1 Research Question

Using ClinVar and UCSC Genome Browser as tools, how are pathogenic mutations distributed across the *DMD gene* based on genomic coordinates, and how do their frequencies relate to their genomic locations (exonic vs. intronic) and mutation types (single nucleotide variants, deletions, duplications, and insertions)?

2 Background Information

Duchenne Muscular Dystrophy (DMD) is an X-linked recessive genetic neuromuscular disorder caused by mutations in the *DMD* gene. The *DMD* gene encodes a large muscle protein dystrophin, which is essential for maintaining the structural integrity of muscle fibers during contraction and relaxation (*OMIM*). DMD primarily affects males: among every 3,500 to 5,000 newborns, there is one affected individual (*Johns Hopkins Medicine*). Boys with DMD are usually late walkers and prone to falling; most die in their 20s due to respiratory muscle weakness or cardiomyopathy. (*Muscular Dystrophy Association*). The *DMD gene* is the largest known human gene, consisting of 2.4 million base pairs, 79 exons, and encoding 3,685 amino acids (Kumar et al.). As shown in *Figure 1*, the DMD gene is located in the short arm of the X chromosome, at locus Xp21.2 (*Muscular Dystrophy Association*). In the GRCh38 human genome assembly, the DMD gene spans the coordinates 31119222–33211549 on the X chromosome (*UCSC Genome Browser*).

The mutations in this gene vary. Some studies show 50-70% of *DMD* mutations are large deletions, about 20% include insertion and nucleotide point mutation; about 12% are duplications, and the rest fall into subexonic insertions, deletions, and missense mutations(Limback et al.). Research also shows that there is no correlation between the size of the gene deletions in *DMD* and the severity of the disease (Limback et al.).

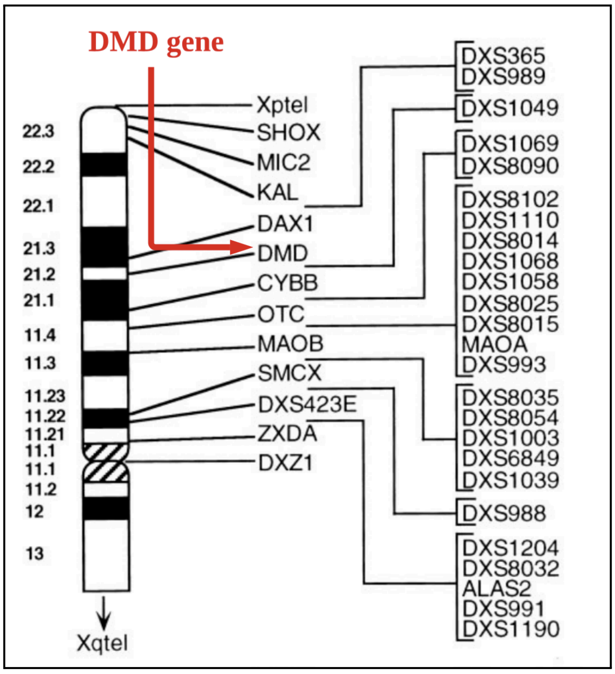


Figure 1. Location of the DMD gene (Ogata et al.)

3 Hypothesis

The hypothesis posits that mutation locations are expected to be distributed randomly rather than clustering in specific genomic coordination. This is predicted because, during protein folding, the different locations on the *DMD* gene may fold together. (Kumar et al.). Therefore, no significant hotspots are expected to be present on the gene.

On the other hand, mutations are predicted to occur more frequently in exonic regions than in intronic regions. As exons are the coding regions of a gene and are directly involved in protein synthesis, they can lead to changes to amino acid sequences or potentially disrupt protein function, hence they have a closer relationship to protein structure (Clancy). In contrast, introns are non-coding regions, and most of these regions are spliced out before translation to protein structure (Clancy).

As for mutation types, deletions are expected to be the most common in both exonic and intronic regions as deletions can bring the most severe consequences as these mutations often result in frameshift mutations, which can change the entire reading frame of codons, leading to a completely different protein product (Darras et al.). While insertions also cause frameshift effects, studies show that deletions are particularly detrimental in Duchenne muscular dystrophy, often leading to function loss in the dystrophin protein (Limback et al.).

4 Variables

Independent/Dependent Variables

The research question can be categorized into two sub-questions.

Table 1. Data Table Showing the Independent and Dependent Variables for Sub-Questions on Mutation Distribution, Frequency, and Mutation Type Analysis Across the Human DMD Gene

	Sub-Question 1	**Sub-Question 2**
Question	How are the mutations distributed across the DMD gene based on genomic coordination?	How do the frequencies of mutations differ in exonic/intronic regions, and how do the frequencies of different mutation types differ between these regions?
Independent Variable	The genomic coordinates. The precise coordinates on the X chromosome, based on GRCh38 range from 31119222–33211549.	The genomic regions (exon vs intron). Whether the mutation occurs in the exonic or intronic region of the gene. The mutation types. The four types of mutation include SNV, deletion, duplication, and insertion.
Dependent Variable	The patterns of mutation occurrence based on genomic coordinates. Whether the mutations are randomly distributed or clustered in specific areas.	Frequency of mutations in each region. The percentage of mutations identified in either the exonic or intronic region. Frequency of specific mutation types in each genomic region. The percentage of each mutation type in the exon and intron region, calculated separately.

Controlled Variables

Table 2. Controlled Variables, Methods of Controlling, Significance of Controlling

Controlled Variable	Method of Controlling	Significance of Controlling
Primary database	This study exclusively uses data extracted from ClinVar.	Data from a consistent source can improve reliability and minimize biases and mismatch information that may arise when using different sources.
Data Visualization Tool	This study uses the UCSC Genome Browser, linked from ClinVar, to visualize the precise genomic coordinates of each mutation.	Using a single genome browser ensures objectivity and accuracy in the data observed. Some other browsers may not provide detailed data such as the canonical form.
Mutation Classification from the Primary Database	In ClinVar, the *DMD* gene mutations are filtered to include only those classified as pathogenic and associated with only Duchenne Muscular Dystrophy.	Limiting the DMD gene collected ensures relevance to the study's goal, as these mutations are clinically significant for the disorder.
Human Genome Assembly	Uses GRCh38 (hg38), the 38th and most up-to-date version of the human genome sequence, as the reference map for all mutations.	Ensuring all genomic coordinates collected are based on the same reference map to maintain accuracy and consistency as different versions of genomic reference vary in information.
Types of Mutation	Only mutations that are deletions, SNVs, duplications, or insertions are included. Other types, like indels, were excluded because the mutation location on GRCh38 was not provided.	Limiting the mutation types creates a standard basis for comparison and prevents extraneous variables that complicate data interpretation. Poorly represented mutations like indels could introduce variability, and reduce the accuracy of the study.
Canonical Isoform	Only the canonical isoform ("default" version) of the gene's transcript is used to define exon/ intron boundaries.	Using the canonical isoform provides a standard comparison among different genes as it is the longest, most prevalent, conserved, and expressed gene.

5 Tools and Resources

5.1 Primary Database

This study uses ClinVar, a public archive of reports on human genetic variations associated with disease, as the primary data source. ClinVar aggregates data from laboratories, research institutions, and clinical sources, assigning unique accession numbers for reference. Each variant is reviewed for accuracy and transparency, allowing researchers to track consensus and conflicting clinical interpretations (*ClinVar*). In this study, ClinVar is used to filter pathogenic *DMD* gene mutations, identify mutation types, ensure that the conditions associated with *DMD* mutations are classified as (and only) Duchenne muscular dystrophy, locate *DMD* gene mutations on the X chromosome while providing their precise genomic coordinates based on the GRCh38 human genome assembly, and link to the UCSC Genome Browser for further visualization and localization of mutations.

5.2 Bioinformatics Tools

The UCSC Genome Browser, hosted by the University of California, Santa Cruz, serves as the main visualization tool for this study. It provides fast and interactive access to genomic data and also allows users to query genetic information, upload custom datasets, as well as visualize genomic data at in-depth levels (*UCSC Genome Browser*). In this study, this tool is used to visualize the mutated *DMD* gene, ensuring alignment with ClinVar's reference version for accurate data comparison, and giving a specific exon or intron number of the mutated point.

5.3 Research Relevance

This study investigates the distribution and frequency of pathogenic mutations in the human *DMD gene*. Identifying the hotspots of mutations could aid in future therapeutic methods, such as target gene therapy. Additionally, understanding the distribution of mutations in genomic regions can provide valuable insights into the DMD gene's architecture. Lastly, analyzing the mutation types could also reveal which mutations are more likely to disrupt the protein function, potentially influencing the severity of Duchenne muscular dystrophy.

6 Methodology

6.1 Preliminary Trial: Designing Data Collection Method

A preliminary trial was conducted to select the most effective tools and approaches for collecting pathogenic *DMD* genes from ClinVar and different bioinformatics tools to identify the exon/intron number and the genomic coordination.

a. Testing Sampling Approaches in ClinVar

Initially, mutation data were collected sequentially on ClinVar's results page. However, this approach led to a higher occurrence of specific mutation types due to the way ClinVar organizes data. To address this issue, random sampling was adopted as the primary approach.

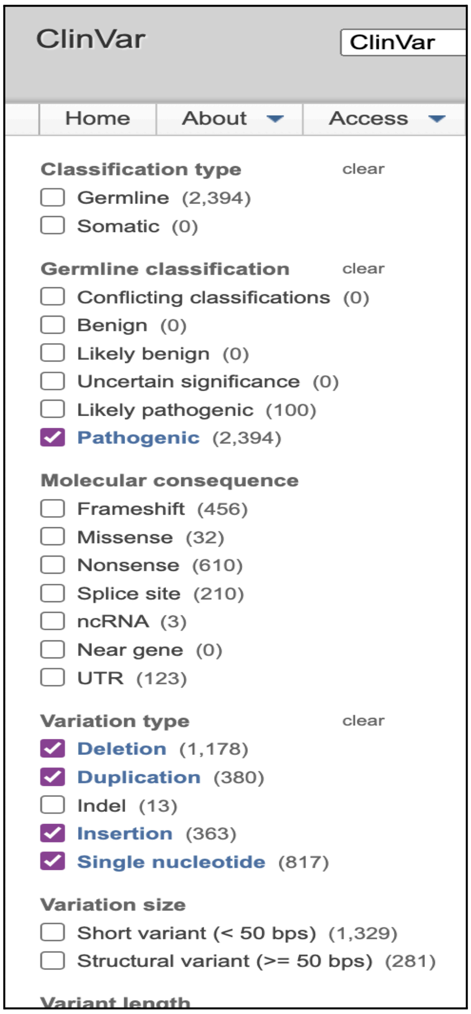
b. Using the NCBI Variation Viewer as a tool to visualize mutated gene

The NCBI Variation Viewer was used to locate mutations in the DMD gene. However, this tool proved to be overly complicated and lacked clarity in identifying exon/intron regions and showing the canonical isoform of genes, which slowed down the process. To improve efficiency, the UCSC Genome Browser, with its user-friendly interface and clear exon and intron annotations in the canonical isoform, was adopted as the primary visualization tool.

6.2 Procedure: Finalized Data Collection Method

1. Accessing the ClinVar database with filtering search
    a. Go to https://www.ncbi.nlm.nih.gov/clinvar/.
    b. Type "DMD gene" in the search box.

c. On the left side of the research page, under Germline Classification, select "Pathogenic." Under Variation Type, select "Deletion," "Duplication," "Insertion," and "Single nucleotide." See *Figure 2.*



ClinVar

ClinVar

Home | About ▾ | Access ▾

Classification type — clear

□ Germline (2,394)
□ Somatic (0)

Germline classification — clear

□ Conflicting classifications (0)
□ Benign (0)
□ Likely benign (0)
□ Uncertain significance (0)
□ Likely pathogenic (100)
☑ Pathogenic (2,394)

Molecular consequence

□ Frameshift (456)
□ Missense (32)
□ Nonsense (610)
□ Splice site (210)
□ ncRNA (3)
□ Near gene (0)
□ UTR (123)

Variation type — clear

☑ Deletion (1,178)
☑ Duplication (380)
□ Indel (13)
☑ Insertion (363)
☑ Single nucleotide (817)

Variation size

□ Short variant (< 50 bps) (1,329)
□ Structural variant (>= 50 bps) (281)

Variant length

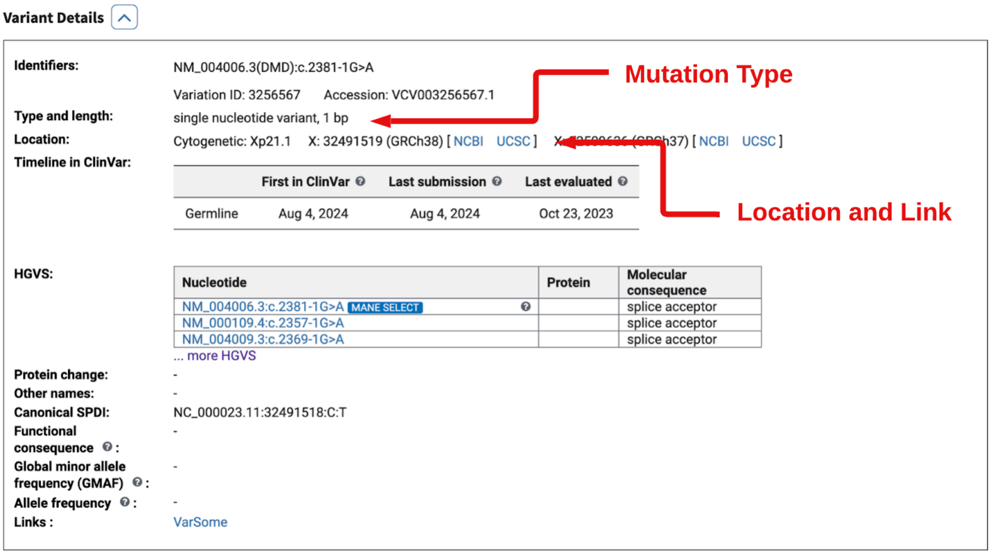
Figure 2. Filter Search For Pathogenic DMD gene (NCBI)

2. <u>Selecting a mutation entry and recording genomic location and mutation type</u>

a. From the filtered results, randomly select a mutation entry.

b. In the variant details, locate "Type and length" for mutation type; then locate "Location" for the specific position on the X chromosome (GRCh38). See *Figure 3.*

c. Locate “Timeline in ClinVar”, and press on “UCSC”, this directs to the UCSC genome browser. See *Figure 3.*



Variant Details

Identifiers:	NM_004006.3(DMD):c.2381-1G>A Variation ID: 3256567 Accession: VCV003256567.1
Type and length:	single nucleotide variant, 1 bp
Location:	Cytogenetic: Xp21.1 X: 32491519 (GRCh38) [NCBI UCSC] X: 32509636 (GRCh37) [NCBI UCSC]
Timeline in ClinVar:	

	First in ClinVar	Last submission	Last evaluated
Germline	Aug 4, 2024	Aug 4, 2024	Oct 23, 2023

HGVS:

Nucleotide	Protein	Molecular consequence
NM_004006.3:c.2381-1G>A MANE SELECT		splice acceptor
NM_000109.4:c.2357-1G>A		splice acceptor
NM_004009.3:c.2369-1G>A		splice acceptor

... more HGVS

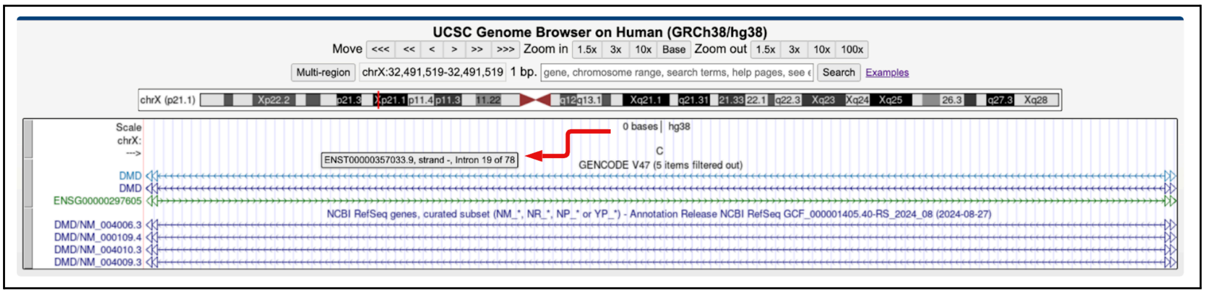
Protein change:	-
Other names:	-
Canonical SPDI:	NC_000023.11:32491518:C:T
Functional consequence:	-
Global minor allele frequency (GMAF):	-
Allele frequency:	-
Links :	VarSome

Figure 3. Mutation Type, Location, and UCSC Link on ClinVar Entry

3. <u>Visualizing Mutation Location in UCSC Genome Browser</u>

a. In the UCSC Genome Browser, find the light blue region, it represents the canonical isoform of the DMD gene. See *Figure 4.*

b. Hover over the light blue gene for data of exon/intron number. See *Figure 5.*



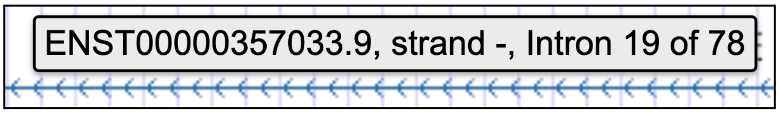
UCSC Genome Browser on Human (GRCh38/hg38)

ENST00000357033.9, strand -, Intron 19 of 78

GENCODE V47 (5 items filtered out)

NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release NCBI RefSeq GCF_000001405.40-RS_2024_08 (2024-08-27)

Figure 4. Hover Over the Light Blue Region



ENST00000357033.9, strand -, Intron 19 of 78

Figure 5. Intron Number (Zoom-In From Figure 4)

7 Risk Assessment

Table 3. Safety, Ethical, and Environmental Precautions

Category	Potential Concern	Mitigation Strategy
Safety	Entering unsafe websites can lead to malware attacks.	Used data with trusted origins; avoid downloading data from unknown websites.
Ethical	Using data beyond the allowed scope can lead to copyright and compliance issues.	Reviewed and adhered to websites' terms of use; Made sure data collected was only used for educational purposes.
Environmental	Lab-based experiments use chemicals that contribute to waste.	Used online public databases and tools instead of doing physical lab experiments.

8 Raw Data Table

Table 4 presents 30 unique pathogenic mutations in the DMD gene extracted from ClinVar, which focuses on the variation themselves rather than the individual cases.

Table 4. Raw Data of Pathogenic DMD Gene Mutations in Genomic Regions, Mutation Types, and Genomic Coordinate

Mutation ID	Genomic Region (Exon/Intron)	Mutation Type	Genomic Coordinate (GRCh38)
1	Exon (76/79)	Single Nucleotide Variant	X: 31146399
2	Intron (17/78)	Single Nucleotide Variant	X: 32545158
3	Intron (54/78)	Deletion	X: 31627653-31679606
4	Intron (48/78)	Duplication	X: 31836698-31968534
5	Exon (39/79)	Deletion	X: 32345970
6	Exon (33/79)	Deletion	X: 32386420
7	Exon (71/79)	Insertion	X: 31177933-31177934
8	Exon (36/79)	Insertion	X: 32364656-32364657
9	Intron (6/78)	Single nucleotide variant	X: 32809612
10	Exon (25/79)	Deletion	X: 32463533-32463545
11	Exon (65/79)	Deletion	X: 31209503-31209504
12	Exon (57/79)	Deletion	X: 31496932
13	Intron (50/78)	Single nucleotide variant	X: 31774194
14	Exon (33/79)	Duplication	X: 32386356-32386357
15	Exon (56/79)	Deletion	X: 31507303
16	Exon (35/79)	Deletion	X: 32365062
17	Exon (11/79)	Single nucleotide variant	X: 32644286
18	Exon (51/79)	Single nucleotide variant	X: 31774065
19	Exon (33/79)	Deletion	X: 32386453-32386454
20	Exon (55/79)	Deletion	X: 31627762
21	Exon (42/79)	Deletion	X: 32310168
22	Exon (38/79)	Deletion	X: 32348515
23	Exon(44/79)	Deletion	X: 32216950
24	Exon(58/79)	Insertion	X: 31479073-31479074
25	Exon(9/79)	Duplication	X: 32697928-32697929
26	Exon(6/79)	Duplication	X: 32816567-32816568
27	Intron (67/78)	Single Nucleotide Variant	X: 31203959
28	Exon(26/79)	Single Nucleotide Variant	X: 32454721
29	Exon(59/79)	Single Nucleotide Variant	X: 31478267
30	Intron (8/78)	Single Nucleotide Variant	X: 32698184

9 Data Analysis

9.1 Data Processing (Sub-Question 1)

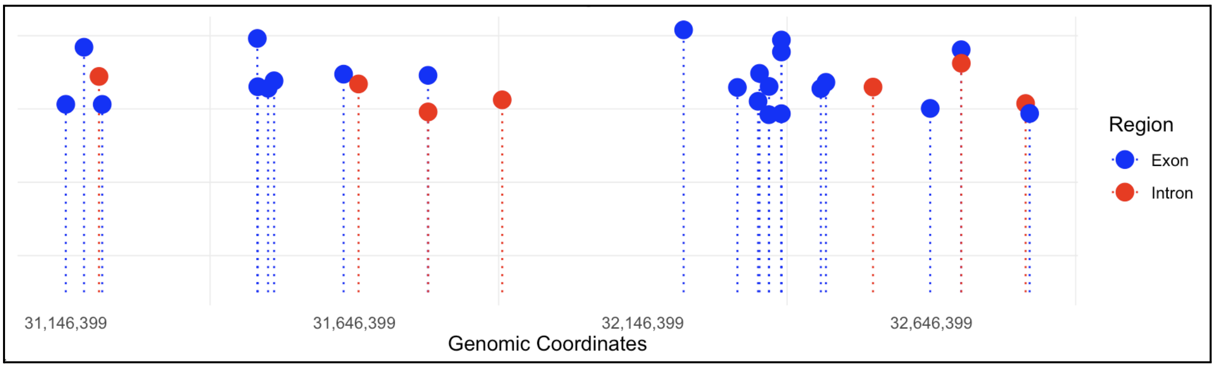
Genomic Coordinate: Extracted from *Table 4*. For values that have a range, the midpoint is calculated.

Table 5. Processed Data Table for Distribution of Mutation Points Based on Genomic Coordinates Across the Human DMD Gene (Full Data in Appendix)

Mutation ID	Genomic Coordinate (X chromosome)	Location (Exon/Intron)
1	31146399.0	Exon
2	32545158.0	Intron
3	31653629.5	Intron
…	…	…
29	31478267.0	Exon
30	32698184.0	Intron

The data presented in *Table 5* is visualized on a lollipop chart in *Figure 6*, with jittering applied. Jittering is the act of adding random noise to data points, preventing overlapping data from being hidden on the graph (Wicklin). The overlapping mutations are represented at different heights, and these mutations span the genomic coordinates from 31119222 to 33211549 (*UCSC Genome Browser*).

Figure 6. Mutation Distribution Map with Jittering Data



9.1.1 Runs Test for Randomness *(Python codes in appendix)*

A Runs Test for randomness was conducted in Python to determine whether the distribution of mutations is random or exhibits clustering. The test will be conducted with Python's `statsmodels` library, which provides the Z-statistic and P-value, to determine whether all the mutation points on the gene are randomly distributed across the genomic coordinates or there are patterns such as hotspot clustering. The null hypothesis (H_0) states the mutations are randomly distributed, while the alternative hypothesis (H_1) states that the mutations are more likely to cluster in certain regions based on genomic coordinates.

The genomic coordinates were normalized using min-max normalization, scaling the values to a range from 0 to 1. Then, the median of the normalized data was calculated. Each data point was compared to the median: if the normalized value was below the median, it was assigned a value of 0; if exceeded, the value of 1 was assigned. Python then analyzed the sequence of 0s and 1s to check for randomness, outputting both a Z-statistic and a P-value. Results:

$$\text{Z-Statistic: } 1.486446705914413 \approx 1.49$$

$$\text{P-Value: } 0.13716100358037725 \approx 0.137$$

The Z-statistic is a measure of how far the observed data is from the expected data, while the P-value indicates whether the observed deviation is statistically significant or due to random chance. The results show that the observed data are 1.49 standard deviations away from what would be expected in a random distribution of mutations. The P-value exceeds 0.05, which fails to reject the null hypothesis. This suggests that the mutations based on genomic coordinations, appear to be random across the DMD gene, with no significant evidence of clustering or hotspots. From a biological perspective, given the vast length of the DMD gene and the dispersed nature of mutations, random distribution may be a result of the gene's extensive sequence.

9.2 Data Processing (Sub-Question 2)

Table 5. Processed Data Table Showing the Frequency of Distribution of Each Mutation Type in Genomic Regions across the DMD Gene.

Mutation Type	Exon Counts	Exon Percentage (of Total Mutations)	Intron Counts	Intron Percentage(of Total Mutations)	Total Counts	Total Percentage
SNV	5	16.7 %	5	16.7 %	10	33.4 %
Deletion	12	40.0 %	1	3.3 %	13	43.3 %
Duplication	3	10.0 %	1	3.3 %	4	13.3 %
Insertion	3	10.0 %	0	0.0 %	3	10.0 %
Total	23	76.7 %	7	23.3 %	30	100.0 %

Intron/Exon Counts: Counting how many mutations occurred in intronic/exonic regions separately from *Table 4*.

Exon/Intron Percentage (of Total Mutations): Dividing the number of mutations in either region (exonic/intronic) by the total number of mutations across both regions then times 100.

Total percentage of mutations in the exonic region:

$$\frac{23}{(23+7)} \times 100 = 76.6666666667 \approx 76.7\ (\%)$$

Intronic regions:

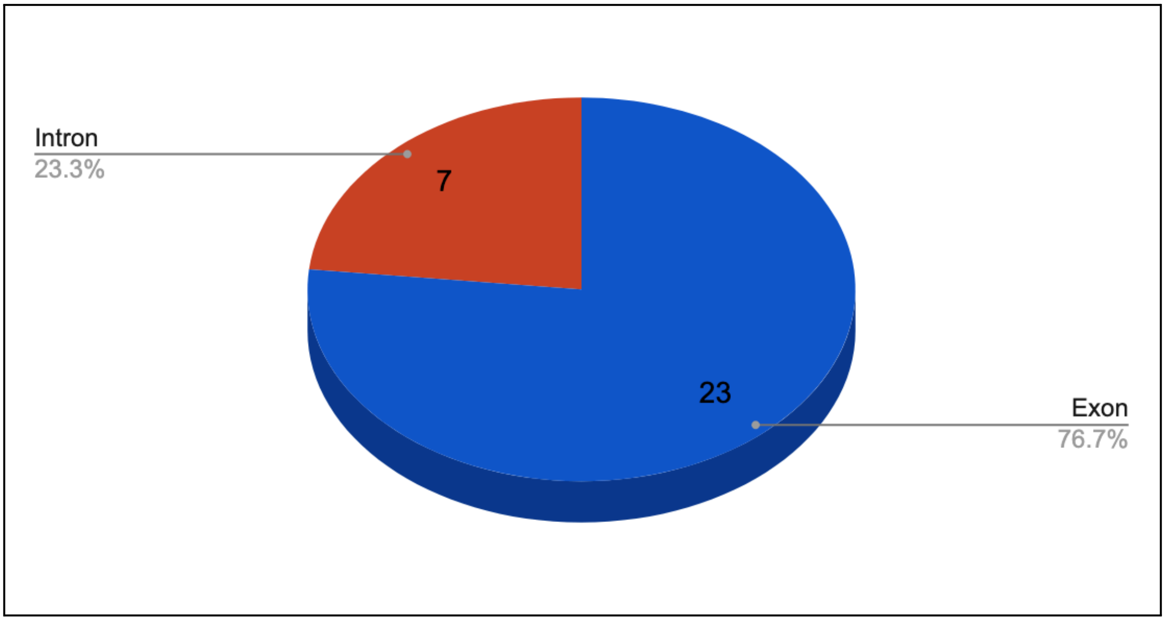
$$\frac{7}{(23+7)} \times 100 = 23.3333333333 \approx 23.3\ (\%)$$

or

$$100 - 76.7 = 23.3\ (\%)$$

Uncertainty: None, the counts were directly taken from the raw data table without measurement.

Figure 7. Pie Chart Showing the Frequency of Mutations in Exonic vs. Intronic Regions across the Human DMD gene



The pie chart clearly shows that both counts and percentage of mutations occurring in the exonic region are higher than in the intronic regions.

9.2.1 Chi-square Test

A Chi-square test was conducted to help observe whether there is a significant difference in the mutation distribution between exonic and intronic regions. The null hypothesis (H_0) posits that mutations occur randomly, with an equal probability (50%) of occurring in exons and introns. The alternative hypothesis (H_1) states that mutations are more likely to happen in the exonic regions.

The chi-square value's calculation is shown as:

$$x^2 = \frac{(23.3-50)^2}{50} + \frac{(76.7-50)^2}{50} = 28.5156 \approx 28.5$$

The degree of freedom (df):

$$df = 2 - 1 = 1$$

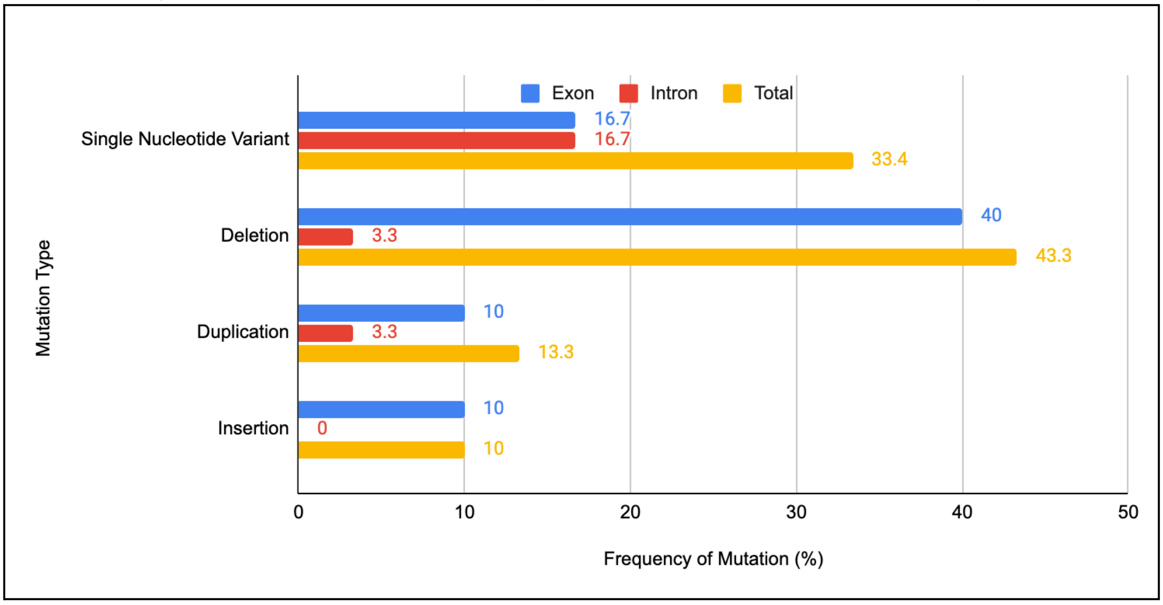
Next, compare the calculated chi-square value with the critical value from the chi-square distribution table for $p = 0.05$ and $df = 1$:

$$28.5 > 3.84$$

The calculated chi-square value exceeds the critical value, indicating that the result rejects the null hypothesis, high mutation counts in exonic regions compared to intronic regions are not due to chance and are statistically significant.

To further answer sub-question 2, the data are collected in *Table 5*, and presented in *Figure 8*.

Figure 8. Stacked Bar Graph Showing the Frequency of Mutation Types in Exons vs. Introns regions and Total Mutation Frequencies across the Human DMD gene



Total Percentage of Each Mutation Type: Calculated by adding the total mutation percentage from both exon and intron regions from the raw data table.

Calculator of SNV's total mutation percentage:

$$16.7 + 16.7 = 33.4\ (\%)$$

Uncertainty: None, as it is based on raw counts.

Figure 8 shows that deletions constitute the largest proportion, as they account for 43.3%, followed by SNV at 33.4%, then insertion at 13.3%, and duplication at 10.0%. *Figure 8* also shows that the most common mutation type in exonic regions is deletions, with a 40% probability. In intronic regions, SNVs have the highest frequency, which equals the frequency of SNVs in exonic regions. As for the least, in exonic regions, there is only a 10% chance of duplication and insertion mutation; in intronic regions, there are no insertion mutations.

10 Conclusion

Sub-Question 1: How are the mutations distributed across the DMD gene?

The investigation showed that all mutations collected were distributed rather than clustered in a particular area. The data presented in *Figure 6* supports the hypothesis of widespread distribution, as the calculated P-value from the runs test of randomness fails to reject the null hypothesis, accepting that the distribution of mutations is random based on genomic coordinates. However, studies have identified two deletion hotspots at exon 43–55 and exon 10–23 by calculating exon-deletion events of certain exon intervals (Chen et al.). These studies focused on specific exon deletions, which may not have been captured in my broad analysis. This highlights the difference in scale between more focused hotspot research and broad genomic research. Furthermore, the DMD gene's large size and complexity (Kumar et al.) may also explain the observed discrepancy, as it might contribute to both evenly distributed mutations and localized hotspots. Thus, while this paper's findings did not detect these hotspots, it does not reject the possibility of the existence of hotspots in smaller, more targeted studies.

Sub-Question 2: How do the frequencies of mutations differ in exonic and intronic regions, and how do the frequencies of different mutation types differ between these regions?

The investigation confirms that despite the mutation types, more mutations occurred in the exonic regions. The data collected in *Figure 7* supported the hypothesis, that 76.7% of mutations occur in exonic regions, while intronic mutations represent only 23.3%. This is reinforced by the chi-square test, which rejects the null hypothesis of a random distribution ($p < 0.05$), indicating that the genomic location is statistically significant. Gatto et al. found that most patients, approximately 80%, exhibit mutations across multiple exons. Additionally, The findings present in this paper also align with the biological context where exons encode protein products (Clancy). While introns experience alternative splicing during mRNA processing, only a few intronic that are not spliced out can be mutated on the gene. Hence, it is expected that mutations in exons occur more frequently. This paper also showed that deletions (40.0 %) are the most common in exonic regions, while SNVs (16.7%) are the most common in intronic regions in *Figure 8*. The data from *Figure 8* partially supports the hypothesis, it suggests that while exonic mutations are predominantly deletions, intronic mutations are more commonly SNVs. In Bladen et al's studies, they observed that 69% of total DMD mutations were deletions due to the severe consequences of frameshift mutations (Bladen et al.), which aligns with my findings in exonic regions. As for intronic regions, they are often less disruptive due to alternative splicing, allowing for more single nucleotide variants (Clancy). In *Journal of Neurology*, a study identified 19 different deep-intronic DMD variants in 30 patients, of which 15 were single nucleotide variants (Xie et al.). This finding is consistent with the data of this study where SNVs represent the most frequent mutation type in intronic regions.

In sum, this study showed that mutations on the *DMD* gene are distributed randomly based on genomic coordinates, supported by the Runs Test of randomness. It also showed that mutations occurred more in exonic regions, with deletions being the most common; single nucleotide variants were found to have the highest frequency in intronic regions, aligning with findings from previous studies.

11 Evaluation

11.1 Strengths

The primary database, ClinVar, is a reliable source for collecting human genetic variation data, as it includes reports of human variations classified for disease with supporting evidence. Moreover, the filter for specific data criteria is clear, ensuring the analysis is focused on clinically significant mutations relevant to DMD. Additionally, the use of the UCSC Genome Browser allows for precise visualization of mutation locations and coordinates across the gene. This enables highly detailed analysis and localization of mutations to analyze the pattern of mutation frequency.

11.2 Weaknesses and Limitations

Table 7. Weaknesses, Impact, and Suggested Improvements

Weakness	Impact on Investigation	Suggested Improvement
Small Sample Size	Only 30 pathogenic DMD mutations were collected. With such a small dataset, it is not representative enough to draw broad conclusions about mutations in this gene. Moreover, it also reduces statistical power, making it harder to detect significant patterns or trends.	Systematically increase the sample size from the ClinVar database or other databases. A larger sample dataset will improve statistical power, and reliability and allow for broader and definitive conclusions about the mutation frequency and distribution in specific regions across the gene.
Selection Bias in Data Collection	Although data were selected randomly through different ClinVar pages, true randomness was not achieved. As there are 27 pages of filtering results, not every page was used. This results in potential selection bias, with a tendency toward certain types of mutations; hence, the analysis of mutation type frequency in the gene could be biased.	Implement a Python script with imported math functions (e.g., `random.sample( )`) to automatically randomly select mutations within filtering criteria across all pages. This would enable a truly random sampling process, minimizing selection bias and resulting in a more representative dataset.
Precision in Mutation Location for Ranges	For mutations that are represented as a range, finding the midpoints approximation to map their distribution reduces accuracy. This can negatively affect the result of mutation distribution, especially for hotspot analysis. For instance, if a mutation is closer to the beginning or end of the range, using the midpoint could lead to inaccurate placement, potentially skewing the observed distribution and affecting the identification of mutation clustering.	Categorize range mutations separately, and conduct separate analyses for precise and range-based mutations to improve accuracy for distribution (hotspot) analysis. This minimizes oversimplification and ensures that both types of mutations are accurately represented.
Limited Severity Analysis	Without severity data indicating how impactful the mutation is (some data shows “Uncertain significance”), the study cannot draw correlations between mutation frequency and the severity of mutations in DMD. This limits the ability to determine whether specific types of mutations or mutation locations impact disease severity outcomes.	Use other databases that provide both mutation locations and severity data for DMD mutations, such as ClinGen. This would allow for a more comprehensive analysis of the relationship between clinical severity mutation location, and type, improving the clinical relevance of this study.

12 Extensions

Further studies can be conducted to analyze the DMD gene in depth. For instance, analyzing how specific mutations map to functional domains on the dystrophin protein, or investigating correlations between mutation types and disease severity—in other words, examining the genotype-phenotype relationship. Applying alignments of the protein sequence (or DNA) from patients diagnosed with Duchenne muscular dystrophy and healthy individuals can further validate or identify the hotspots in this disease. Moreover, further study could also use AI-based protein model prediction tools, such as AlphaFold, to map mutations to the functional domains of dystrophin. This would allow for findings on how mutations at different gene locations can affect various functional domains of the protein, thereby enhancing the understanding of genotype-phenotype relationships. While this study acknowledges the importance of these aspects, the lack of data on disease severity prevented a more in-depth exploration of these relationships.

13 Bibliography

Bladen, Catherine L., et al. "The TREAT‐NMD DMD Global Database: analysis of more than

7,000 Duchenne muscular dystrophy mutations." *Human mutation* 36.4 (2015): 395-402.

“Causes/Inheritance - Duchenne Muscular Dystrophy (DMD) .” *Muscular Dystrophy Association*, www.mda.org/disease/duchenne-muscular-dystrophy/causes-inheritance. Accessed 9 Dec. 2024.

Chen, Chen, et al. “Screening of Duchenne Muscular Dystrophy (DMD) Mutations and Investigating Its Mutational Mechanism in Chinese Patients.” *Screening of Duchenne Muscular Dystrophy (DMD) Mutations and Investigating Its Mutational Mechanism in Chinese Patients*, Public Library of Science, 22 Sept. 2014, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0108038

Clancy, Suzanne. “RNA Splicing: Introns, Exons, and Spliceosome.” *Scitable by Nature Education*, Nature Education, 2008, www.nature.com/scitable/topicpage/rna-splicing-introns-exons-and-spliceosome-12375/.

“ClinVar.” *U.S. National Library of Medicine*, National Institutes of Health, 2024, https://www.ncbi.nlm.nih.gov/clinvar/

Darras, Basil T, et al. “Dystrophinopathies.” *GeneReviews®*, U.S. National Library of

Medicine, 20 Jan. 2022, www.ncbi.nlm.nih.gov/books/NBK1119/.

“Debug my code for Runs test in Python.” prompt. ChatGPT, GPT-4o mini, OpenAI, Dec 2024, https://chat.openai.com/

“Duchenne Muscular Dystrophy.” *Johns Hopkins Medicine*, www.hopkinsmedicine.org/health/conditions-and-diseases/duchenne-muscular-dystrophy#:~:text=Duchenne%20muscular%20dystrophy%20is%20caused,in%203%2C500%20to%205%2C000%20newborns.

“DYSTROPHIN; DMD.” *Online Mendelian Inheritance in Man*, www.omim.org/entry/300377?search=DMD%20gene&highlight=dmd%2Cgene.

Gatto, Francesca, et al. “The complex landscape of DMD mutations: Moving towards personalized medicine.” *Frontiers in Genetics*, vol. 15, 26 Mar. 2024, https://doi.org/10.3389/fgene.2024.1360224.

“Give biological context why exonic mutations occurred more than intronic mutations in the DMD gene.” prompt. ChatGPT, G*PT-4o mini*, OpenAI, 13 Oct 2024, https://chat.openai.com/.

“How can I use the Runs Test for randomness in order to analyze mutation distributions, and what do the Z-statistic and P-value indicate about the randomness of the data?” prompt. *ChatGPT, GPT-4o mini*, OpenAI, 2 Dec 2024, https://chat.openai.com/.

How To Reject a Null Hypothesis Using 2 Different Methods, Indeed, 16 Aug. 2024, www.indeed.com/career-advice/career-development/reject-null-hypothesis.

Kumar, Shalini H, et al. "Comprehensive Genetic Analysis of 961 Unrelated Duchenne Muscular Dystrophy Patients." *PLOS ONE*, 19 June 2020, journals.plos.org/plosone/article/file?type=printable&id=10.1371/journal.pone.0232654.

Limback, Kylie, et al. "Biotechnology Journal International." *A Comprehensive Review of Duchenne Muscular Dystrophy: Genetics, Clinical Presentation, Diagnosis, and Treatment*, vol. 26, no. 6, 29 Dec. 2022, pp. 1–31, https://doi.org/10.9734/bji.

Ogata, Tsutomu, et al. "Turner syndrome and XP deletions: Clinical and molecular studies in 47 patients." *Turner Syndrome and Xp Deletions: Clinical and Molecular Studies in 47 Patients*, Nov. 2001, https://doi.org/10.1210/jcem.86.11.8058.

"Please fix my grammar and spelling." prompt. ChatGPT, G*PT-4o with canvas*, OpenAI, 16 Nov 2024, https://chat.openai.com/.

"Runs Test of Randomness in Python." *GeeksforGeeks*, GeeksforGeeks, 8 June 2020, www.geeksforgeeks.org/runs-test-of-randomness-in-python/.

"Signs and Symptoms of Duchenne Muscular Dystrophy (DMD) - Diseases." *Muscular Dystrophy Association*, www.mda.org/disease/duchenne-muscular-dystrophy/signs-and-symptoms.

"The NCBI Minute: Find All Variants with ClinVar." *YouTube*, YouTube, 29 July 2015, www.youtube.com/watch?v=H09-0pP48Us&t=20s.

UCSC Genome Browser Gateway, UCSC, genome-asia.ucsc.edu/cgi-bin/hGateway?redirect=manual&source=genome.ucsc.edu.

UCSC Genome Browser. "How Do I Identify Exon Numbers with the UCSC Genome Browser?" *YouTube*, 24 Jan. 2015, www.youtube.com/watch?v=gK8B6sjzhmM.

"What Is Clinvar?" *ClinVar*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/clinvar/intro/#:~:text=ClinVar%20is%20a%20freely%20accessible,drug%20responses%2C%20with%20supporting%20evidence.

Wicklin, Rick. "Jittering to Prevent Overplotting in Statistical Graphics." *Jittering to Prevent Overplotting in Statistical Graphics*, SAS Institute Inc, 5 July 2011, blogs.sas.com/content/iml/2011/07/05/jittering-to-prevent-overplotting-in-statistical-graphics.html#:~:text=Jittering%20is%20the%20act%20of,prevent%20overplotting%20in%20statistical%20graphs.

Xie, Zhiying, et al. "Clinical, muscle imaging, and genetic characteristics of dystrophinopathies with deep-intronic DMD variants." *Journal of Neurology*, vol. 270, 2 Nov. 2022, pp. 925–937, https://doi.org/10.1007/s00415-022-11432-0.

14 Appendix

Full Data presented in Table 5:

Mutation ID	Genomic Coordinate (X chromosome)	Location (Exon/Intron)
1	31146399	exon
2	32545158	intron
3	31653629.5	intron
4	31902616	intron
5	32345970	exon
6	32386420	exon
7	31177933.5	exon
8	32364656.5	exon
9	32809612	intron
10	32463544	exon
11	31209503.5	exon
12	31496932	exon
13	31774194	intron
14	32386356.5	exon
15	31507303	exon
16	32365062	exon
17	32644286	exon
18	31774065	exon
19	32386453.5	exon
20	31627762	exon
21	32310168	exon
22	32348515	exon
23	32216950	exon
24	31479073.5	exon
25	32697928.5	exon
26	32816567.5	exon
27	31203959	intron
28	32454721	exon
29	31478267	exon
30	32698184	intron

The code below is for the Runs Test for Randomness, written in Python:

```python
# Importing Packages and Models
from statsmodels.sandbox.stats.runs import runstest_1samp
import numpy as np
# Mutations' genomic coordinations
mutation_locations = np.array([
   31146399, 32545158, 31653629.5, 31902616, 32345970, 32386420,
   31177933.5, 32364656.5, 32809612, 32463544, 31209503.5,
   31496932, 31774194, 32386356.5, 31507303, 32365062, 32644286,
   31774065, 32386453.5, 31627762, 32310168, 32348515, 32216950,
   31479073.5, 32697928.5, 32816567.5, 31203959, 32454721,
   31478267, 32698184
])
# Normalize to range 0 to 1
normalized_locations = (mutation_locations - mutation_locations.min()) / (
   mutation_locations.max() - mutation_locations.min())
# Convert to binary data (above/below median)
median = np.median(normalized_locations)
binary_data = (normalized_locations > median).astype(int)
# Perform the runs test
z_stat, p_value = runstest_1samp(binary_data, correction=False)
# Output results
print(f"Runs Test Z-Statistic: {z_stat}")
print(f"Runs Test P-Value: {p_value}")
# Interpretation
if p_value < 0.05:
   print("The mutations are not random.")
else:
   print("The mutations appears random.")
```