancer Sci, 2012. **103**(5): p. 837-45.
19. Dotti, M., et al., *A Rett syndrome MECP2 mutation that causes mental retardation in men.* Neurology, 2002. **58**: p. 226-30.
20. Cheever, A., E. Blackwell, and S. Ceman, *Fragile X protein family member FXR1P is regulated by microRNAs.* Rna, 2010. **16**(8): p. 1530-9.

21. Asim, A., et al., *"Down syndrome: an insight of the disease".* J Biomed Sci, 2015. **22**: p. 41.
22. Franzoni, E., et al., *miR-128 regulates neuronal migration, outgrowth and intrinsic excitability via the intellectual disability gene Phf6.* Elife, 2015. **4**.
23. Chai, M., et al., *Identification of a thymus microRNA-mRNA regulatory network in Down syndrome.* Mol Med Rep, 2019. **20**(3): p. 2063-2072.
24. Lim, J.H., et al., *Integrative analyses of genes and microRNA expressions in human trisomy 21 placentas.* BMC Med Genomics, 2018. **11**(1): p. 46.
25. Siew, W.H., et al., *MicroRNAs and intellectual disability (ID) in Down syndrome, X-linked ID, and Fragile X syndrome.* Front Cell Neurosci, 2013. **7**: p. 41.
26. Zablotskaya, A., et al., *Mapping the landscape of tandem repeat variability by targeted long read single molecule sequencing in familial X-linked intellectual disability.* BMC Med Genomics, 2018. **11**(1): p. 123.
27. Smedley, D., et al., *The BioMart community portal: an innovative alternative to large, centralized data repositories.* Nucleic Acids Res, 2015. **43**(W1): p. W589-98.
28. Kozomara, A., M. Birgaoanu, and S. Griffiths-Jones, *miRBase: from microRNA sequences to function.* Nucleic Acids Res, 2019. **47**(D1): p. D155-d162.
29. Chen, C., et al., *IDGenetics: a comprehensive database for genes and mutations of intellectual disability related disorders.* Neurosci Lett, 2018. **685**: p. 96-101.
30. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
31. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
32. Carithers, L.J. and H.M. Moore, *The Genotype-Tissue Expression (GTEx) Project.* Biopreserv Biobank, 2015. **13**(5): p. 307-8.
33. Kanehisa, M., et al., *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Res, 2017. **45**(D1): p. D353-d361.
34. Bailey, T.L., et al., *MEME SUITE: tools for motif discovery and searching.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W202-8.
35. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Res, 2014. **42**(Database issue): p. D980-5.
36. Liu, X., et al., *dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs.* Hum Mutat, 2016. **37**(3): p. 235-41.
37. Chen, W., et al., *Mutation screening of brain-expressed X-chromosomal miRNA genes in 464 patients with nonsyndromic X-linked mental retardation.* Eur J Hum Genet, 2007. **15**(3): p. 375-8.
38. Karagkouni, D., et al., *DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions.* Nucleic Acids Res, 2018. **46**(D1): p. D239-d245.
39. Guo, S., et al., *Identification and analysis of the human sex-biased genes.* Brief Bioinform, 2018. **19**(2): p. 188-198.
40. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life.* Nucleic Acids Res, 2015. **43**(Database issue): p. D447-52.
41. Curie, A., et al., *Basal ganglia involvement in ARX patients: The reason for ARX patients very specific grasping?* Neuroimage Clin, 2018. **19**: p. 454-465.

42. Gonçalves, T.F., et al., *Network Profiling of Brain-Expressed X-Chromosomal MicroRNA Genes Implicates Shared Key MicroRNAs in Intellectual Disability.* J Mol Neurosci, 2019. **67**(2): p. 295-304.
43. Vakilian, H., et al., *DDX3Y, a Male-Specific Region of Y Chromosome Gene, May Modulate Neuronal Differentiation.* J Proteome Res, 2015. **14**(9): p. 3474-83.
44. Fernandez, B.A. and S.W. Scherer, *Syndromic autism spectrum disorders: moving from a clinically defined to a molecularly defined approach.* Dialogues Clin Neurosci, 2017. **19**(4): p. 353-371.