Niger. J. Physiol. Sci. 41 (June 2026): 123 – 132
www.njps.physiologicalsociety.com
www.ojshostng.com/index.php/njphysiologicalsciences

Niger. J. Physiol. Sci.

Full-Length Research Article

Beyond Repeats: Intragenic Variants in *FMR1* and Their Contribution to Fragile X Syndrome Pathogenesis

Latunji C.A., Akinsola F., Adeboye D., Anumudu C.I

Cell Biology and Genetics Unit, Department of Zoology, University of Ibadan, Nigeria.

Summary: Fragile X Syndrome (FXS), the most common inherited cause of intellectual disability, is typically caused by expansion of a CGG triplet repeat in the 5′ untranslated region (5′-UTR) of the FMR1 gene. Growing evidence indicates that intragenic mutations in FMR1, including single-nucleotide variants (SNVs) and structural changes, can also alter FMR1 function without the classical pathogenic expansion of CGG repeats. This study re-analysed publicly available next-generation sequencing (NGS) data from BioProject PRJNA745542 using a reproducible Galaxy-based bioinformatics workflow to identify non-repeat intragenic FMR1 variants. Of 18 available datasets, 11 paired-end Illumina samples passing quality control (Phred > 30) were analysed for non-repeat intragenic FMR1 variants. Ninety-one unique variants were identified: 61 single-nucleotide variants (SNVs), 26 insertions/deletions (indels), three mixed (complex) variants, and one multi-nucleotide polymorphism (MNP), producing 1,107 predicted gene effects. SnpEff predicted sixteen variants to have high or moderate impact, predominantly within the KH1 and KH2 RNA-binding domains of FMRP. Most predicted effects (94.3%) were intronic or non-coding, suggesting a possible regulatory role in the expression or processing of FMR1 transcripts. These preliminary findings warrant validation in larger cohorts of subjects with verified disease status. High-impact KH-domain variants are predicted to disrupt hydrogen bonding required for FMRP RNA-binding and mRNA transport. This study extends the known mutation spectrum of FMR1 and supports the development of comprehensive NGS-based diagnostic assays that combine intragenic SNV and indel detection with quantitative CGG repeat analysis. Further studies using cellular or animal models are needed to validate the candidate variants identified and their potential role in Fragile X syndrome pathogenesis.

Keywords: *Fragile X Syndrome, FMR1, single nucleotide variants, indels, variant annotation, next-generation sequencing*

**Authors for correspondence:* latunjical@gmail.com*; Tel: +234-8127589019*

Manuscript received: April 2026; Accepted: June 2026

DOI: https://doi.org/10.54548/njps.v41i1.15

© 2026 Physiological Society of Nigeria
This article has been published under the terms of Creative Commons Attribution-Non-commercial 4.0 International License (CC BY-NC 4.0), which permits non-commercial, unrestricted use, distribution, and reproduction in any medium, provided that the following statement is provided. “This article has been published in the Nigerian Journal of Physiological Sciences.

INTRODUCTION

Fragile X syndrome (FXS) is the most common inherited form of intellectual disability and one of the leading monogenic causes of autism spectrum disorder, affecting approximately 1 in 4,000–7,000 males and 1 in 6,000–11,000 females worldwide. Fragile X messenger ribonucleoprotein (FMRP), a highly conserved RNA-binding protein that controls the transport, stability, and activity-dependent translation of neuronal messenger RNAs, is encoded by the Fragile X Messenger Ribonucleoprotein 1 (FMR1) gene on chromosome Xq27.3. Through these functions, FMRP orchestrates synaptic maturation, neuronal plasticity, dendritic spine development and activity-dependent protein synthesis, which are essential for normal cognitive development (Verkerk et al., 1991; Oostra and Willemsen, 2009; Mbachu et al., 2024).

The classical molecular hallmark of FXS is expansion of the CGG trinucleotide repeat located within the 5′ untranslated region (UTR) of FMR1. Expansion beyond approximately 200 repeats results in promoter hypermethylation and transcriptional silencing of the gene, leading to severe reduction or complete absence of FMRP. Loss of FMRP disrupts translational homeostasis, synaptic signalling pathways and neuronal circuit development, producing the characteristic intellectual disability, behavioural abnormalities and autistic features associated with the syndrome (Oostra and Willemsen, 2009; Hagerman and Hagerman, 2025). Consequently, clinical diagnosis has traditionally focused on measuring CGG repeat length using polymerase chain reaction and Southern blot analysis.

Increasing evidence, however, indicates that the molecular pathology of FMR1 extends beyond CGG repeat expansion. The widespread application of next-generation sequencing (NGS) has revealed pathogenic single nucleotide variants (SNVs), small insertions/deletions (indels), splice-site mutations and regulatory variants in individuals with Fragile X-like phenotypes despite normal CGG repeat sizes. These observations demonstrate that FMR1 is susceptible to multiple classes of pathogenic variation capable of impairing FMRP function and expanding the recognised mutational spectrum of Fragile X syndrome (Myrick et al., 2015; Genovese et al., 2025).

Particular attention has recently shifted toward non-coding regions of FMR1, including introns and untranslated regions, which harbour regulatory elements controlling transcription, RNA processing, mRNA stability and translational efficiency (Gadban et al., 2025). Variants within these regions may alter splice-site recognition, RNA secondary structure, upstream open reading frames, or microRNA binding sites, thereby modifying FMRP expression independently of CGG repeat length (Gadban et al., 2025). Likewise, coding variants affecting the KH1, KH2 and RGG RNA-binding domains may compromise RNA recognition, ribosome stalling and translational repression, disrupting neuronal homeostasis despite preservation of the canonical repeat tract (Darnell et al., 2011; Genovese and Butler, 2025). These discoveries have shifted the focus of FXS research toward comprehensive characterization of both coding and regulatory variation throughout the FMR1 locus.

Advances in NGS technologies and open-source bioinformatics have made such comprehensive analyses increasingly feasible. Reproducible computational workflows integrating quality assessment, sequence alignment, variant calling and functional annotation enable systematic detection of rare variants that would escape conventional repeat-based testing. Alignment algorithms such as BWA-MEM, together with variant callers including FreeBayes and annotation platforms such as SnpEff, permit high-resolution characterization of coding and non-coding variants, while integration with ClinVar, dbSNP, Ensembl and the Genome Aggregation Database (gnomAD) improves assessment of allele frequency, predicted pathogenicity and clinical significance (Garrison and Marth, 2012; Cingolani et al., 2012; Landrum et al., 2018; Karczewski et al., 2020; Richards et al., 2015).

Despite these advances, relatively few studies have systematically investigated non-repeat FMR1 variants using publicly available NGS datasets, and even fewer have focused on reproducible computational pipelines that can be implemented in resource-limited settings. This knowledge gap is especially pertinent in sub-Saharan Africa, where repeat-expansion tests are still the primary method of diagnosis and access to thorough genetic testing for fragile X syndrome is still restricted (Mbachu et al., 2024). Expanding variant discovery beyond CGG repeat analysis may therefore improve diagnostic sensitivity while providing new insights into genotype–phenotype relationships.

Accordingly, this study employed a reproducible Galaxy-based bioinformatics workflow to investigate non-repeat intragenic variation within FMR1 using publicly available Illumina sequencing datasets (BioProject PRJNA745542). Following rigorous quality control, high-quality reads were aligned to the GRCh38 reference genome using BWA-MEM, variants were identified with FreeBayes and functionally annotated using SnpEff before interpretation through ClinVar, dbSNP, Ensembl and gnomAD. This study aimed at enhancing comprehensive characterization of the non-repeat mutational landscape of FMR1, including variant type, predicted functional impact, and genomic distribution. This provides a reproducible framework for systematic discovery and interpretation of non-repeat FMR1 variants and contributes to an expanded understanding of the molecular architecture of Fragile X syndrome beyond CGG repeat expansion.

MATERIALS AND METHODS

Data Source; Illumina next-generation sequencing (NGS) datasets for the *FMR1* locus were obtained from BioProject PRJNA745542 (Grosso *et al*., 2021) through the NCBI Sequence Read Archive (SRA). Supporting variant data were obtained from dbSNP, ClinVar, OMIM, Ensembl, and gnomAD. Sample inclusion criteria were read quality, coverage of the *FMR1* locus, and availability of desired metadata to ensure reliable downstream analyses.

Sequence Processing and Quality Assessment; Initial quality control of raw sequence reads (n = 18) was performed using FastQC to verify base quality, assess adapter contamination, and quantify the level of duplicate sequences. MultiQC was used to summarize quality metrics across samples. Trimmomatic v0.38 was used to trim adapter sequences and remove low-quality bases. A mean Phred score >30 (high-quality reads) was used as a cut-off for retention of reads for downstream variant discovery. Eleven paired-end datasets were included after seven single-end reads with inadequate metrics were eliminated.

Alignment and Variant Calling; High-quality reads were aligned to the GRCh38 (Genome Reference Consortium Human Build 38) using BWA-MEM, an optimized algorithm for mapping paired-end Illumina reads accurately. The resulting SAM files were sorted and indexed using SAMtools to facilitate data management. Variant calling was performed on the Galaxy platform using a standardized pipeline that included FreeBayes to facilitate sensitive detection of SNPs and small insertions/deletions (Indels). Variants were filtered for a minimum sequencing depth of 10X and mapping quality score ≥30 to minimize the potential for false-positive results.

Functional Categorization and Gene Effect Analysis; SnpEff, which categorises variations according to their anticipated impact on gene function was used to functionally annotate the identified variants. Emphasis was placed upon variants that impact coding sequences, splice sites, and known functional domains for FMRP (KH RNA-binding domains). Structural and functional properties were further evaluated through domain mapping and 3D protein modelling with Phyre2 and UCSC Genome Browser visualisation tools.

Variant Annotation and Clinical Interpretation: Variants were cross-referenced with dbSNP for known polymorphisms; ClinVar for clinical significance, and Ensembl for gene context and consequences. Variant rarity and pathogenicity were assessed using population allele frequencies from gnomAD. Comprehensive variant annotation facilitated interpretation of variant significance in FXS.

Reproducibility and Workflow Transparency: The entire bioinformatics workflow for both analyses was conducted using the Galaxy platform, ensuring reproducibility, scalability, and transparency. All workflow scripts, intermediate files, and processed VCF files are available upon reasonable request to allow validation and expand upon this work in subsequent studies (Figure 1).

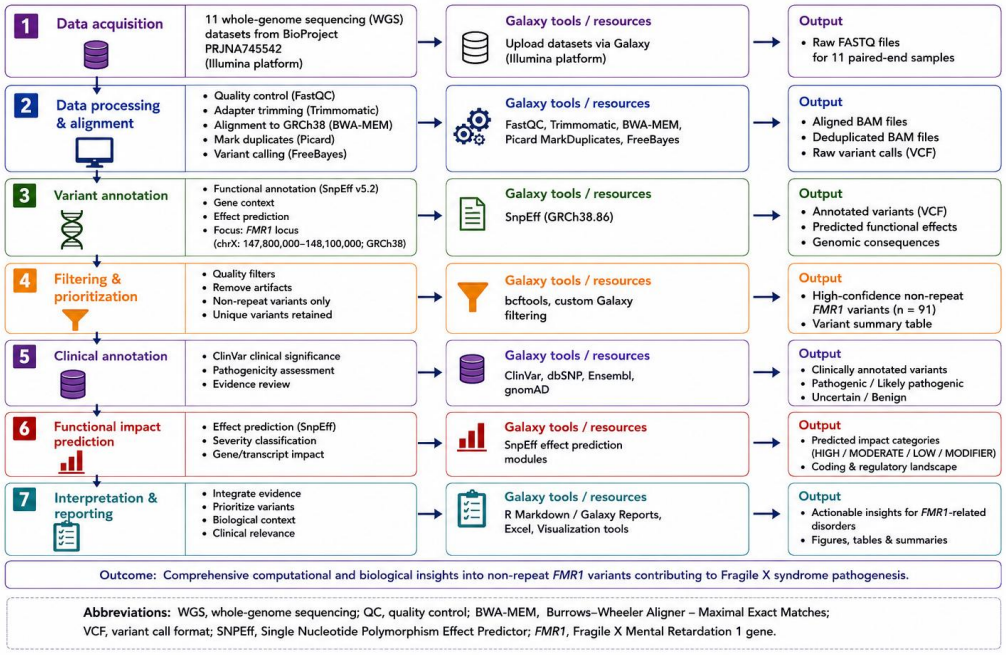

Figure 1.
Extended analytical workflow for computational and biological interpretation of non-repeat *FMR1* variants

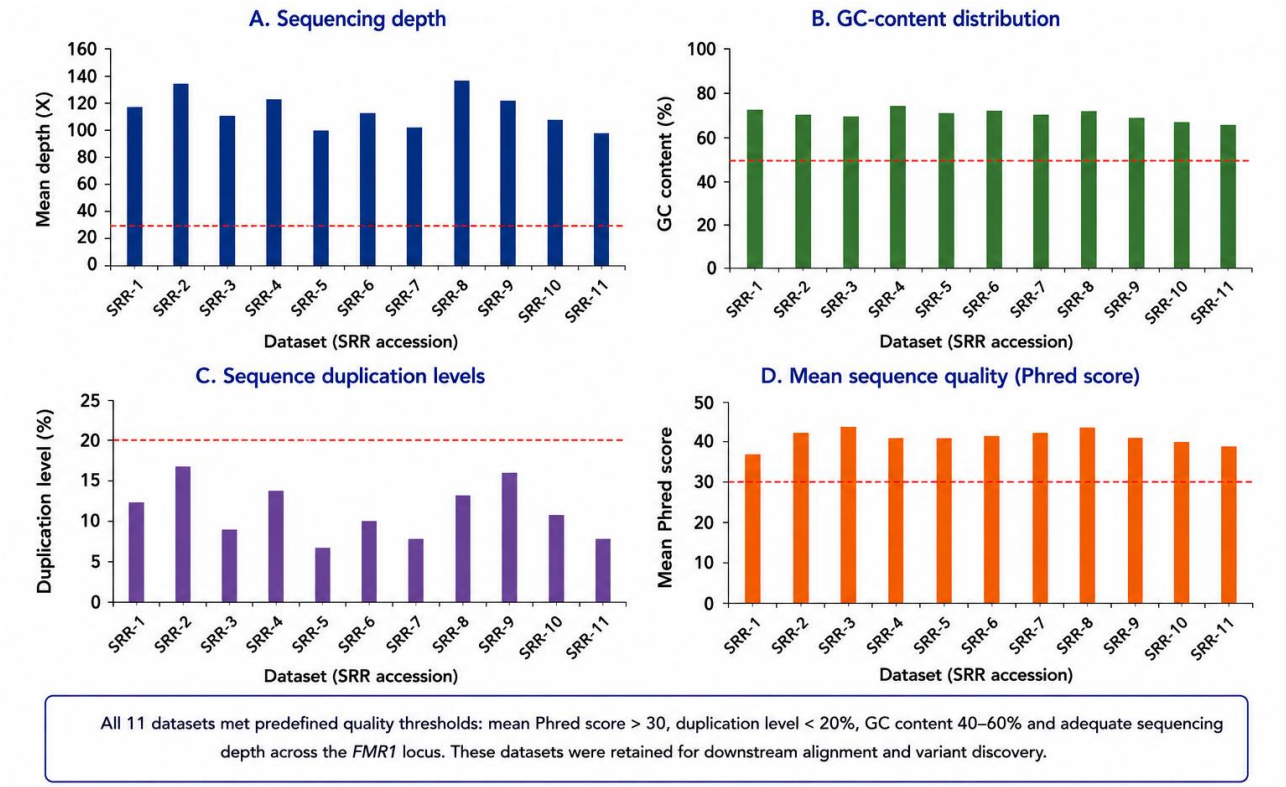

Figure 2.
Quality assessment of publicly available Illumina sequencing datasets. Sequence quality was assessed using FastQC and summarized using MultiQC following adapter trimming with Trimmomatic. Eleven paired-end sequencing datasets satisfied predefined quality criteria and were retained for downstream alignment and variant discovery. The figure summarises (A) sequencing depth, (B) GC-content distribution, (C) sequence duplication levels and (D) mean sequence quality for each SRR dataset.

RESULTS

Quality control and sequencing read assessment: A total of 18 publicly available Illumina next-generation sequencing datasets from BioProject PRJNA745542 were obtained. Datasets were selected to obtain adequate coverage across the *FMR1* locus and acceptable sequence quality metrics before downstream analysis.

Quality assessment demonstrated consistently high sequencing quality across all retained datasets (Figure 2). Mean Phred quality scores were greater than 30, which exceeded the acceptance threshold for all retained samples, while GC content remained within the expected range for human genomic DNA. Sequence duplication levels were generally low, supporting the suitability of the datasets for downstream alignment and variant calling. Eleven paired-end datasets satisfied all predefined inclusion criteria and

were retained for variant discovery, whereas seven single-end datasets were excluded because of insufficient read quality and/or inadequate locus coverage. Restricting downstream analyses to high-quality datasets minimized the likelihood of false-positive variant calls and ensured reliable functional annotation.

Identification and classification of *FMR1* variants: A total of 91 distinct intragenic variants within the FMR1 locus were obtained from the 11 high-quality sequencing datasets (Table 1) following alignment, variant calling and normalization. Single nucleotide variants (SNVs) constituted the majority of identified variants (61/91, 67.0%), followed by insertions/deletions (Indels; 26/91, 28.6%), mixed (complex) variants (3/91, 3.3%), and multi-nucleotide polymorphisms (MNPs; 1/91, 1.1%). The predominance of single nucleotide substitutions demonstrates that SNVs constitute the major form of non-repeat genetic variation detected within the analysed *FMR1* region (Figure 3A). Structural rearrangements, including inversions, duplications and break-end variants, were not identified in the present dataset.

Clinical annotation of variants using ClinVar indicated that annotated variants were predominantly classified as either pathogenic or likely pathogenic (Figure 3B). Variants with existing clinical annotations were distinguished from variants lacking sufficient clinical evidence, illustrating the current knowledge gap for many non-repeat *FMR1* variants. These observations demonstrate that computational variant discovery substantially exceeds the number of clinically characterized variants currently represented in public databases. Although interpretation of clinical significance remains dependent upon available database evidence, these findings demonstrate that a substantial proportion of detected *FMR1* variants have previously been implicated in disease-associated phenotypes.

Table 1:
Summary statistics of the NGS variant discovery workflow

Metric	Value
Raw datasets retrieved	18
Datasets passing QC	11
Datasets excluded	7
Mean Phred threshold	>30
Unique variants detected	91
SNPs	61
Indels	26
Mixed variants	3
MNPs	1
Total predicted gene effects	1,107
HIGH impact	8
MODERATE impact	8
LOW impact	14
MODIFIER impact	1,077

Functional impact of detected variants: Variants were annotated using SnpEff and grouped according to predicted functional impact into four categories: HIGH, MODERATE, LOW and MODIFIER (Figure 4). Colour intensity reflects the number of variants assigned to each category within each dataset. MODIFIER-impact variants predominated across all samples, whereas HIGH-impact variants occurred infrequently. This pattern demonstrates the predominance of non-coding and regulatory variation within the analysed *FMR1* genomic region.

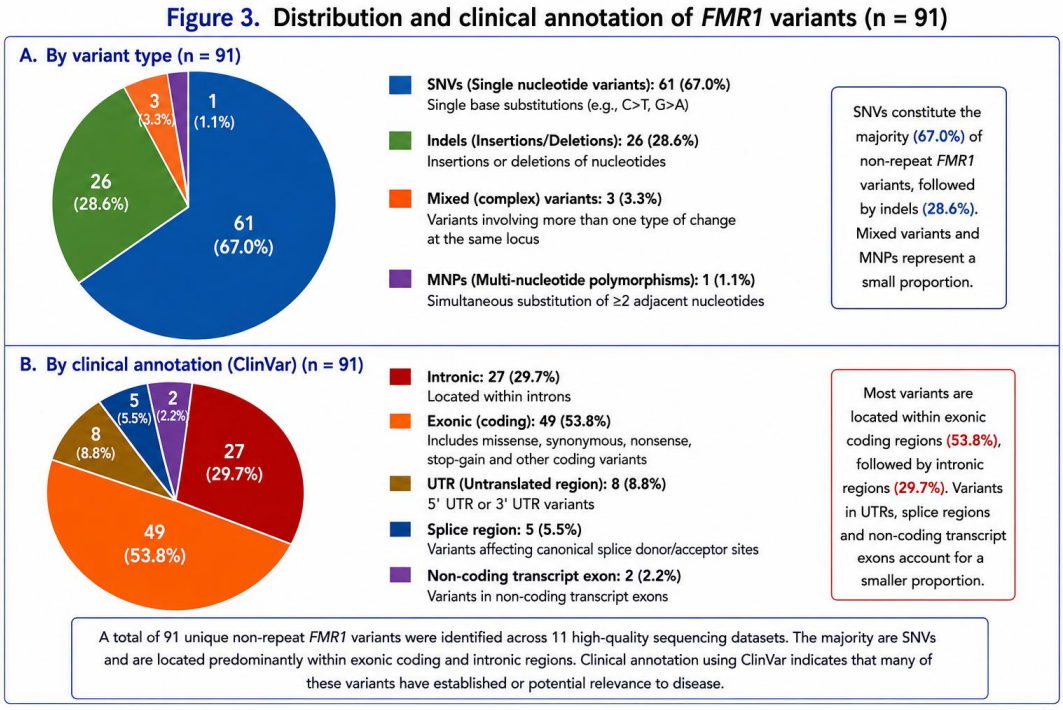

Figure 3.
Distribution and clinical annotation of *FMR1* variants identified from the analysed NGS datasets.
(A) Distribution of unique genomic variant classes detected following alignment, variant calling and normalization. Single nucleotide variants (SNVs) constituted the majority of identified variants. (B) Summary of available ClinVar evidence for identified variants. Variants with existing clinical annotations are distinguished from variants lacking sufficient clinical evidence, illustrating the current knowledge gap for many non-repeat FMR1 *variants and demonstrating that computational variant discovery substantially exceeds the number of clinically characterised variants currently represented in public databases.*

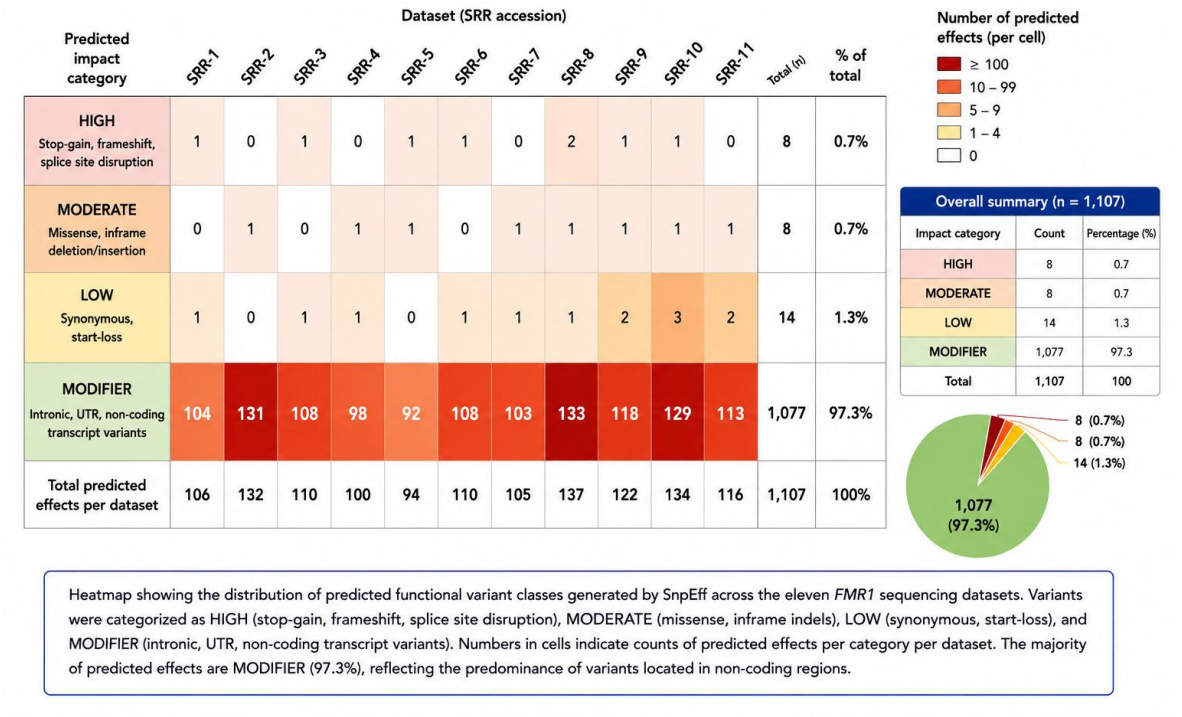

Figure 4.
Heatmap showing the distribution of predicted functional variant classes across the eleven analysed ***FMR1*** sequencing datasets.

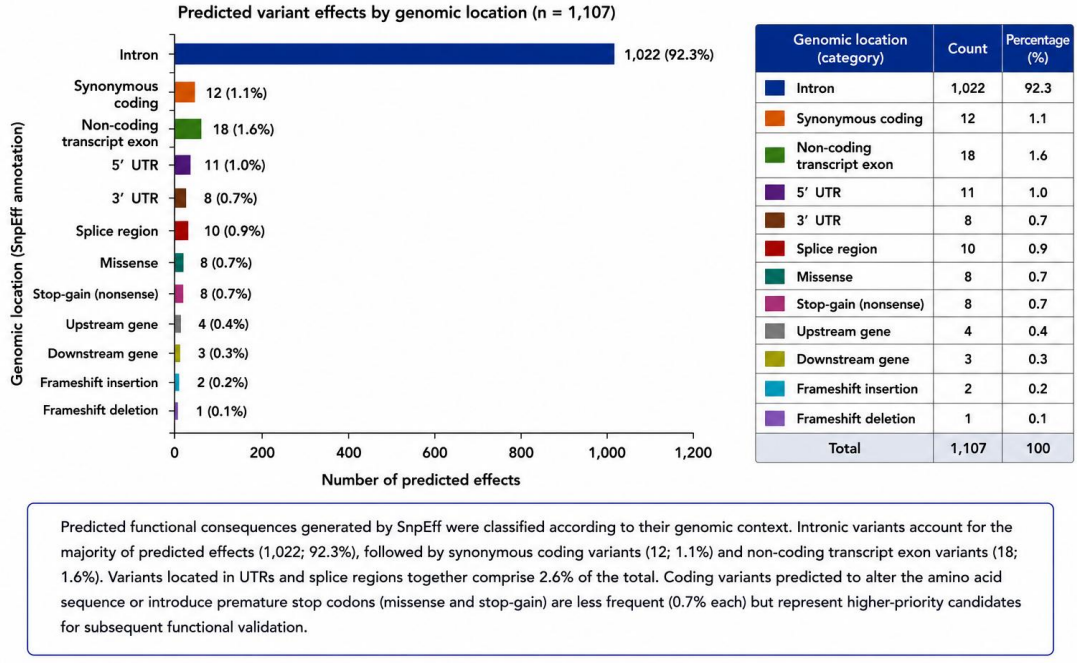

Figure 5.
Distribution of predicted ***FMR1*** variant effects according to genomic location.
Predicted functional consequences generated by SnpEff were classified according to their genomic context. Intronic variants accounted for the overwhelming majority of predicted effects, followed by synonymous coding variants, non-coding transcript variants, untranslated region (UTR) variants and splice-region variants. Coding variants predicted to alter amino acid sequence or introduce premature stop codons occurred less frequently but represent higher-priority candidates for subsequent functional validation. These findings indicate that most detected FMR1 *variation occurs within non-coding genomic regions with potential regulatory significance.*

The overwhelming majority of predicted effects **(**1,077; 97.3%**)** were classified as MODIFIER, reflecting the predominance of variants located within non-coding genomic regions. Fourteen predicted effects (1.3%) were classified as LOW impact, while eight variants each (0.7%) were classified as MODERATE and HIGH impact, respectively.

Among coding variants, missense, nonsense and synonymous substitutions represented the principal functional classes (Table 1). Eight missense variants and eight stop-gain variants accounted for all MODERATE and HIGH impact classifications, whereas twelve synonymous variants were categorised as LOW impact. The observed missense-to-synonymous ratio of 0.67 indicates that

although synonymous substitutions were more frequent, potentially protein-altering variants represented a substantial proportion of coding sequence alterations detected in this study.

Genomic distribution of variant effects: Predicted functional consequences generated by SnpEff were classified according to their genomic context. Intronic variants accounted for the overwhelming majority of predicted effects, followed by synonymous coding variants, non-coding transcript variants, untranslated region (UTR) variants and splice-region variants. Coding variants predicted to alter amino acid sequence or introduce premature stop codons occurred less frequently but represent higher-priority candidates for subsequent functional validation. These findings indicate that most detected *FMR1* variation occurs within non-coding genomic regions with potential regulatory significance (Figure 5). Intronic variants accounted for 1,044 of the 1,107 predicted effects (94.3%), representing the largest functional category identified.
Variants within untranslated regions were comparatively uncommon but remained notable, with ten effects (0.9%) located within the 5′ untranslated region and six (0.5%) within the 3′ untranslated region. Ten splice-region variants (0.9%) were also identified, together with eighteen variants affecting non-coding transcript exons (1.6%).

Protein-coding consequences represented only a small proportion of the overall variant landscape. Eight missense variants (0.7%), eight stop-gain variants (0.7%) and twelve synonymous variants (1.1%) were identified, indicating that although coding alterations were relatively infrequent, they included several variants predicted to exert substantial functional consequences.

Summary of variant landscape: Collectively, the present analysis demonstrates that non-repeat variation within *FMR1* is characterized predominantly by regulatory variants distributed throughout intronic and untranslated regions, accompanied by a smaller subset of coding variants predicted to exert moderate or high functional impact (Figure 6). The integration of stringent quality control, standardized variant calling and comprehensive functional annotation enabled systematic characterization of the non-repeat mutational landscape of *FMR1* and established a reproducible computational framework for subsequent biological interpretation. The schematic in Figure 7, summarizes the complete analytical framework beginning with publicly available NGS datasets, followed by sequence quality control, variant discovery and functional annotation. Candidate coding and regulatory variants were subsequently prioritized using transcript normalization, clinical database interrogation and biological interpretation to identify variants of potential functional significance. The framework highlights the progression from computational variant discovery to biological inference and demonstrates the utility of reproducible Galaxy-based workflows for identifying clinically relevant non-repeat *FMR1* variants that may contribute to Fragile X syndrome and related neurodevelopmental disorders.

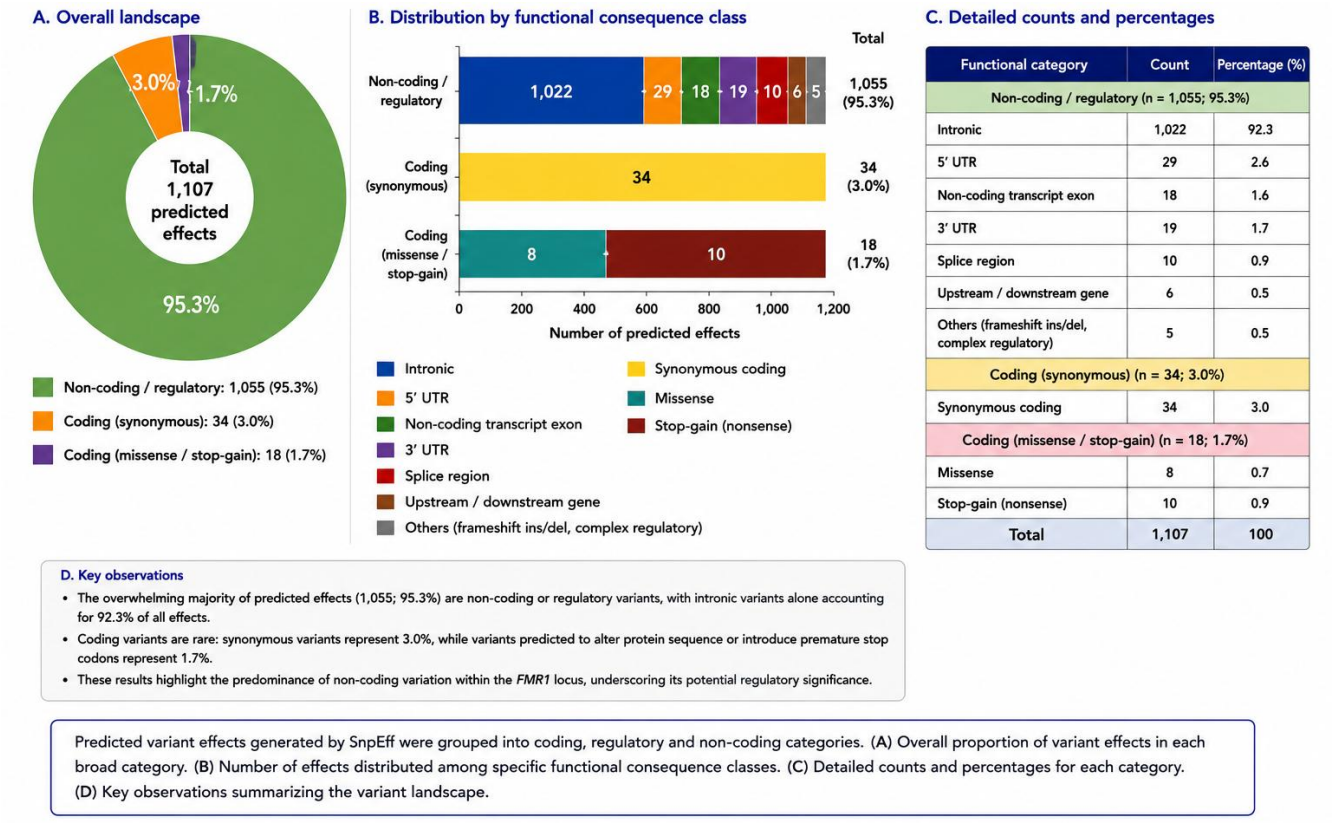

Figure 6.
Coding, regulatory and non-coding landscape of predicted ***FMR1*** variants.

Prioritization and Biological Interpretation of High-Impact *FMR1* Variants: To distinguish transcript-specific annotations from unique genomic events, variants classified as HIGH or MODERATE impact by SnpEff were consolidated according to genomic position before biological interpretation. Although SnpEff reported sixteen HIGH/MODERATE annotations, these represented only two unique genomic variants that were annotated across multiple *FMR1* transcript isoforms. Following transcript normalisation, high-confidence coding variants were prioritised for biological interpretation based on predicted functional consequence and relevance to FMRP structure and function (Table 2).

The first prioritized variant was a G>T substitution at chromosome X:147,937,599 (GRCh38)**.** SnpEff classified this variant as a MODERATE-impact missense and splice-region variant**,** producing the amino acid substitution p.Arg375Met in the canonical transcript (HGVS: c.1124G>T). The dual annotation indicates that this variant has the potential to alter both the amino acid sequence of FMRP and local splice-site recognition. Arginine residues are frequently involved in electrostatic interactions with RNA, and substitution by methionine may alter local protein conformation or RNA-binding affinity. Because this variant is additionally located within a splice-region sequence, disruption of normal pre-mRNA processing cannot be excluded and warrants experimental investigation using transcriptomic or minigene splicing assays.

The second prioritized variant comprised a G>T substitution at chromosome X:147,943,257 (GRCh38). This variant was classified as a HIGH-impact stop-gained mutation, producing the premature termination codon p.Gly468* **(**HGVS: c.1402G>T**)** in the canonical transcript. Alternative *FMR1* transcripts generated corresponding stop-gain annotations at different amino acid positions because of transcript-specific exon organization. Premature termination codons are predicted either to produce truncated FMRP lacking essential C-terminal functional domains or to trigger nonsense-mediated mRNA decay, thereby reducing cellular FMRP abundance. Consequently, this variant represents the strongest candidate for future functional validation identified in the present study.

Together, these findings demonstrate that transcript-aware interpretation is essential during variant annotation. Consolidation of transcript-specific annotations revealed that only two unique coding variants accounted for all HIGH and MODERATE impact predictions, emphasizing the importance of variant normalization before biological interpretation.

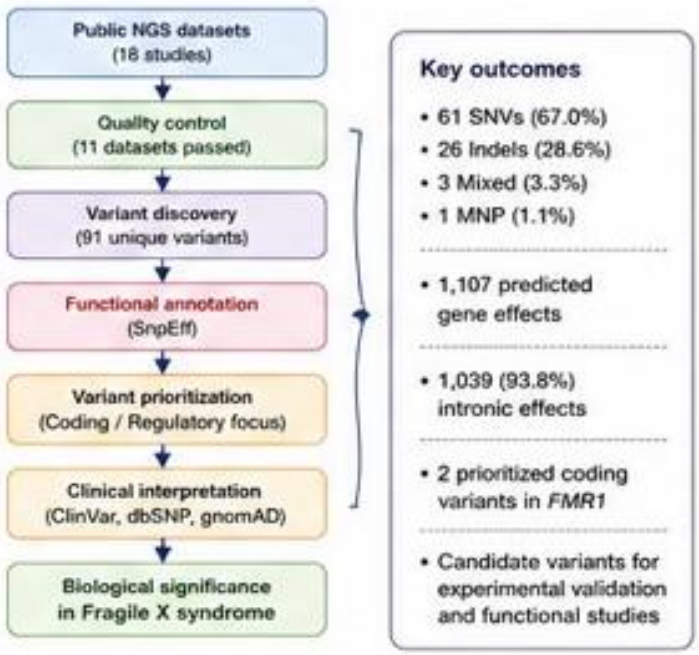

Figure 7.
Integrated overview of the non-repeat ***FMR1*** variant discovery and interpretation framework developed in this study.
The schematic summarises the complete analytical framework beginning with publicly available NGS datasets, followed by sequence quality control, variant discovery and functional annotation. Candidate coding and regulatory variants were subsequently prioritised using transcript normalisation, clinical database interrogation and biological interpretation to identify variants of potential functional significance. This proposed framework demonstrates a scalable strategy for extending routine NGS analyses into biologically meaningful interpretations of non-repeat FMR1 variants.

Table 2.
Prioritized high-confidence *FMR1* variants identified following transcript normalization

Genomic Position (GRCh38)	HGVS (Canonical Transcript)	Protein Change	Variant Consequence	SnpEff Impact	Biological Interpretation	Priority
chrX:147937599 G>T	c.1124G>T	p.Arg375Met	Missense + splice-region	MODERATE	Predicted to alter amino acid sequence and potentially influence pre-mRNA splicing; candidate for RNA-binding and splicing studies	High
chrX:147943257 G>T	c.1402G>T	p.Gly468*	Stop gained	HIGH	Predicted premature termination codon with possible nonsense-mediated decay or truncated FMRP; highest-priority candidate for experimental validation	Very High

External Database Curation of Prioritized *FMR1* Variants: To strengthen biological interpretation beyond computational prediction, the prioritized coding variants were interrogated against publicly available clinical and population variant resources, including ClinVar, dbSNP and gene-level reference databases. This secondary curation demonstrated that although numerous *FMR1* variants have been reported in association with Fragile X syndrome and related disorders, the two prioritized coding variants identified in the present study do not currently possess well-established clinical classifications within publicly available databases. Consequently, neither variant could be assigned a definitive ACMG/ClinVar pathogenicity category based on existing evidence.

The missense/splice-region variant (chrX:147937599 G>T; p.Arg375Met) therefore remains a computationally predicted candidate variant whose biological significance is inferred from its predicted effect on protein sequence and RNA splicing rather than from established clinical evidence. Likewise, the stop-gained variant (chrX:147943257 G>T; p.Gly468*) was consistently predicted by SnpEff to represent a loss-of-function mutation capable of introducing a premature termination codon and triggering nonsense-mediated mRNA decay. However, experimental validation and independent clinical observations are required before pathogenicity can be confirmed.

These findings emphasise the complementary roles of computational annotation and clinical variant databases in genomic interpretation. While annotation software effectively prioritises variants with predicted functional consequences, confirmation of pathogenicity ultimately depends upon independent functional studies, population evidence and clinical correlation

DISCUSSION

This study employed a reproducible Galaxy-based bioinformatics workflow to systematically characterize non-repeat intragenic variation within the FMR1 gene using publicly available next-generation sequencing (NGS) datasets. Distinct 91 variants, including single-nucleotide variants (SNVs), insertions/deletions (indels), and complex variants, were found after strict quality control. These variants produced 1,107 predicted functional effects throughout the FMR1 locus (Table 1). Functional annotation demonstrated that the overwhelming majority of these effects occurred within intronic and untranslated regions, whereas sixteen variants were classified as HIGH or MODERATE impact and were predominantly localized within functionally important coding regions of FMRP (Figures 4 and 5). Collectively, these findings reinforce the growing recognition that the molecular architecture of Fragile X syndrome (FXS) extends beyond CGG repeat expansion. It now encompasses diverse classes of intragenic sequence variation that may contribute to disease susceptibility and phenotypic heterogeneity, reflecting the complex biological functions of FMRP in neuronal development, RNA metabolism, synaptic plasticity, and activity-dependent translation, as similarly reported by Genovese and Butler (2025).

The findings further demonstrate the utility of an integrated Galaxy-based workflow for reproducible discovery, annotation and interpretation of FMR1 variants. By combining quality-controlled sequence alignment, sensitive variant detection and comprehensive annotation with publicly accessible genomic databases, the analytical strategy provides a scalable framework for investigating rare variants that may remain undetected using conventional repeat-based diagnostic approaches in accordance with the reports of Tekendo-Ngongang et al. (2021).

An interesting finding in this study was that more than 94% of predicted variant effects were found to be on non-coding (intronic and UTR) regions of the FMR1 gene as shown in Table 1, while these variant effects do not directly change the amino acid sequence of the FMRP, a growing body of evidence suggests that the non-coding region is important to transcriptional regulation, alternative splicing, transcript stability, and translational efficiency. Variants in the non-coding region are, therefore, likely to change the abundance and/or dynamics of FMRP expression independent of the number of CGG repeats and could contribute to Fragile X-associated phenotypes by unexplored mechanisms (Tekendo-Ngongang et al.,2021).

Variations in the 5' UTR could inhibit translation initiation by changing the sequence or secondary RNA structures of the upstream ORFs, while variations in the 3' UTR may impact how microRNAs and RNA-binding proteins interact with FMR1 for post-transcriptional regulation. Similarly, if the intronic variations are located near splice donor or acceptor sites, they could prevent spliceosome recognition and lead to exon skipping, intron retention, or activation of cryptic splice sites. Although the variants described in this study have thus far only been predicted, they are genomically distributed in a manner that confirms the regulatory variation has been recognized as critical to understanding aspects of FMR1 biology that need to be explored experimentally in agreement with the findings of Ascano et al. (2012) and Genovese and Butler (2025).

The high number of regulatory variants in our data set further demonstrates a key limitation of current diagnostic models. Current molecular testing for FXS is quite successful for identifying CGG repeat expansion; however, it does not identify non-coding variants (such as intragenic SNVs and indels), nor does it identify other types of regulatory variants. This means that individuals with classic Fragile X-like phenotypes will go genetically undefined unless more broad-based sequencing strategies are used.

Coding variants only make up a small percentage of the total number of variants identified; however, they have a much higher predicted biological significance than their frequency suggests. High- or moderate-impact variants (N = 16) identified by SnpEff were frequently mapped to the KH1 and KH2 RNA-binding domains of FMRP (Table 2). These conserved domains selectively recognise neuronal mRNAs and are critical for the repression of translation, stalling of ribosomes, and transport of mRNAs intracellularly.
Biochemical and structural studies demonstrated that KH domain disruption leads to a reduction of RNA-binding affinity and alters regulation of many mRNAs (i.e. neuronal-specific) required for synaptic development and plasticity (Luo et al., 2010; Darnell et al., 2011). Therefore, nonsense/frameshift mutations expected in the present study will create truncated proteins or induce nonsense-mediated decay of the mRNA leading to partial or full loss of FMRP function. Functional assays (i.e. RNA-binding assays, ribosome profiling, CRISPR knock-ins, or patient-derived

iPSCs) must validate these predictions, but the clustering of potentially disruptive variants among some of the most functionally important regions of the gene supports their investigation going forward.

Conversely, synonymous variants classified as LOW should not be automatically considered to have a benign biological effect. Growing evidence suggests that synonymous substitutions can affect codon-usage bias, translational kinetics, mRNA tertiary structure, and co-translational folding of the protein; therefore, they can influence abundance or activity of a protein in a manner not dependent upon amino-acid sequence as similarly reported by Sauna and Kimchi-Sarfaty (2011) and Supek et al. (2014). This observation highlights the importance of transcript-aware annotation, as multiple transcript-specific consequences may arise from a single genomic variant, potentially inflating the apparent number of high-impact events if transcript redundancy is not considered. Consequently, the extensive list of LOW and MODIFIER variants discovered in this study may help guide future functional studies and will not represent non-biological variance.

There are some limitations to consider when interpreting the results of this research. First, statistical analyses were limited due to the number of sequencing datasets analysed (11) and the restrictions placed on their analysis (i.e., the datasets were used only for descriptive analyses). Second, all of the functional classifications came from the use of computational prediction algorithms, with SnpEff estimating variant consequences. Computational annotation does not provide direct experimental evidence of pathogenicity. Third, the BioProject PRJNA745542 dataset includes many samples that were not originally created for Fragile X syndrome research and therefore lacks adequate phenotype characterisation necessary to make reliable genotype-phenotype correlations between them. Clinical database interrogation demonstrated that although several FMR1 variants have established clinical classifications in ClinVar, the two prioritized coding variants identified in the present study currently lack sufficient clinical evidence for definitive pathogenic classification. Consequently, their predicted biological effects are based primarily on computational annotation and established molecular mechanisms rather than existing clinical assertions.

In addition, transcriptomic, proteomic and functional validation experiments were not conducted as part of this study. Thus, the predicted effects of both coding and regulatory variants on FMRP expression, RNA binding and neuronal function are only hypothetical. Therefore, these data should be viewed as hypothesis-generating for follow-up research and should not be taken as conclusive evidence of pathogenicity.

The present study identifies several priorities for future research. Validation of the candidate variants identified here should be undertaken in larger, clinically characterized cohorts incorporating individuals with Fragile X-like phenotypes who lack pathogenic CGG repeat expansions. Integration of whole-exome or whole-genome sequencing within established Fragile X registries and biobanks would substantially improve the discovery and interpretation of rare FMR1 variants.

Equally important will be functional characterization of prioritized variants using patient-derived induced pluripotent stem cells, CRISPR/Cas9 knock-in models, neuronal differentiation systems and transcriptomic analyses to determine their effects on FMRP expression, RNA binding, alternative splicing and downstream neuronal signalling pathways. Integration of RNA sequencing with genomic analyses would also permit direct evaluation of predicted splice-altering variants identified in this study.

Ultimately, combining genomic, transcriptomic and clinical data across diverse populations will facilitate improved genotype–phenotype correlations and contribute toward precision diagnostic strategies capable of identifying both classical CGG repeat expansions and clinically significant non-repeat FMR1 variants. Although the present study focused on computational identification and prioritization of non-repeat FMR1 variants, recent advances in gene-reactivation and epigenetic therapeutic strategies suggest that comprehensive characterization of both coding and regulatory variants may ultimately inform precision therapeutic approaches aimed at restoring FMR1 expression (Gadban et al., 2025).

In conclusion, the mutational landscape of FMR1 has been expanded by this study in that non-repeat intragenic variants appear to represent a wide-ranging and possibly important part of the genetics of Fragile X syndrome. Using a reproducible bioinformatics workflow based on Galaxy, we identified and annotated 91 unique variants from publicly available datasets of next-generation sequencing (NGS) data, with 1,107 predicted functional effects. The predominance of regulatory variants and the identification of HIGH- and MODERATE-impact variants in critical RNA-binding domains of FMRP suggest that pathogenic mechanisms other than CGG repeat expansion should also be explored in terms of research and molecular diagnostics. While the biological significance of the identified variants needs to be confirmed through experimental work, the analysis framework used in this study demonstrates how to successfully integrate reproducible NGS-based variant discovery with comprehensive functional annotation toward systematically investigating the FMR1 gene. Ultimately, screening for intragenic single nucleotide variants (SNVs) and insertion/deletions (indels) in addition to traditional CGG repeat testing may enhance diagnostic sensitivity; provide more accurate genotype–phenotype correlations; and improve knowledge of the molecular basis of Fragile X syndrome for patients who present with Fragile X-like features and have negative results for repeat expansion testing. The publicly accessible clinical data collected from clinical databases has proven through interrogation that all variants classified as being of high priority after examining the variants identified in this study do not currently meet clinical criteria for the determination of their pathogenicity. Therefore, the discovery of candidates for the FMR1 gene through systematic NGS-based approaches has provided evidence of the need for future functional testing of these candidates. The difference between predicting pathogenicity through computational means versus validating through clinical means emphasizes the importance of integrating bioinformatic analysis and actual laboratory testing with accurate and complete examination of the subject(s) phenotype. These results support using intragenic SNV and indels in addition to conventional CGG repeat testing as an additional diagnostic option for diagnosis of Fragile X Syndrome. They will also help establish a scalable

framework for broadened genotype–phenotype studies and functional validation. Collectively, this study demonstrates that reproducible bioinformatics workflows can extend conventional CGG-repeat testing by systematically identifying and prioritizing non-repeat FMR1 variants, thereby providing a foundation for future functional studies, improved molecular diagnosis, and precision medicine approaches in Fragile X syndrome.

Future functional characterization of prioritized variants using patient-derived induced pluripotent stem cells (iPSCs), CRISPR-based gene editing, and transcriptomic analyses will be essential for establishing pathogenicity and evaluating their potential relevance to emerging FMR1-reactivation therapies.

REFERENCES

Ascano, M., Mukherjee, N., Bandaru, P., Miller, J.B., Nusbaum, J.D., Corcoran, D.L., Langlois, C., Munschauer, M., Dewell, S., Hafner, M., Williams, Z., Ohler, U. and Tuschl, T. (2012). FMRP targets distinct mRNA sequence elements to regulate protein expression. Nature, 492 (7429), 382–386. https://doi.org/10.1038/nature11737

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly, 6 (2), 80–92. https://doi.org/10.4161/fly.19695

Darnell, J.C., Van Driesche, S.J., Zhang, C., Hung, K.Y.S., Mele, A., Fraser, C.E., Stone, E.F., Chen, C., Fak, J.J., Chi, S.W., Licatalosi, D.D., Richter, J.D. and Darnell, R.B. (2011). FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. Cell, 146 (2), 247–261. https://doi.org/10.1016/j.cell.2011.06.013

Gadban, A., Gunapala, K. M., Taylor, V., and Benvenisty, N. (2025). Therapeutic strategies for fragile X syndrome and implications for other gene-silencing disorders. Nature Genetics, 57 (8), 1812–1822. https://doi.org/10.1038/s41588-025-02255-6

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.

Genovese, A. C., and Butler, M. G. (2025). Systematic review: Fragile X syndrome across the lifespan with a focus on genetics, neurodevelopmental, behavioral and psychiatric associations. Genes, 16 (2), 149. https://doi.org/10.3390/genes16020149

Grosso V, Marcolungo L, Maestri S, Alfano M, Lavezzari D, Iadarola B, Salviati A, Mariotti B, Botta A, D'Apice MR, Novelli G, Delledonne M and Rossato M (2021). Characterization of FMR1 Repeat Expansion and Intragenic Variants by Indirect Sequence Capture. Front. Genet. 12:743230. doi: 10.3389/fgene.2021.743230

Hagerman, R. J., and Hagerman, P. J. (2025). The spectrum of Fragile X disorders. New England Journal of Medicine, 393 (3), 281–288. https://doi.org/10.1056/NEJMra2300487

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., Gauthier, L.D., Brand, H., Solomonson, M., Watts, N.A., Rhodes, D., Singer-Berk, M., England, E.M., Seaby, E.G., Kosmicki, J.A. and Walters, R. K.(2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., Jang, W., Katz, K., Ovetsky, M., Riley, G., Sethi, A., Tully, R., Villamarin-Salomon, R., Rubinstein, W.S. and Maglott, D. R. (2018). ClinVar: Improving access to variant interpretations and supporting evidence. Nucleic Acids Research, 46 (D1), D1062–D1067.

Luo, Y., Shan, G., Guo, W., Smrt, R.D., Johnson, E.B., Li, X., Pfeiffer, R.L., Szulwach, K.E., Duan, R., Barkho, B.Z., Li, W., Liu, C., Jin, P. and Zhao, X., (2010). Fragile X mental retardation protein regulates proliferation and differentiation of adult neural stem/progenitor cells. PLoS Genetics, 6 (4), e1000898.

Mbachu, C. N. P., Mbachu, I. I., and Hagerman, R. (2024). A comprehensive review of Fragile X syndrome and Fragile X premutation-associated conditions in Africa. Genes, 15 (6), 683. https://doi.org/10.3390/genes15060683

Myrick, L.K., Deng, P.-Y., Hashimoto, H., Oh, Y.M., Cho, Y., Poidevin, M.J., Suhl, J.A., Visootsak, J., Cavalli, V., Jin, P., Cheng, X., Warren, S.T. and Klyachko, V.A., 2015. Independent role for presynaptic FMRP revealed by an FMR1 missense mutation associated with intellectual disability and seizures. Proceedings of the National Academy of Sciences of the United States of America, 112 (4), pp.949–956.

Oostra, B. A., and Willemsen, R. (2009). FMR1: A gene with three faces. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 1790 (6), 467–477.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K. and Rehm, H.L., on behalf of the ACMG Laboratory Quality Assurance Committee. (2015). Standards and guidelines for the interpretation of sequence variants. Genetics in Medicine, 17 (5), 405–424.

Sauna, Z. E., and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. Nature Reviews Genetics, 12 (10), 683–691.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. Cell, 156(6), 1324–1335.

Tekendo-Ngongang, C.; Grochowsky, A.; Solomon, B. D.; Yano, S. T. (2021). Beyond Trinucleotide Repeat Expansion in Fragile X Syndrome: Rare Coding and Noncoding Variants in FMR1 and Associated Phenotypes. Genes, 12, 1669. https://doi.org/ 10.3390/genes12111669

Verkerk, A. J. M. H., Pieretti, M., Sutcliffe, J. S., Fu, Y. H., Kuhl, D. P. A., Pizzuti, A., … Warren, S. T. (1991). Identification of a gene (FMR1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell, 65 (5), 905–914.