Systematic Analysis of Genetic Variation of Duchenne Muscular Dystrophy and Implication for Cancer

Author: Hubert Chen

West Windsor Plainsboro High School South

New Jersey, USA

Mentor: Dr. Pingzhang Wang, Peking University

Betty Wang, IvyMind Academy

December 2020

Table of Contents

Abstract....1
1. Introduction of Duchenne Muscular Dystrophy....2
2. Research Workflow, Materials and Methods....3
2.1 Research aims and workflow....3
2.2 Source dbSNP database and SNPs extraction....4
2.3 Functional annotation with wANNOVAR....4
2.4 Retrieve ensembl transcript....5
2.5 Protein-protein interaction map....6
2.6 Integrative analysis genetic alternations in the DMD gene with cancer....6
3. Results....6
3.1 Variant prioritization....6
3.2 DMD gene annotation output....7
3.3 Example of raw annotation data....7
3.4 Exonic SNPs frequency in the DMD gene....8
3.5 DMD gene variants in Human 1000 Genomes Project....11
3.6 SNPs distribution by exons....12
3.7 SNPs distribution by ACMG classification....12
3.8 Network analysis highlights non-random interconnectivity between the genetic modifiers identified in DMD patient....14
3.9 Exon DMD genetic alternations in different tumors....15
3.10 Patients with DMD alterations have poorer overall survival....18
4. Discussions and Conclusions....20
5. Appendix....22
Acknowledgments....26
References....26

Abstract

Duchenne muscular dystrophy (DMD) is a rare, severe, progressive genetic disorder causing disability and premature death. Mutations in the DMD gene encoding the dystrophin protein lead to the dystrophinopathies DMD. Phenotypic variations in DMD may also occur in patients with the same primary mutation due to secondary genetic modifiers. Recent advances in molecular therapies for DMD have need of precise genetic diagnosis, since a large number of therapeutic approaches are mutation specific[1]. Interestingly, it has also been reported that muscular dystrophy patients may be at increased risk of malignancy[2, 3]. Although mutations in the DMD genes have been widely studied, to our knowledge, a systematic genetic analysis of all variants, especially single nucleotide polymorphisms (SNPs), of the gene in human have not been reported. In this study, we performed a systematic analysis of the DMD genetic variants via the Single Nucleotide Polymorphism Database (dbSNP), and annotated the functions of the variants with wANNOVAR [4]. The protein-protein interaction (PPI) network for genetic modifiers identified in DMD patients was explored. DMD genetic alternations in different tumors have also been investigated via cBioPortal [5].

We examined a total of 3,627 exonic SNPs in the DMD gene. SNPs distributed across all exons. The largest category was nonsynonymous account for nearly 64% of all mutations. Exon 19 appeared to have most density of pathogenic SNP distribution. Nonsense mutation (i.e. stopgain) or frameshift mutation likely lead to more pathogenic. Among the genetic modifiers identified in DMD patients, THBS1 has higher network topological parameters, followed by SPP1, ACTN3 and LTBP4. Network analysis highlighted non-random interconnectivity between the genetic modifiers identified in DMD patients, and potentially shed light on new genetic modifiers by their functional coupling to these known genes. Gene therapy is a promising experimental approach to treat genetic disease such as DMD[6]. In addition, our results also suggest DMD gene may serve as a diagnostic and therapeutic target for certain types of cancer.

Keywords:

DMD, dystrophin, SNPs, exon, genetic, mutation, protein-protein interaction, tumor, survival, data mining

1. Introduction of Duchenne Muscular Dystrophy

Duchenne muscular dystrophy (DMD), a rare X-linked disorder, is caused by a genetic mutation that prevents the body from producing dystrophin[7], a protein that enables muscles to work properly. It is one of the most common types of muscular dystrophy.

DMD symptom onset is in early childhood, usually between ages 3 and 5 years. Over time, children with Duchenne will have difficulty walking and breathing, then lead to disability, dependence, and premature death. The disease primarily affects boys, but in rare cases it also can affect girls. The prevalence of DMD is approximately 1 in 3500 to 5000 male births worldwide[8]. Muscle weakness is the principal symptom of DMD and worsen over time, first affecting the proximal muscles and later affecting the distal limb muscles. Patients with DMD progressively lose the ability to perform activities independently and often require a wheelchair by their early teens. As the disease progresses, life-threatening heart and respiratory conditions can occur[9]. In general, patients succumb to the disease in their 20s or 30s[9]; however, disease severity and life expectancy can vary.

DMD is caused by mutations in the DMD gene, which encodes the protein product called dystrophin. DMD gene is one of the largest of the identified human genes, spanning 2.4 Mb of a genomic sequence and corresponding to about 0.1% of the total human genome[10]. The gene consists of 79 exons encoding a 14,000 bp messenger RNA transcript that is translated into dystrophin[11]. The most common mutation responsible for DMD is a deletion spanning one or multiple exons. Such deletions account for 60–70% of all DMD cases. Point mutations are responsible for around 26% of DMD cases[12]. Exonic duplications account for 10 to 15% of all DMD case. Mutations in the DMD gene disrupt the protein's reading frame causing premature stop codons, leaving little or no functional dystrophin protein produced in cells. In addition, phenotypic variations in DMD may also occur in patients with the same primary mutation. A wide range of clinical manifestations suggest that genetic modifiers can influence the severity of DMD disease[13].

Dystrophin protein consists of 3,685 amino acids with molecular weight of 427 kDa and plays a crucial role in muscle function. Cytoplasmic face of the sarcolemma is where dystrophin is located[14]. Based on sequence homology, dystrophin is divided into four distinct structural domains. (i) amino terminal actin binding domain that contains two calponin homology domain which directly interacts with cellular actin cytoskeleton[15]; (ii) a long central rod shaped domain is composed of 24 structurally similar spectrin-type repeats[16]; (iii) cysteine rich domain binds to the intrinsic membrane protein β-dystroglycan; (iv) the carboxy-terminal domain binds to dystrobrevin and syntrophins[17]. Dystrophin is a cytoskeletal protein

that binds actin and associates with the dystrophin–glycoprotein complex (DGC) to link the cytoskeleton to the extracellular matrix (ECM)[18, 19]. The DGC consists of integral and peripheral proteins: dystroglycans, sarcoglycans, and syntrophins [18, 19]. Defects in dystrophin and/or other components of the DGC are responsible for several phenotypes of muscular dystrophy including DMD.

The DMD gene produces different transcripts encoding at least seven protein isoforms of various sizes and tissue distributions, each translated from a unique message initiated from one of the seven distinct promoters in the gene. Three full-length, large molecular weight isoforms (427 kDa), expressed primarily in skeletal muscle cells and cardiomyocytes, cortical neurons and the hippocampus of the brain and cerebellar Purkinje cells [20]. The shorter Dp260 and Dp116 isoforms are expressed primarily in the retina and the peripheral nerve, respectively [21]. Dp140 is expressed in the central nervous system (CNS) and kidney. There have been studies linking the lack of Dp140 expression in some DMD patients to cognitive impairment [22]. Finally, the Dp71 isoform is ubiquitously expressed, with higher levels in the CNS [23]. The literatures usually only describe seven protein coding isoforms. However, the Ensembl.org automatic annotation software predicts the presence of more protein coding splice variants of the dystrophin gene [24].

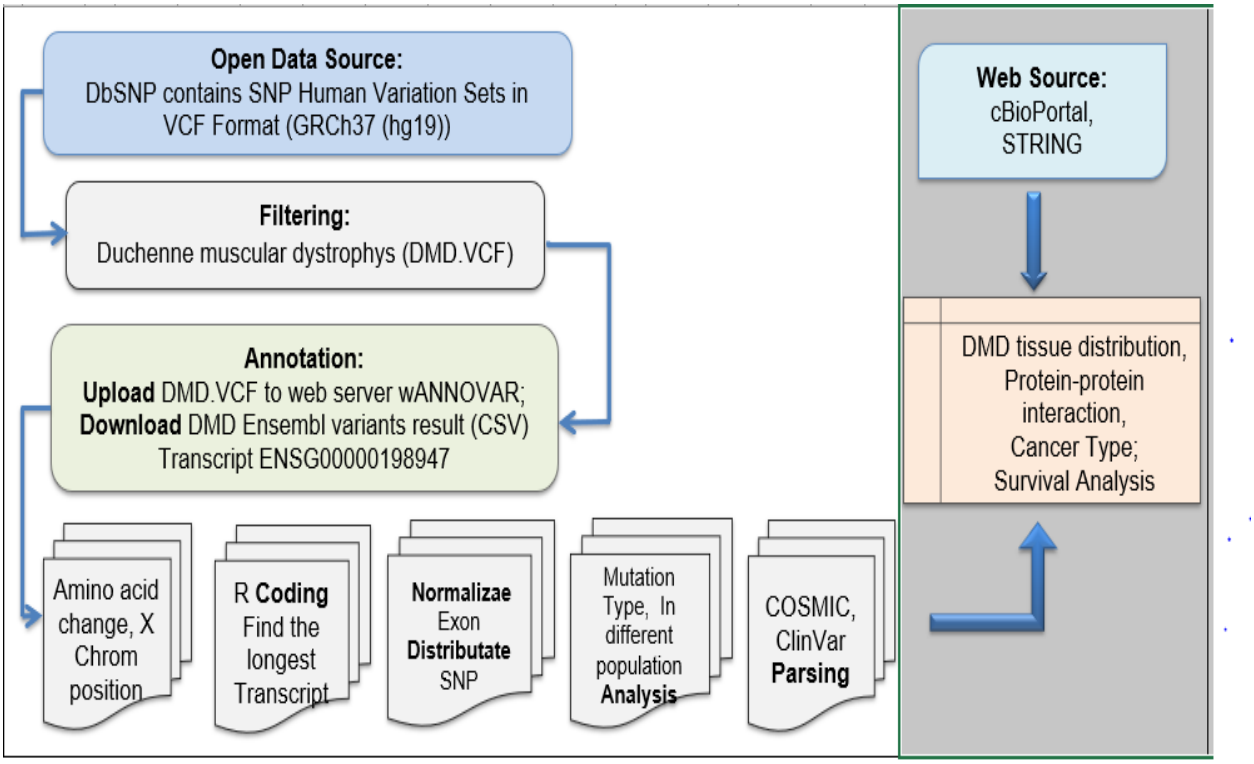
Although mutations in DMD genes have been widely studied, to our knowledge, a systematic genetic analysis of all variants, especially single nucleotide polymorphisms (SNPs), of the gene in human have not been reported. In this study, we carried out a systematic analysis of the DMD genetic variants via the dbSNP database, and annotated functions of these variants with wANNOVAR[4]. Protein-protein interactions for genetic modifiers identified in DMD patients were explored. In addition, potential relationships of genetic alternations in the DMD gene with cancer were also investigated.

2. Research Workflow, Materials and Methods

2.1 Research aims and workflow

The DMD genetic research mainly focus on using public-domain human genome databases to perform a systematic analysis of genetic variants especially SNPs in human populations, explore protein-protein interactions for genetic modifiers identified in DMD patients and examine relationships of genetic alternations in DMD gene and types of cancer. No statements of approval or informed consent were required for our study as we obtained data from an open access database. A research workflow was presented in Figure 1.

Figure 1: Research Workflow



In brief, all the genetic variants of DMD genes were extracted from the open access Single Nucleotide Polymorphism Database (dbSNP) database with variant call format (VCF). Variants using the online wANNOVAR server to generate output results files. These output files were annotated with function, distribution and disease etc[4]. Data mining and visualization were further performed based on the annotation.

2.2 Source dbSNP database and SNPs extraction

The source data for the DMD gene comes from the dbSNP, which was established by National Center for Biotechnology Information (NCBI) in collaboration with the National Human Genome Research Institute (NHGRI) in 1998 for GenBank of publicly available nucleic acid and protein sequences [32]. In order to extract specific human DMD gene data, build-in Unix compress tool "MacCompress" under a Mac environment was selected to unzip the large size vcf.gz file. To extract DMD variants, we used the command 'grep DMD 00-All.vcf > DMD.vcf', then added the header lines of 00-All.vcf into the file. After removing genes with mismatching, the DMD.vcf file was used for next steps of analysis.

2.3 Functional annotation with wANNOVAR

Online server wANNOVAR was used to annotate functional consequences of genetic variation [26]. Online annotating single nucleotide variants, insertions and deletions functions effects on genes, ClinVar database

information(Clinvar_DIS, Clinvar_ID), calculating predicted importance scores (SIFT_pred,Polyphen2) of each variant, retrieving allele frequencies in public databases (the 1000 Genomes Project and Genome Aggregation Database (gnomAD) of 15,708 genomes)[27], and implementing a 'variants reduction' protocol to identify a subset of potentially deleterious variants/genes[26].

2.4 Retrieve ensembl transcript

The Dystrophin gene ENSG00000198947 has different transcripts which described at Ensembl.org[28]. According to wANNOVAR instruction, the most popular approach was to use the longest transcript which provided the most genetic annotation information if multiple transcripts were available [24]. The full-length transcript variant of the dystrophin gene was ensembl transcript (ENST) ID ENST00000357033.9, with13992 length in base pairs (bp) [28]. R studio (R-4.0.2) was used to import "exome summary results.csv". The R program identified the longest transcript start and end, then split long characters string for AAChange.ensGene. It also picked the exon number and variant type, protein coding information (R scripts in appendix 5.1).

Figure 2-a: DMD comprehensive ensembl /gencode gene annotations

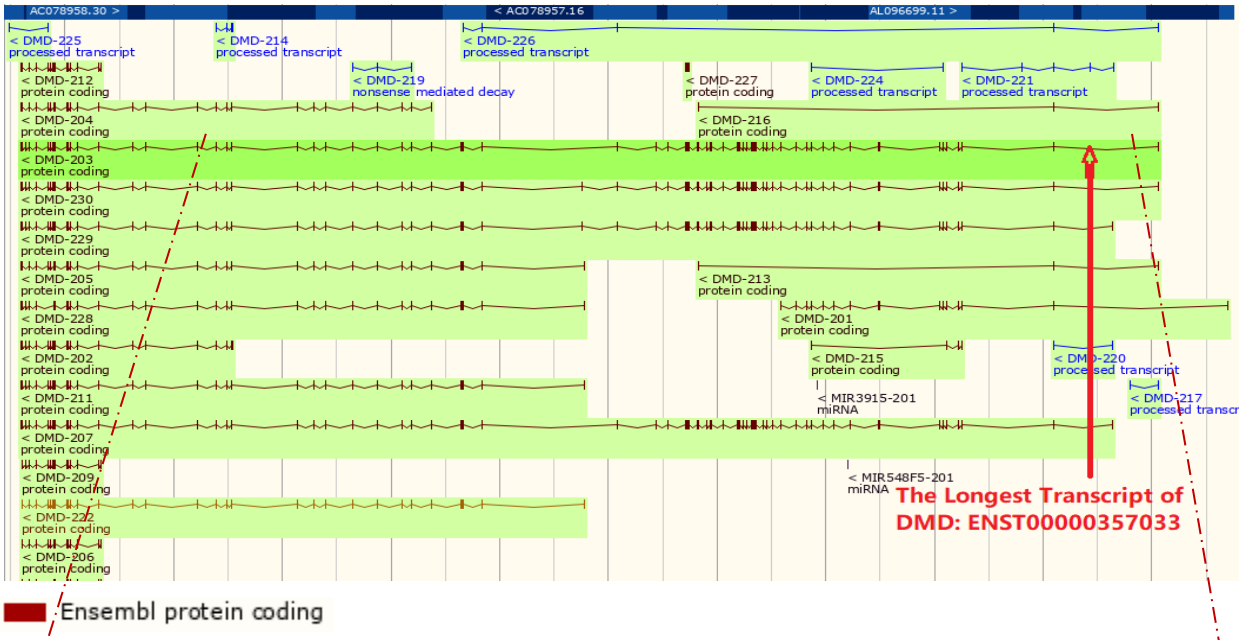


Figure2-b: The Longest DMD gene ENST00000357033.9 (has 79 exons,158 domains and features, 458334 variant alleles zooming from Figure 2-a DMD-203 [28].

2.5 Protein-protein interaction map

Protein-protein interaction map for genetic modifiers identified in DMD patients was constructed using search tool for the retrieval of interacting genes/proteins (STRING v11) [29]. Active interaction sources were restricted to "Textmining," "Experiments," and "Databases." Only interactions with confidence score over 0.9 were mapped to the network. The network obtained from STRING was subsequently analyzed using Cytoscape 3.8.1 plugin Network Analyzer [30]. The nodes in the PPI network represented the genes/proteins, and the edges between the nodes represented the interactions between them. Nodes with high degree and betweenness centrality (BC) value were considered as key parameters to analyze the network. The node with a high degree was deemed with an essential biological function.

2.6 Integrative analysis genetic alternations in the DMD gene with cancer

Genetic alternations in the DMD gene with cancer was examined by using Cancer Genomics Portal (cBioPortal) [5]. This web portal provides query interface combined with customized data enabled us to interactively explore genetic alterations across DMD samples, genes mutation site and frequency. Kaplan-Meier curves stratified by genotype were plotted and comparisons were tested using the Log-rank test.

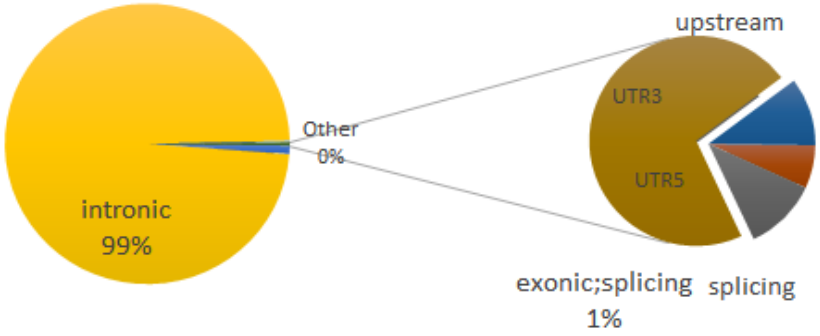
3. Results

3.1 Variant prioritization

Online wANNOVAR submission for DMD.vcf processed a total of 382160 variants. Variants located mainly in intronic, exonic, and UTR3 variant regions (Figure 3). There are two DMD functional annotation output files: "genome summary results.csv" and "exome summary results.csv". In this study, genetic variants in exons rather than intron was focused on, because variants in exonic (coding) region often can alter the protein function. Therefore, "exome summary results.csv" was selected.

Figure 3: Catalog of genomic regions in the DMD gene

Func.ensGene	Count
downstream	151
exonic	3626
exonic;splicing	1
intergenic	203
intronic	376096
ncRNA_exonic	340
ncRNA_intronic	251
ncRNA_splicing	1
splicing	99
upstream	167
UTR3	1069
UTR5	156
Total	*382160*



"genome summary results.csv has all processed variants 382160 output, 99% is intronic. "exome summary results.csv" only contain exonic output. The pie chart depicts the frequency of variant location.

3.2 DMD gene annotation output

DMD "exome summary results.CSV" file contains a total of 3627 rows and 140 columns, which includes information such as function class of exonic variant. We presented some selected key information (Table 1) in the annotated file that is crucial for our research.

Table 1: Example of selected key columns in the annotated file.

Columns Name	Description	Annotation
Start	Start nucleotide number on X chromosome.	31140001
End	End nucleotide number on X chromosome.	31140003
Ref	Original nucleotide(s) present before mutation.	C
Alt	Alternative nucleotide(s) present after mutation.	G
Gene.ensGene	Gene Variant number.	ENSG00000198947
Func.ensGene	Regions (e.g., exonic, intronic, non-coding RNA)) that one variant hit	
ExonicFunc.ensGene	Exonic variant function, e.g., synonymous, nonsynonymous, frameshift insertion.	nonsynonymous SNV
AAChange.ensGene	Amino Acid Change	ENST00000357033.9: exon45:c.G6502C:p.E2168Q ENST00000378677: exon45:c.G6490C:p.E2164Q
COSMIC_DIS	Distribution in Catalogue of Somatic Mutations in Cancer	1(prostate)
ClinVar_SIG	ClinVar uses standard terms for clinical significance recommended by an authoritative source when available. These standards include five terms for diseases [33]	Pathogenic
ClinVar_DIS	Clinical significance parameter	Duchenne muscular dystrophy
SIFT_score SIFT_pred	SIFT score (deleterious or tolerated) [31] D: Deleterious (sift<=0.05); T: Tolerated (sift>0.05)	*0.239* *D*
Polyphen2 HDIV_score	Impact on amino acid sequence and protein function.	0.291
1000_exome_ALL	1000 Genomes Project dataset with allele frequencies in six populations including ALL, AFR (African), AMR (Admixed American), EAS (East Asian), EUR (European), SAS (South Asian). These are whole-genome variants[31].	0.005

3.3 Example of raw annotation data

The output DMD "exome summary results.csv" file with raw data can be opened by Excel or Imported into R Studio. Each row represents one variant, and each column represents one annotation task. Users can

specify columns or annotation, and also select multiple gene-definition systems. By filtering out variants, and then selecting key columns such as AAChange, 1000G_AMR, ExonicFun, SFIT_Score, ClinVar_SIG to analyze the genetic alterations.

Table 2 -a: Example of variants in DMD Gene "exome summary results.csv" output

Chr	Start	End	Ref	Alt	Func.en	Gene.ensGene	ExonicFunc.ensGene	AAChange.ensGene
X	31139958	31139958	C	G	exonic	ENSG00000198947	nonsynonymous SNV	ENSG00000198947:ENST00000378680:exon13:c.G1570C:p
X	31139961	31139963	CTT	-	exonic	ENSG00000198947	nonframeshift deletion	ENSG00000198947:ENST00000378680:exon13:c.1565_156
X	31139963	31139963	T	C	exonic	ENSG00000198947	nonsynonymous SNV	ENSG00000198947:ENST00000378680:exon13:c.A1565G:p
X	31139975	31139975	-	TGAC	exonic	ENSG00000198947	frameshift insertion	ENSG00000198947:ENST00000378680:exon13:c.1552_155
X	31139978	31139983	ACTGAT	-	exonic	ENSG00000198947	nonframeshift deletion	ENSG00000198947:ENST00000378680:exon13:c.1545_155
X	31140001	31140013	TCTGCCCA	-	exonic	ENSG00000198947	frameshift deletion	ENSG00000198947:ENST00000378680:exon13:c.1515_152
X	31165594	31165595	TC	CCCCACTT	exonic	ENSG00000198947	stopgain	ENSG00000198947:ENST00000378680:exon10:c.1060_106

Table 2 -b: "*AAchange. enGene*" of zooming for Amino Acid Change Information

		AAChange.ensGene
Exonic		ENSG00000198947:ENST00000378705:exon2:c.106_113del:p.R36Gfs*2,ENSG00000198947:ENST00000361471:exon6:c.532_539del:p.R178Gfs*1,ENSG00000198947:ENST00000378680:exon6:c.532_539del:p.R178Gfs*1,ENSG00000198947:ENST00000378702:exon6:c.532_539del:p.R178Gfs*1,ENSG00000198947:ENST00000378723:exon6:c.532_539del:p.R178Gfs*1,ENSG00000198947:ENST00000358062:exon20:c.2824_2831del:p.R942Gfs*1,ENSG00000198947:ENST00000343523:exon24:c.2356_2363del:p.R786Gfs*1,ENSG00000198947:ENST00000359836:exon24:c.2356_2363del:p.R786Gfs*1,ENSG00000198947:ENST00000378707:exon24:c.2356_2363del:p.R786Gfs*1,ENSG00000198947:ENST00000474231:exon24:c.2356_2363del:p.R786Gfs*1,ENSG00000198947:ENST00000541735:exon24:c.2356_2363del:p.R786Gfs*1,ENSG00000198947:**ENST00000357033:exon67:c.9736_9743del:**p.R3246Gfs*1,ENSG00000198947:ENST00000378677:exon67:c.9724_9731del:p.R3242Gfs*1

3.4 Exonic SNPs frequency in the DMD gene

We choose wANNOVAR's default definitions of exonic functional categories in order of precedence: deletions, insertions or substitutions for frameshift; stop gain; start loss; deletions, insertions, or substitutions for nonframeshift; nonsynonymous and synonymous (Figure 4). A total of 3627 exonic SNPs in the DMD gene has been examined. The largest category was nonsynonymous account for nearly 64% of all mutations. Synonymous accounted for nearly 43% and followed by stop gain account for nearly 6.7% of all mutations.

Figure 4: Variants Type and Frequency in the DMD Gene

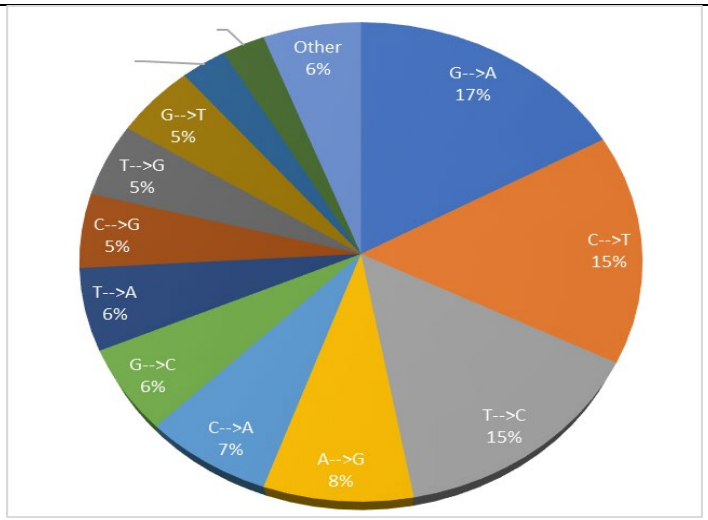
Variants Definition	Variants Type	Count	Frequency
Insertion, deletion or substitution of one or more nucleotides that cause frameshift changes in protein coding sequence.	frameshift insertion	47	1.30%
	frameshift deletion	116	3.20%
	frameshift substitution	3	0.08%
Variant that leads to the immediate creation of stop codon at the variant site;	stopgain	242	6.67%
Variant that leads to the immediate elimination of stop codon at the variant site	startloss	1	0.03%
	nonframeshift insertion	4	0.11%

Insertion, deletion or substitution of 3 or multiples of 3 nucleotides that do not cause frameshift changes in protein coding sequence.	nonframeshift deletion	27	0.74%
	nonframeshift substitution	1	0.03%
A single nucleotide change that cause an amino acid change	nonsynonymous	2322	64.02%
A single nucleotide change that does not cause an amino acid change	synonymous	864	23.82%
	Total variants	**3627**	

High frequency nucleotide substitutions have been observed in the DMD gene. The largest category was Guanine (G) to Adenine (A) with 17%, followed by Cytosine (C) to Thymine (T) with 15% and Thymine (T) to Cytosine (C) with 15%, respectively. (Figure 5).

Figure 5: Summary of high frequency of amino acid change in the DMD gene

Count and Percentage of High frequency AAChange		
G-->A	614	17%
C-->T	561	15%
T-->C	533	15%
A-->G	303	8%
C-->A	259	7%
G-->C	213	6%
T-->A	203	6%
C-->G	182	5%
T-->G	178	5%
G-->T	177	5%
A-->T	100	3%
A-->C	93	3%
Other	211	6%
Toatl	*3627*	



Examples of DMD gene mutation have been presented in Table 3. In brief, the variants table fields are listed as positional variant start, positional variant end, reference allele sequence at variant locus, alternate allele sequence at variant locus. Same as Figure 4, the largest category is nonsynonymous, follow by synonymous and stop gain, respectively.

Table 3: Examples overview and summary of DMD gene mutation

frameshift deletion				*frameshift deletion*				*frameshift substitution*				*startloss*				*stopgain*			
Chr Start	Chr End	Ref	Alt	Chr Start	Chr End	Ref	Alt	Chr Start	Chr End	Ref	Alt	Chr Start	Chr End	Ref	Alt	Chr Start	Chr End	Ref	Alt
31140001	31140013	TCTGCCCAAATCA	-	31139975	31139975	-	TGAC	31838158	31838160	GGT	AAAC	31526352	31526352	C	A	31165538	31165538	G	A
31140027	31140039	AAAGACTTCCTAC	-	31165537	31165537	-	GG	32466574	32466580	TCCAAAG	CC	31526352	31526352	C	T	31165594	31165595	TC	CCCCACTTTAAAGTTTCTTTAAAGTTT
31165415	31165415	T	-	31165586	31165586	-	T	33229415	33229416	TT	A	31526354	31526354	T	C	31165617	31165617	A	T
31165464	31165464	C	-	31187706	31187706	-	TT					32430174	32430174	T	C	31187609	31187609	C	A
31187665	31187666	GA	-	31196052	31196052	-	A					32834744	32834744	A	G	31200887	31200887	-	TTAC
31190498	31190498	T	-	31950285	31950285	-	AAAC					32834745	32834745	T	C	32583907	32583907	-	ACTT
31196081	31196093	AACGGGACTGCAA	-	31950287	31950287	-	GAAGT									32583911	32583911	T	A
Total: 116				*Total: 47*				*Total: 3*				*Total: 1*				*Total: 242*			

nonframeshift deletion				*nonframeshift insertion*				*nonframeshift substitution*				*nonsynonymous SNV*				*synonymous SNV*			
Chr Start	Chr End	Ref	Alt	Chr Start	Chr End	Ref	Alt	Chr Start	Chr End	Ref	Alt	Chr Start	Chr End	Ref	Alt	Chr Start	Chr End	Ref	Alt
31191709	31191714	GGACGA	-	31462643	31462643	-	TTATACGGTGAGAGC	32503201	32503203	GCC	AAT	31139958	31139958	C	G	31139977	31139977	G	T
31196906	31196908	TCT	-	31792099	31792099	-	TGA					31139958	31139958	C	T	31139983	31139983	T	C
31196910	31196912	CTC	-	31950265	31950265	-	ATC					31139963	31139963	T	C	31139998	31139998	C	T
31514959	31514964	CTGCCG	-	31986593	31986593	-	ACA					31139988	31139988	A	T	31140004	31140004	G	A
31747784	31747792	TGGTCTT	-									31139991	31139991	A	G	31164421	31164421	A	G
32235148	32235150	CGC	-									31139997	31139997	T	A	31152229	31152229	A	T
32364097	32364102	TTTATC	-									31139999	31139999	G	A	31187658	31187658	C	G
Total: 27				*Total: 4*				*Total: 1*				*Total: 2322*				*Total: 864*			

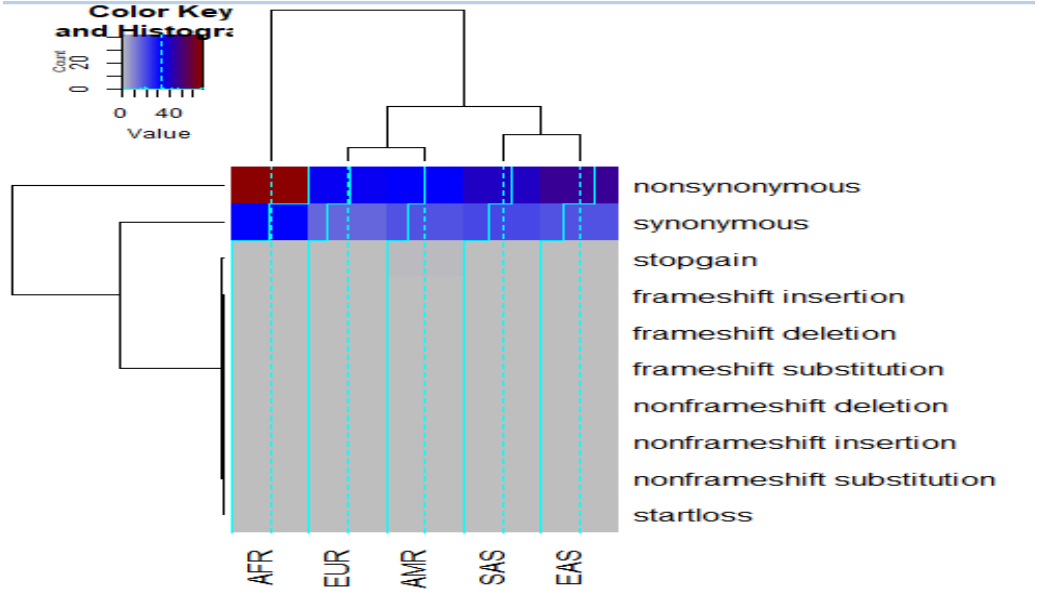
3.5 DMD gene variants in Human 1000 Genomes Project

DMD ensemble in this project contains the 1000 Genomes Project data in annotation. 1000G was designed to provide a comprehensive description of human genetic variation through sequencing multiple individuals. Populations were classified into 5 major continental groups: Africa (AFR), America (AMR), Europe (EUR), East Asia (EAS), and South Asia (SAS)[34].The DMD gene variants frequency from several different continents were showed in Figure 6-a and Table 6-b. Approximately 7% nonsynonymous mutations were observed from 1000 Genomes project, with relatively higher alteration frequency in African (~3%). Similar frequency distributions were observed among America, Europe, East Asia, and South Asia. No frameshift and nonframeshift mutations and startloss were reported in 1000 Genomes Project.

Figure 6-a: DMD gene variants frequency in Human 1000 Genomes Project

Variants Type	All DMD Exome	1000G ALL	1000G AFR	1000G AMR	1000G EAS	1000G EUR	1000G SAS
nonsynonymous	2320	165 (7.1%)	67 (2.9%)	34 (1.6%)	47 (2%)	35 (1.5%)	41 (1.8%)
stopgain	242	2 (0.8%)	0	1 (0.4%)	0	0	0
synonymous	735	83(11.3%)	33 (4.5%)	19 (2.6%)	19 (2.6%)	16 (2.2%)	21 (2.9%)
frameshift deletion	95	0	0	0	0	0	0
frameshift insertion	37	0	0	0	0	0	0
frameshift substitution	2	0	0	0	0	0	0
nonframeshift deletion	22	0	0	0	0	0	0
nonframeshift insertion	2	0	0	0	0	0	0
nonframeshift substitution	1	0	0	0	0	0	0
startloss	1	0	0	0	0	0	0
Total	***3297***	***250 (7.6%)***	***100 (3%)***	***154 (4.7%)***	***66 (2%)***	***41 (1.2%)***	***63 (1.9%)***

Figure 6-b: DMD gene variants frequency from different continents. Different colors represent for different levels of frequency values by R Heatmap Script in the appendix 5.2).



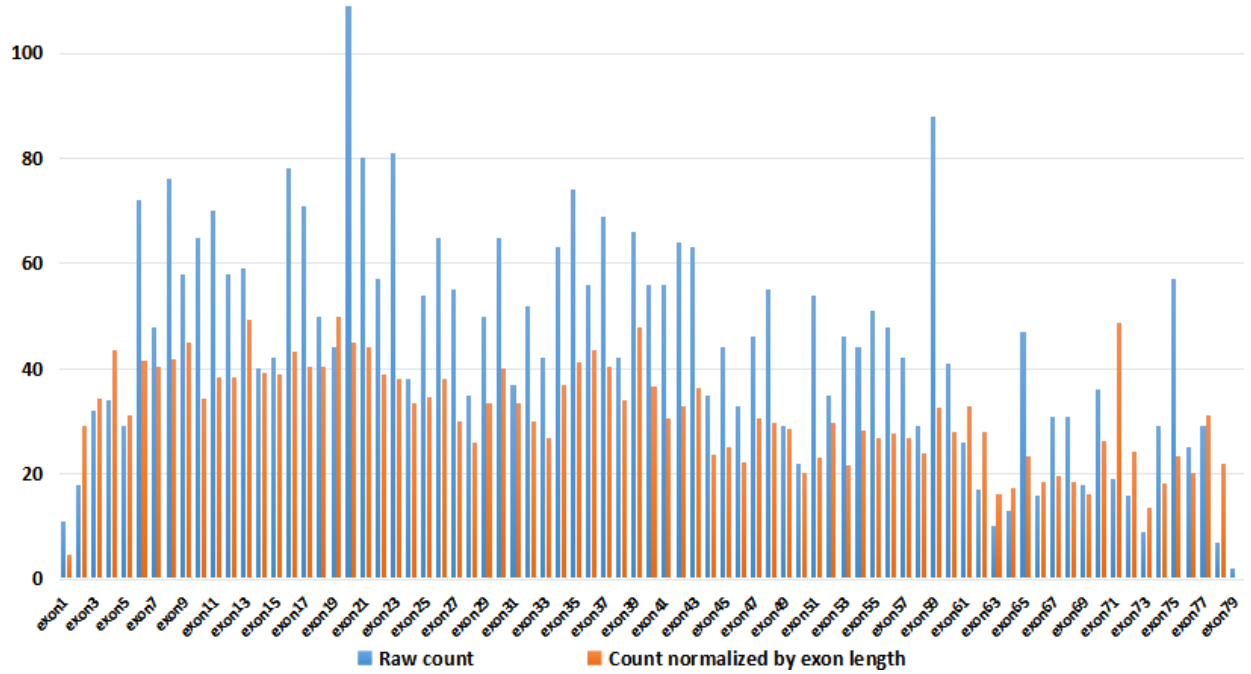
3.6 SNPs distribution by exons

In the DMD gene, SNPs were distributed across all the exons. Exon 79 is the longest with 2703 bp in length. Exon 78 is the shortest with 32 bp. To a better understanding of SNPs distribution density, we normalized the length of the exon, and as such high-frequency SNPs were observed in exon 19, 13, 71, respectively (Figure 7). Interestingly, although Exon 79 is the longest exon, there is a low distribution of genetic variants in this gene region. A further investigation revealed that the majority region of the exons belongs in the 3'untranslated region (3'-UTR). With consideration for the low SNP frequency in the 5'-UTR, demonstrates that the genetic variation of DMD gene mainly occurs in the protein-coding region. This implies a potential genetic interaction between variation and protein function.

Figure 7-a: Exon 78 and Exon 79 length example of DMD gene ENST00000357033.9

X	protein_coding	transcript	Start	End	Length	gene_id "ENSG00000198947"; transcript_id "ENST00000357033"; gene_name "DMD"; gene_source "ensembl_h
X	protein_coding	exon	31144759	31144790	32	gene_id "ENSG00000198947"; transcript_id "ENST00000357033"; exon_number "78"; gene_name "DMD"; gene_
X	protein_coding	exon	31137345	31140047	2703	gene_id "ENSG00000198947"; transcript_id "ENST00000357033"; exon_number "79"; gene_name "DMD"; gene_

Figure 7-b: The frequency and distribution density of genetic variants in the exons of DMD gene



The x-axis only shows exon odd-number due to space

3.7 SNPs distribution by ACMG classification

Examples of DMD gene variants by American College of Medical Genetics and Genomics (ACMG) Classifications presented in Table 4. Variants according to ACMG Standards and Guidelines for clinical

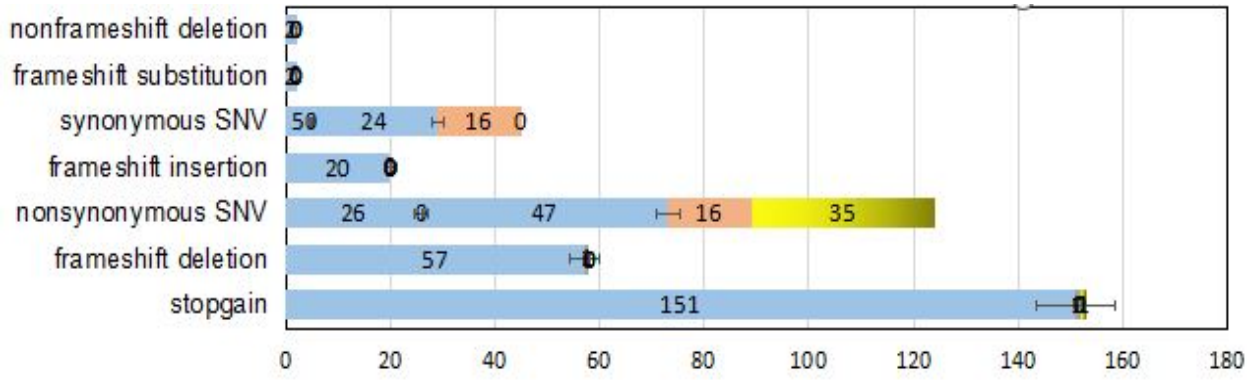
laboratories were classified as: 'Benign', 'Likely benign', 'Pathogenic', 'Likely pathogenic'[35] in table 4.

Table 4: Examples of DMD gene variants by ACMG-AMP classifications

Chr	Start	Ref	Alt	ExonicFunc.ensGene	Exon	1000G AFR	1000G AMR	1000G EAS	1000G EUR	1000G SAS	COSMIC_DIS	ClinVar_SIG	ClinVar_DIS
X	31196876	T	-	frameshift deletion	exon70	.	.	.	.	.	large_intestine	Pathogenic	Duchenne_muscular_dystrophy
X	33229415	TT	A	frameshift substitution	exon1	.	.	.	.	.	.	Pathogenic|	Duchenne_muscular_dystrophy
X	32867917	C	G	startloss	exon3	.	.	.	.	.	.	.	.
X	31196906	TCT	-	nonframeshift deletion	exon70	.	.	.	.	.	.	Pathogenic|	Duchenne_muscular_dystrophy
X	31514988	G	A	stopgain	exon57	.	.	.	.	0.001	.	Pathogenic	Dilated_cardiomyopathy_3B
X	32305668	T	A	stopgain	exon43	.	0.002	.	.	.	.	.	.
X	32519950	G	C	nonsynonymous SNV	exon19	.	.	0.0013	.	.	.	Pathogenic|	Duchenne_muscular_dystrophy
X	31462649	C	T	synonymous SNV	exon60	0.001	.	.	.	.	large_intestine	Uncertain si	not_specified
X	31497112	G	C	nonsynonymous SNV	exon58	0.001	.	.	.	.	.	Pathogenic|	Duchenne_muscular_dystrophy
X	31497186	A	G	nonsynonymous SNV	exon58	.	.	0.0013	.	.	.	Uncertain si	not_specified
X	31792190	G	A	nonsynonymous SNV	exon51	0.001	.	.	.	.	stomach	Uncertain si	Cardiovascular_phenotype
X	32305804	A	G	synonymous SNV	exon43	.	.	.	.	0.001	.	Likely benig	not_specified
X	32366608	G	A	nonsynonymous SNV	exon38	.	.	.	0.0013	.	.	Pathogenic|	Duchenne_muscular_dystrophy
X	32456488	C	T	nonsynonymous SNV	exon29	.	.	0.0013	.	.	central nervous system	Benign	Duchenne_muscular_dystrophy
X	32472815	A	G	synonymous SNV	exon26	.	0.002	.	.	.	.	Benign	not_specified
X	32481613	C	T	synonymous SNV	exon25	.	.	0.0013	.	.	large_intestine	Uncertain si	not_specified

Distribution of SNPs by ACMG Classifications in the DMD Gene presented in Figure 8. Stopgain caused 57% pathogenic, followed by frameshift (22%). In general, most of synonymous mutation have no functional consequence. Interestingly, we observed a few cases with synonymous mutation (2%) also associated with pathogenic DMD in the current analysis.

Figure 8: Distribution of SNPs by ACMG-AMP classifications in the DMD gene

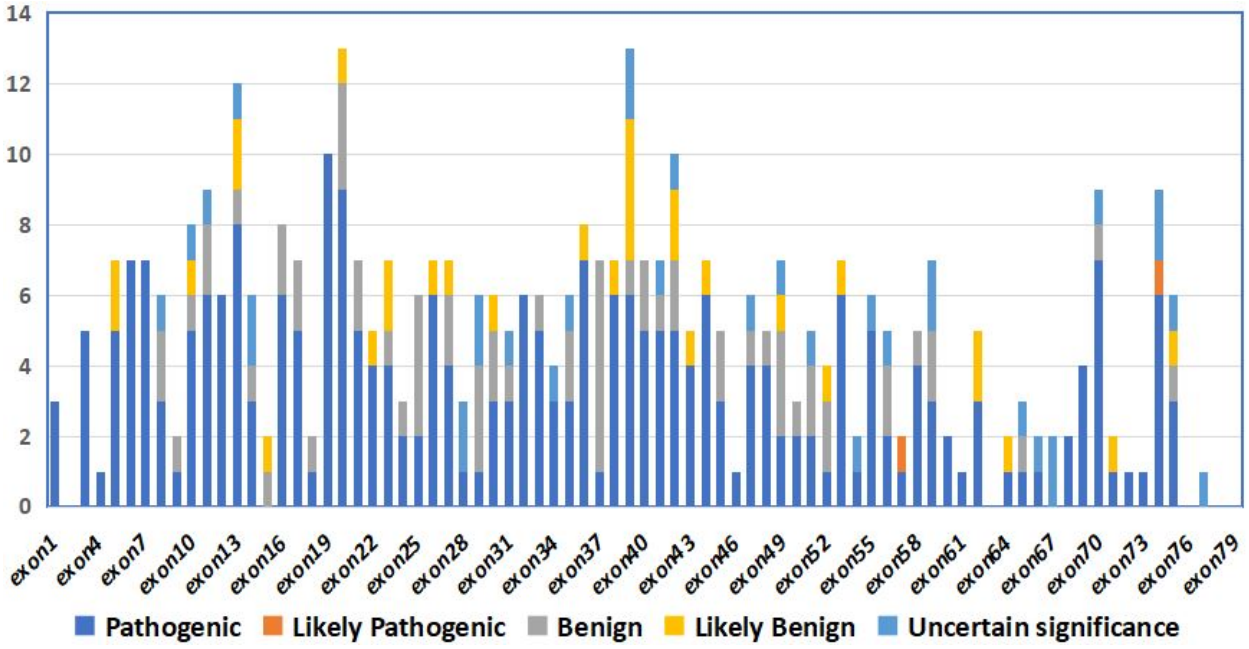


	stopgain	frameshift deletion	nonsynonymous SNV	frameshift insertion	synonymous SNV	frameshift substitution	nonframeshift deletion
Pathogenic	151	57	26	20	5	2	2
Likely pathogenic	1	1	0	0	0	0	0
Benign	0	0	47	0	24	0	0
Likely benign	0	0	16	0	16	0	0
Uncertain significance	1	0	35	0	0	0	0

Distribution of SNPs by exonic region and ACMG classification in the DMD gene presented in figure 9. In general, pathogenic variants were distributed across almost all exons. In many exons (i.e. exon 19), most

variants were pathogenic. However, in some other exons (i.e. exon 37), benign variants occurred more frequently. Exon 19 has most density of pathogenic SNP distribution and followed by exon 20.

Figure 9: Distribution of SNPs by exonic region and ACMG classification in the DMD gene



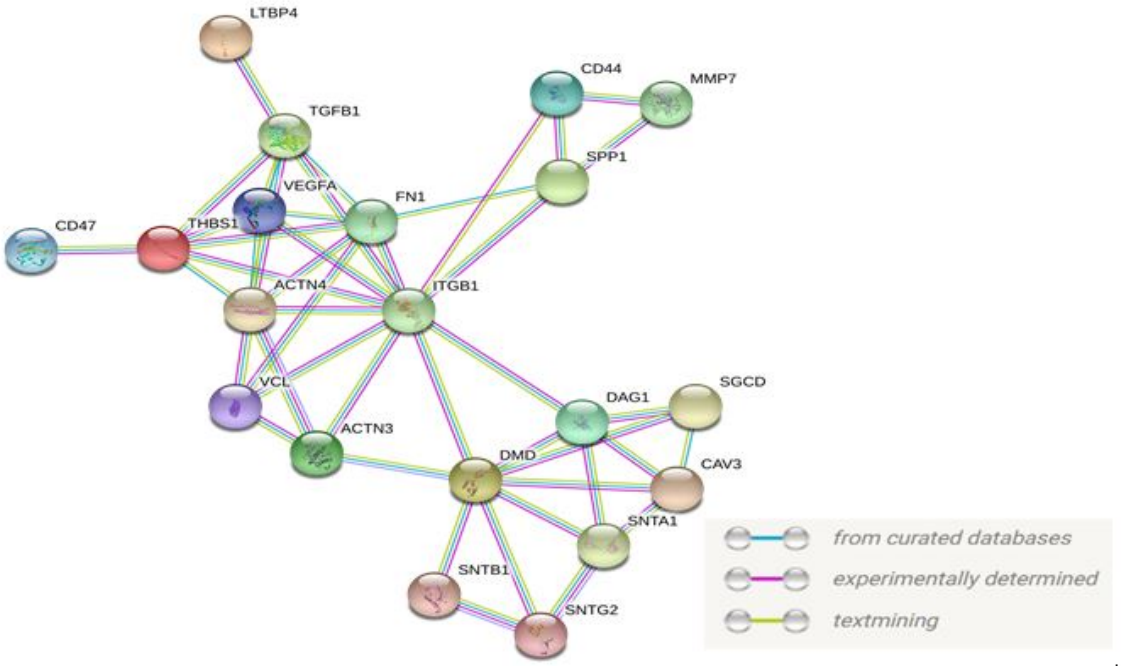
The x-axis only shows every third exon number due to space

3.8 Network analysis highlights non-random interconnectivity between the genetic modifiers identified in DMD patient

Genetic modifiers have been associated with variability in human DMD sub-phenotypes[13]: SPP1, LTBP4, and THBS1 are mapped to the TGF-β pathway, and implicated in several interconnected molecular pathways regulating inflammatory response to muscle damage, regeneration, and fibrosis. ACTN3 deficiency triggered an increase in oxidative muscle metabolism through activation of calcineurin, and then ameliorated the progression of dystrophic pathology. We considered SPP1, LTBP4, ACTN3 and THBS1 as the "seed" genes/proteins to construct the PPI network associated with DMD (Figure 10) [36]. Fifteen additional interactors were allowed in the network to identify the most significant interactions and achieve a meaningful size for network analysis. PPI analyses showed that ACTN3 had direct interaction with DMD. SPP1 may interact with DMD through ITGB1, which has the highest node degree and BC values in the network. ITGB1 is one of the most common forms in muscle. Disruptions of integrins are responsible for a further class of muscular dystrophies[37]. Among the "seed" genetic modifiers, THBS1 has higher network topological parameters, followed by SPP1, ACTN3 and LTBP4. The network enrichment p-value was < 3.09e-08, meaning that this connected network has significantly more interactions than expected at

random, and that the genetic modifiers have more interactions among themselves than what would be expected for a random set of genes/proteins of similar size. Such enrichment also indicates that these genetic modifiers are, at least partially, biologically connected. In addition, the PPI network may be possible to shed light on new genetic modifiers by their functional coupling to these known "seed" genes.

Figure 10. Interaction network resulting from the genetic modifiers identified in DMD patients



3.9 Exon DMD genetic alternations in different tumors

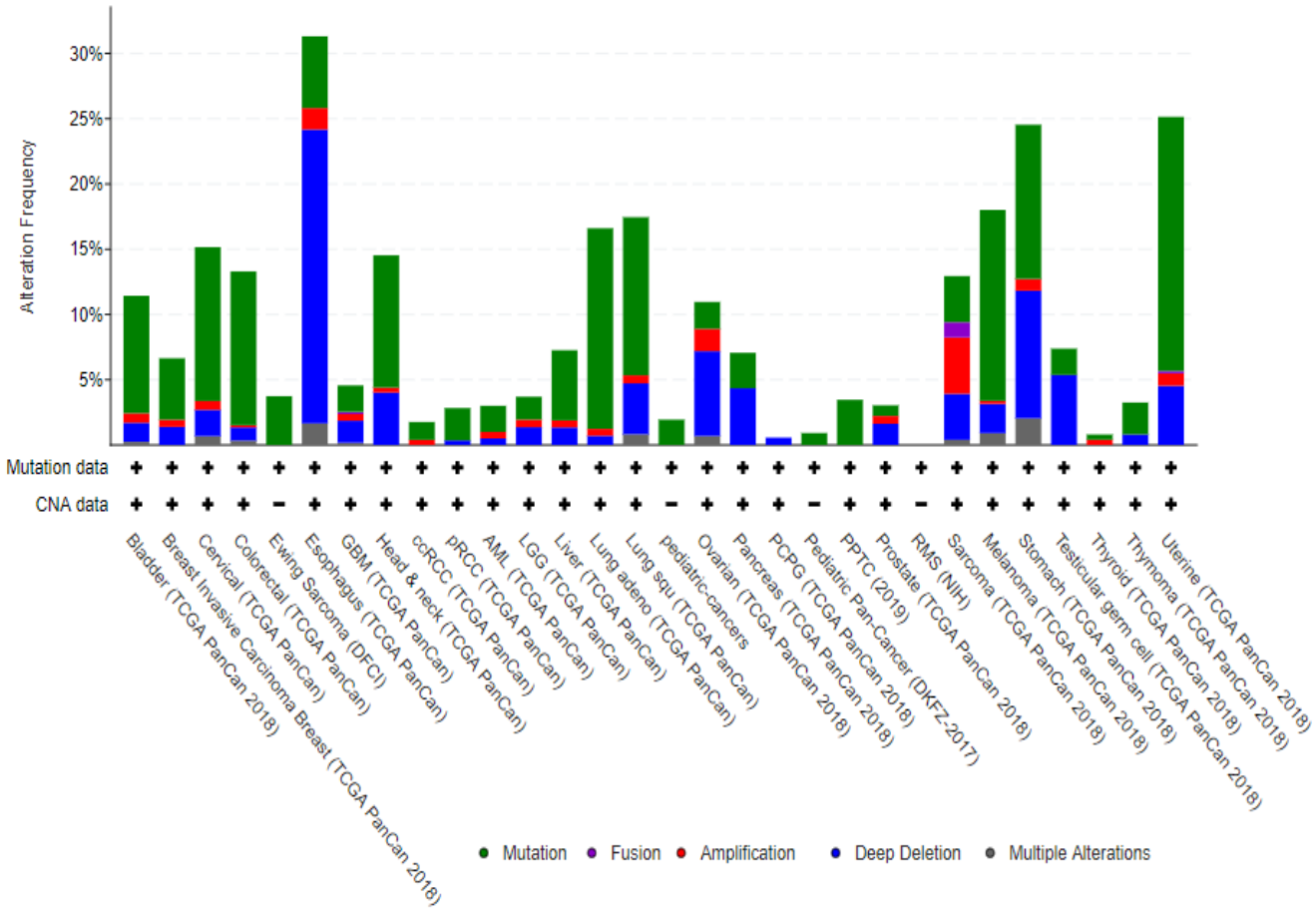
After examining SNPs distribution by ACMG Classification from DMD functional annotation output file, it appeared genetic alternations in the DMD gene associated with different types of cancers. It has also been reported that DMD gene is involved in tumor development and progression[38-41]. Therefore, we further analyzed data from 25 published cancer studies from The Cancer Genome Atlas (TCGA) and 4 pediatric cancer studies that included a minimum of 100 samples in the cBioPortal database. Since people with DMD could develop rare and aggressive type of muscle cancer (i.e.rhabdomyosarcoma), we also included one study that analyzed 43 rabdomyosarcomas cases. Total 11927 patients (age from ~ 3 years to 90 years; ~ 48% male and ~ 46% female; ~ 60% white, ~7% black or africa america and ~ 5% asian) and 11977 samples from 30 studies have been included in the analysis. Approximately 10% of the cases have an alteration in the DND gene, consisting mainly of missense mutation and truncating mutation. The studies analyzed were listed in the appendix 5.4

Table 5: Example of DMD gene mutations with cancers

Cancer Type	Functional Impact	Mutation Type	Variant Type	HGVSc	Exon
Adrenocortical Carcinoma	MutationAssessor: NA;SIFT: impact: tolerated, score: 0.05;Polyphen-2: impact: probably_damaging,	Missense_Mutation	SNP	ENST00000357033.	53/79
Adrenocortical Carcinoma	MutationAssessor: NA;SIFT: NA;Polyphen-2: NA	Frame_Shift_Ins	INS	ENST00000357033.	57/79
Mucinous Carcinoma	MutationAssessor: NA;SIFT: impact: tolerated, score: 1;Polyphen-2: impact: benign, score: 0.06	Missense_Mutation	SNP	ENST00000357033.	54/79
Cervical Squamous Cell Carcinoma	MutationAssessor: NA;SIFT: impact: deleterious, score: 0.03;Polyphen-2: impact: benign, score: 0.06	Missense_Mutation	SNP	ENST00000357033.	31/79
Cervical Squamous Cell Carcinoma	MutationAssessor: NA;SIFT: impact: deleterious, score: 0;Polyphen-2: impact: probably_damaging,	Missense_Mutation	SNP	ENST00000357033.	33/79
Colon Adenocarcinoma	MutationAssessor: NA;SIFT: impact: tolerated, score: 0.27;Polyphen-2: impact: benign, score: 0.005	Missense_Mutation	SNP	ENST00000357033.	37/79
Colon Adenocarcinoma	MutationAssessor: NA;SIFT: NA;Polyphen-2: NA	Frame_Shift_Del	DEL	ENST00000357033.	27/79
Mucinous Adenocarcinoma of the Colon an	MutationAssessor: NA;SIFT: NA;Polyphen-2: NA	Frame_Shift_Del	DEL	ENST00000357033.	27/79
Stomach Adenocarcinoma	MutationAssessor: NA;SIFT: NA;Polyphen-2: NA	Frame_Shift_Ins	INS	ENST00000357033.	70/79
Intestinal Type Stomach Adenocarcinoma	MutationAssessor: NA;SIFT: NA;Polyphen-2: NA	Nonsense_Mutation	SNP	ENST00000357033.	32/79
Stomach Adenocarcinoma	MutationAssessor: NA;SIFT: impact: deleterious, score: 0.01;Polyphen-2: impact: possibly_damagin	Missense_Mutation	SNP	ENST00000357033.	31/79
Stomach Adenocarcinoma	MutationAssessor: NA;SIFT: impact: deleterious, score: 0;Polyphen-2: impact: probably_damaging,	Missense_Mutation	SNP	ENST00000357033.	31/79
Tubular Stomach Adenocarcinoma	MutationAssessor: NA;SIFT: impact: tolerated, score: 0.08;Polyphen-2: impact: benign, score: 0.018	Missense_Mutation	SNP	ENST00000357033.	57/79
Papillary Stomach Adenocarcinoma	MutationAssessor: NA;SIFT: NA;Polyphen-2: NA	Frame_Shift_Del	DEL	ENST00000357033.	54/79
Signet Ring Cell Carcinoma of the Stomach	MutationAssessor: NA;SIFT: impact: deleterious, score: 0;Polyphen-2: impact: probably_damaging,	Missense_Mutation	SNP	ENST00000357033.	65/79

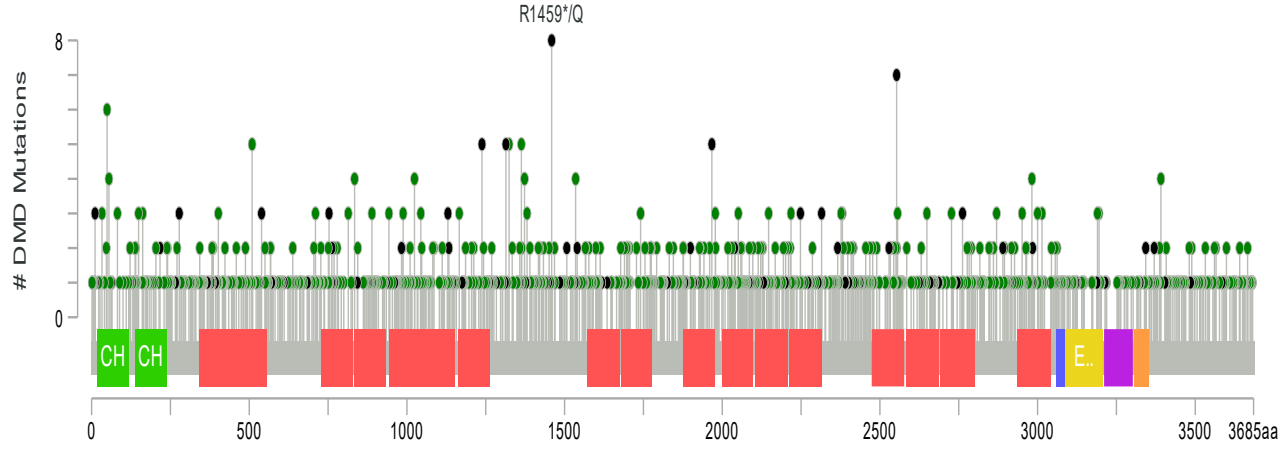
The majority of DMD genetic alterations corresponded to mutations and deep deletions, and a low frequency of gene amplifications. The occurrence of DMD alterations varied across the studies/tumor types (Figure 11). The highest ratio of genetic alternations was approximately 31% (Esophagus, TCGA PanCan), and followed by 25% (Uterine TCGA PanCan 2018). There were approximately 967 missense mutations and 246 truncating mutation in the tumor samples, with highest frequency (8 of nonsense mutation) being R1459*/Q in the kinase domain. There appeared to be no hot spots of alteration in the gene region (Figure 12). Interestingly, DMD alterations were not found in rabdomyosarcomas samples.

Figure 11: Frequency of genetic alternations in the DMD gene in different types of tumors



The x-axis shows the types of cancer (color coded), availability of mutation and copy number variation data, and the study abbreviation

Figure 12: Mutation diagram of DMD gene in the tumor samples

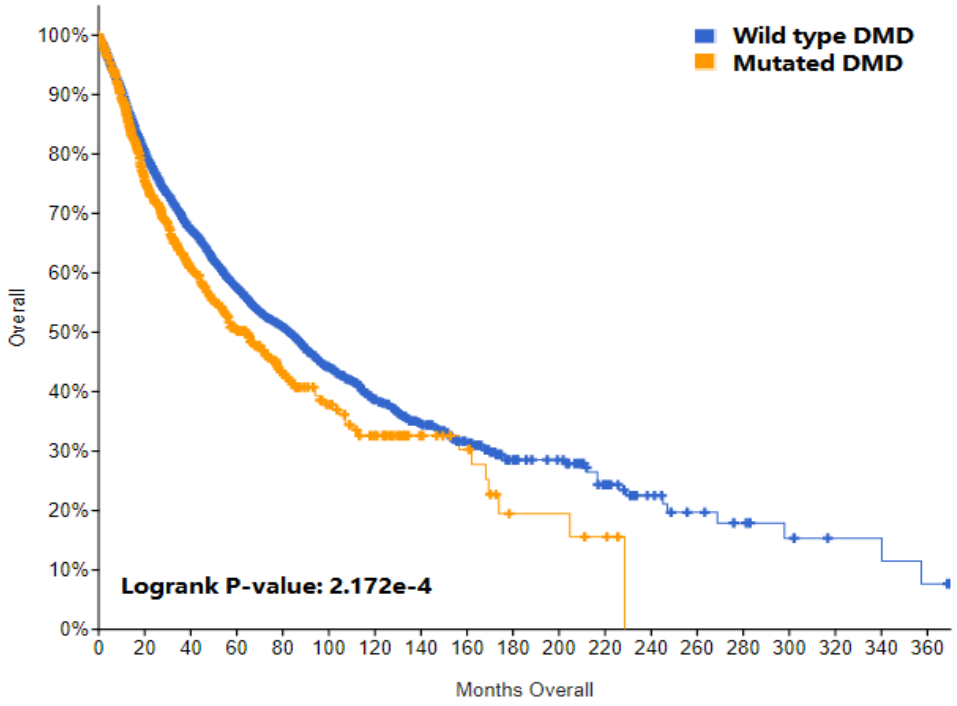


- ***Missense Mutations***
- ***Truncating Mutations****: Nonsense, Nonstop, Frameshift deletion, Frameshift insertion, Splice site*

3.10 Patients with DMD alterations have poorer overall survival

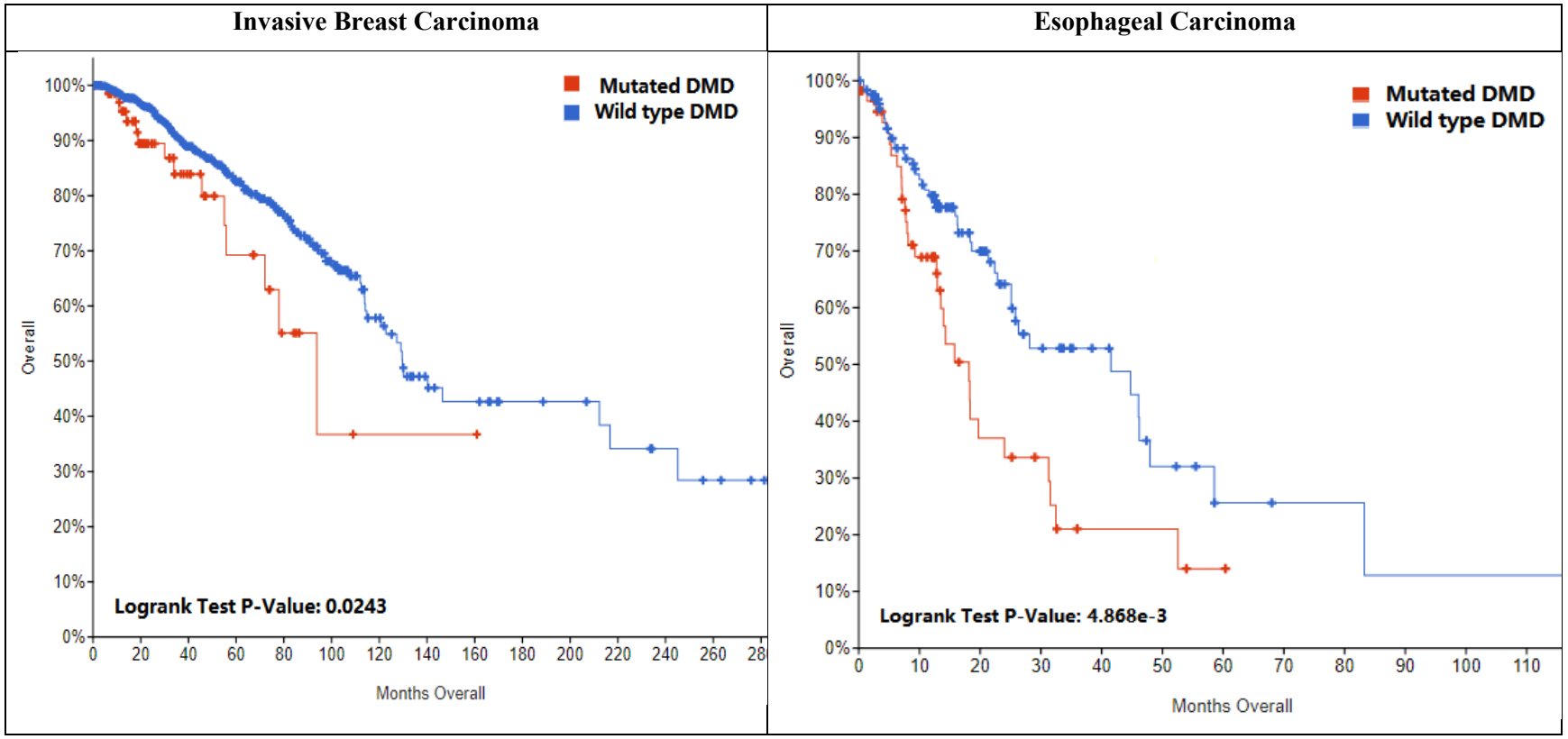
To study the overall survival (OS) of patients with and without DMD genetic alterations, Kaplan-Meier curves stratified by genotype were plotted and the comparisons were tested using the Log-rank test[5]. OS analysis has been initially conducted using cBioPortal with pooled all patient data from 30 different cancer studies as mentioned in previous section (Appendix 5). We observed that patients with genetic alterations in DMD had significantly shorter OS (63.8 months) compared to patients with wild-type DMD (82.9 months) (Figure 13).

Figure 13: Overall survival analyses using cBioPortal data.



Additional OS analysis has been performed by various tumor types with greater than 5% DMD genetic alteration frequency[5]. Invasive breast carcinoma and esophageal carcinoma that revealed significant differences between by wild-type DMD or mutated DMD groups (Figure 14). Others such as ovarian carcinomas showed similar trend, but not statistically significant (Figures not presented). OS analysis of breast cancer patients with low or high expression of DMD gene has been further conducted by using Kaplan-Meier Plotter (www.kmplot.com/analysis) [42]. Likewise, low DMD expression is associated with poorer survival in breast cancer patients (Figures not presented). These results demonstrate that the relationship between DMD genetic status and prognosis may be tumor-type specific. Moreover, its biological function in tumorigenesis as well as prognosis is complicated and co-regulated with other factors.

Figure 14: Overall survival for invasive breast carcinoma and esophageal carcinoma using cBioPortal data.



4. Discussions and Conclusions

DMD is a rare, severe, progressive genetic disorder causing disability and premature death. DMD caused by mutations in the dystrophin-encoding DMD gene, which is one of the largest of the identified human genes. There are currently no curative therapies for DMD. Gene therapy is a promising experimental method that uses genes (the fundamental units of heredity) to treat disorders that result from genetic mutations. Currently, several kinds of gene-based therapies including exon skipping are being developed to treat DMD.

Although mutations in DMD genes have been widely studied, to our knowledge, a systematic genetic analysis of all variants, especially SNPs, of the gene in human have not been reported. In general, Variants in exonic (coding) region can alter the protein function. Therefore, we focused on a total of 3627 exonic SNPs. In the DMD gene, SNPs distributed across all exons. The largest category was nonsynonymous account for nearly 64% of all mutations. Exon 19 appeared to be most pathogenic region. As expected, nonsense mutation (i.e. stopgain) or frameshift mutation likely lead to more pathogenic DMD. Of note, the limitation of current database may not perfectly represent the DMD patient population, since the database collects all genetic variations from various individuals, but not merely DMD patients. Therefore, the result in the study reveals richer variations in human populations than in DMD patients. Interestingly, some pathogenetic variants were also observed in healthy individuals. For example, a few nonsynonymous and stop gain cases were observed in the 1000 Genomes project, although the allele frequency was very low. Healthy individuals may also carry DMD related genetic alternations. However, for each gene, there are homologous alleles from the mother and father, so the abnormality of one of the alleles will not lead to disease. This may explain why normal individuals carry DMD mutations without clinic symptoms.

Phenotypic variations in DMD may also occur in patients with the same primary mutation. A wide range of clinical manifestations suggest that genetic modifiers such as SPP1, LTBP4, ACTN3, and THBS1 can modify the clinical severity of DMD disease [13]. As expected, our PPI analysis highlighted non-random interconnectivity between the genetic modifiers identified in DMD patients, and potentially shed light on new genetic modifiers by their functional coupling to these known genes.

The genomic, functional and clinicopathological evidence demonstrate dystrophin tumor suppressor roles in human cancers. People with DMD may increase risks to develop aggressive and rare muscle cancer such as rhabdomyosarcoma[38, 43]. In the current study, we have used the cBioPortal for cancer genomics as a tool for visualizing, exploring and analyzing the biological and clinical characteristics of DMD genetic alterations with cancer. Total 11927 patients and 11977 samples from 30 studies with various types of

tumors have been included in the analysis. Approximately 10% of the cases had an alteration in the DMD gene, consisting mainly of missense mutation and truncating mutation. We observed DMD genetic alterations varied across the studies/tumor types. Patients with DMD genetic alterations appeared to have shorter overall survival, particularly with esophageal carcinoma and invasive breast carcinoma. Similarly, Stephens et al. also reported poorer survival for patients with upper gastrointestinal cancer and low expression of DMD[39]. However, there were no significant correlations between DMD genetic alterations and overall survival in some other types of human cancers. These results demonstrate that the relationship between DMD genetic status and prognosis may be tumor-type specific. Moreover, its biological function in tumorigenesis as well as prognosis is complicated and co-regulated with other factors.

Leonela N. Luce et al conducted paired tumor/normal tissues showed that majority of tumor specimens had lower DMD expression compared to the normal adjacent tissue[44], in concordance to the other reports suggesting the tumor suppressor role of DMD[38]. Contrastingly, overexpression of the DMD gene has also been reported in leukemias, renal carcinomas, ependymomas, and astrocytomas [44]. The function and molecular mechanism involved in altered DMD gene expression in different cancer types remains to be further investigated.

In conclusion, to our knowledge, this is first data mining study with a systematic genetic analysis of all variants, especially SNPs, of the DMD gene in human. SNPs distributed across all Exons. The largest category was nonsynonymous account for nearly 64% of all mutations. Exon 19 appeared to have most density of pathogenic SNP distribution. Nonsense mutation (i.e. stopgain) or frameshift mutation likely lead to more pathogenic. Network analysis highlighted non-random interconnectivity between the genetic modifiers identified in DMD patients, and potentially shed light on new genetic modifiers by their functional coupling to these known genes. In addition, our results also suggest DMD gene may serve as a diagnostic and therapeutic target for certain types of cancer.

5. Appendix

5.1 R Script of "Extract the longest length transcript with Exon Number, amino acid change from DMD gene exome file"

```
# String work
library(stringr)
library(dplyr)
# Read Path where your CSV file is located
DMD<- read.csv(file = 'Read Path where CSV file
located\\DMD_query.output.exome_summary.csv')
head(DMD)
library(stringr)
location<-str_locate(DMD$AAChange.ensGene,"ENST00000357033")
location
startpos<-location[,1]  startpos
endpos<-location[,2] endpos
str_sub(DMD$AAChange.ensGene,startpos,endpos+25)
# Export ENST00000357033 segment CSV
DMD_ENST_Output<-str_sub(DMD$AAChange.ensGene,startpos,endpos+25)
write.csv(DMD_ENST_putput, file = "DMD_ENST_Output.csv")
```

5.2 R Script for Heatmap of DMD variants frequency from different population in 1000 genomes project

```
# Read DMD variants frequency VS 1000 gene population file from ENSG0000019894
DMDgeneExp <- read.csv'Read Path where CSV file located\\heatmap\\1000genomeproject.csv")

install.packages('caTools')
library(gplots)

head(DMDgeneExp )
rownames(DMDgeneExp ) <- make.unique(as.character(DMDgeneExp $Variants.Type))

DMDgeneExp_matrix <- as.matrix(DMDgeneExp [2:6])
head(DMDgeneExp_matrix)

dist_no_na <- function(mat) {
  edist <- dist(mat)
  edist[which(is.na(edist))] <- max(edist, na.rm=TRUE) * 1.1
  return(edist)
}

colors = c(seq(-3,-2,length=100),
           seq(-2,0.5,length=100),
           seq(0.5,6,length=100))
my_palette <- colorRampPalette(c("grey","blue","darkred"))(n = 100)

# Heatmap plotted
heatmap.2(DMDgeneExp_matrix,distfun=dist_no_na,col=my_palette,
          key=TRUE, symkey=FALSE,
          cexRow=0.01,cexCol=1.2,
          offsetRow = 0,
          densadj = 0.15,
          margins=c(4,16),
          lwid = c(5,20))
```

5.3: Amino acid descriptions

One letter code	Three letter code	Amino acid	Possible codons
A	Ala	Alanine	GCA, GCC, GCG, GCT
B	Asx	Asparagine or Aspartic acid	AAC, AAT, GAC, GAT
C	Cys	Cysteine	TGC, TGT
D	Asp	Aspartic acid	GAC, GAT
E	Glu	Glutamic acid	GAA, GAG
F	Phe	Phenylalanine	TTC, TTT
G	Gly	Glycine	GGA, GGC, GGG, GGT
H	His	Histidine	CAC, CAT
I	Ile	Isoleucine	ATA, ATC, ATT
K	Lys	Lysine	AAA, AAG
L	Leu	Leucine	CTA, CTC, CTG, CTT, TTA, TTG
M	Met	Methionine	ATG
N	Asn	Asparagine	AAC, AAT
P	Pro	Proline	CCA, CCC, CCG, CCT
Q	Gln	Glutamine	CAA, CAG
R	Arg	Arginine	AGA, AGG, CGA, CGC, CGG, CGT
S	Ser	Serine	AGC, AGT, TCA, TCC, TCG, TCT
T	Thr	Threonine	ACA, ACC, ACG, ACT
V	Val	Valine	GTA, GTC, GTG, GTT
W	Trp	Tryptophan	TGG
X	X	any codon	NNN
Y	Tyr	Tyrosine	TAC, TAT
Z	Glx	Glutamine or Glutamic acid	CAA, CAG, GAA, GAG
*	*	stop codon	TAA, TAG, TGA

5.4 Cancer studies analyzed from cBioPortal

Type of cancer – study abbreviation	Number
Bladder Urothelial Carcinoma (TCGA, PanCancer Atlas)	411
Colorectal Adenocarcinoma (TCGA, PanCancer Atlas)	594
Breast Invasive Carcinoma (TCGA, PanCancer Atlas)	1084
Brain Lower Grade Glioma (TCGA, PanCancer Atlas)	514
Glioblastoma Multiforme (TCGA, PanCancer Atlas)	592
Cervical Squamous Cell Carcinoma (TCGA, PanCancer Atlas)	297
Esophageal Adenocarcinoma (TCGA, PanCancer Atlas)	182
Stomach Adenocarcinoma (TCGA, PanCancer Atlas)	440
Head and Neck Squamous Cell Carcinoma (TCGA, PanCancer Atlas)	523
Kidney Renal Clear Cell Carcinoma (TCGA, PanCancer Atlas)	512
Kidney Renal Papillary Cell Carcinoma (TCGA, PanCancer Atlas)	283
Liver Hepatocellular Carcinoma (TCGA, PanCancer Atlas)	372
Lung Adenocarcinoma (TCGA, PanCancer Atlas)	566
Lung Squamous Cell Carcinoma (TCGA, PanCancer Atlas)	487
Acute Myeloid Leukemia (TCGA, PanCancer Atlas)	200
Ovarian Serous Cystadenocarcinoma (TCGA, PanCancer Atlas)	585
Pancreatic Adenocarcinoma (TCGA, PanCancer Atlas)	184
Prostate Adenocarcinoma (TCGA, PanCancer Atlas)	494
Skin Cutaneous Melanoma (TCGA, PanCancer Atlas)	448
Pheochromocytoma and Paraganglioma (TCGA, PanCancer Atlas)	178
Sarcoma (TCGA, PanCancer Atlas)	255
Testicular Germ Cell Tumors (TCGA, PanCancer Atlas)	149
Thymoma (TCGA, PanCancer Atlas)	123
Thyroid Carcinoma (TCGA, PanCancer Atlas)	500
Uterine Corpus Endometrial Carcinoma (TCGA, PanCancer Atlas)	529
Pediatric Pan-Cancer (DKFZ, Nature 2017)	961
Pediatric Ewing Sarcoma (DFCI, Cancer Discov 2014)	107
Pediatric Preclinical Testing Consortium (CHOP, Cell Rep 2019)	261
Pediatric Pan-cancer (Columbia U, Genome Med 2016)	103
Rhabdomyosarcoma (NIH, Cancer Discov 2014)	43

Acknowledgments

I would like to thank my research project mentors, Dr. Pingzhang Wang and Mrs. Betty Wang, for suggesting the research topic, and for providing resources and thorough guidance. Their guidance helped me navigate this paper and allowed me to overcome difficulties in the researching and writing process. They suggested me to get the source data online and discussed with me to analyze and annotate the important gene data into a readable form which was crucial for the paper. They also gave me hint to search various different websites and programs like wANNOVAR, PubMed, NCBI, BioGps, GtexPortal, Ensemble, COSMIC, cBioPortal, RStudio, and EndNote which were the important programs that I utilized to obtain the data, analyze results.

References

1. Mohammed, F., et al., *Mutation spectrum analysis of Duchenne/Becker muscular dystrophy in 68 families in Kuwait: The era of personalized medicine.* PLoS One, 2018. **13**(5): p. e0197205.
2. Gadalla, S.M., et al., *Cancer risk among patients with myotonic muscular dystrophy.* Jama, 2011. **306**(22): p. 2480-6.
3. Win, A.K., et al., *Increased cancer risks in myotonic dystrophy.* Mayo Clin Proc, 2012. **87**(2): p. 130-5.
4. Lab, W.G. *wANNOVAR*. 2010-2020; Available from: http://wannovar.wglab.org/.
5. *cBioPortal for Cancer Genomics*. 2020; Available from: https://www.cbioportal.org/.
6. Nelson, S.F., et al., *Emerging genetic therapies to treat Duchenne muscular dystrophy.* Current opinion in neurology, 2009. **22**(5): p. 532.
7. Salmaninejad, A., et al., *Duchenne muscular dystrophy: an updated review of common available therapies.* Int J Neurosci, 2018. **128**(9): p. 854-864.
8. Wang, R.T. and S.F. Nelson, *What can Duchenne Connect teach us about treating Duchenne muscular dystrophy?* Curr Opin Neurol, 2015. **28**(5): p. 535-41.
9. BioIncept, L. *Potential to effectively protect muscle function, combatting Duchenne muscular dystrophy*. 2020; Available from: https://bioincept.com/product-development/duchenne-muscular-dystrophy/.
10. VanBelzen, D.J., et al., *Mechanism of Deletion Removing All Dystrophin Exons in a Canine Model for DMD Implicates Concerted Evolution of X Chromosome Pseudogenes.* Mol Ther Methods Clin Dev, 2017. **4**: p. 62-71.
11. Lee, B.L., et al., *Genetic analysis of dystrophin gene for affected male and female carriers with Duchenne/Becker muscular dystrophy in Korea.* J Korean Med Sci, 2012. **27**(3): p. 274-80.

12. Gao, Q.Q. and E.M. McNally, *The Dystrophin Complex: Structure, Function, and Implications for Therapy.* Compr Physiol, 2015. **5**(3): p. 1223-39.
13. Bello L, Pegoraro E. *The "Usual Suspects": Genes for Inflammation, Fibrosis, Regeneration, and Muscle Strength Modify Duchenne Muscular Dystrophy*. J Clin Med. 2019 May 10;8(5):649.
14. Porter, G.A., et al., *Dystrophin colocalizes with beta-spectrin in distinct subsarcolemmal domains in mammalian skeletal muscle.* J Cell Biol, 1992. **117**(5): p. 997-1005.
15. Norwood, F.L., et al., *The structure of the N-terminal actin-binding domain of human dystrophin and how mutations in this domain may cause Duchenne or Becker muscular dystrophy.* Structure, 2000. **8**(5): p. 481-91.
16. Muthu, M., K.A. Richardson, and A.J. Sutherland-Smith, *The crystal structures of dystrophin and utrophin spectrin repeats: implications for domain boundaries.* PLoS One, 2012. **7**(7): p. e40066.
17. Broderick, M.J. and S.J. Winder, *Spectrin, alpha-actinin, and dystrophin.* Adv Protein Chem, 2005. **70**: p. 203-46.
18. Henry, M.D. and K.P. Campbell, *Dystroglycan: an extracellular matrix receptor linked to the cytoskeleton.* Curr Opin Cell Biol, 1996. **8**(5): p. 625-31.
19. Blake, D.J., J.M. Tinsley, and K.E. Davies, *Utrophin: a structural and functional comparison to dystrophin.* Brain Pathol, 1996. **6**(1): p. 37-47.
20. Heikoop, J.C., Hogervorst, F.B.L., Meershoek, E.J. *et al. Expression of the Human Dp 71 (Apo-Dystrophin-1) Gene from a 760-kb DMD-YAC Transferred to Mouse Cells*. Eur J Hum Genet **3,** 168–179 (1995).
21. D'Souza, V N et al. *A novel dystrophin isoform is required for normal retinal electrophysiology.* Human molecular genetics vol. 4,5 (1995): 837-42.
22. Taylor, P.J., et al., *Dystrophin gene mutation location and the risk of cognitive impairment in Duchenne muscular dystrophy.* PLoS One, 2010. **5**(1): p. e8803.
23. Lederfein, D. et al*., A 71-kilodalton protein is a major product of the Duchenne muscular dystrophy gene in brain and other nonmuscle tissues*. Proceedings of the National Academy of Sciences of the United States of America vol. 89,12 (1992): 5346-50.
24. Yates, A., et al., *Ensembl 2016.* Nucleic Acids Res, 2016. **44**(D1): p. D710-6.
25. *dbSNP*. 2020; Available from: https://www.ncbi.nlm.nih.gov/snp/.
26. Chang, X. and K. Wang, *wANNOVAR: annotating genetic variants for personal genomes via the web.* J Med Genet, 2012. **49**(7): p. 433-6.
27. *Genomics Lab*. 2010-2020; Available from: http://wannovar.wglab.org/.

28. Institute, E.M.B.L.s.E.B. *DMD ENSG00000198947 Transcripts*. August 2020; Available from: http://Aug2020.archive.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000198947;r=X:31097677-33339441.
29. Szklarczyk, Damian et al. *STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets*. Nucleic acids research vol. 47, D1 (2019): D607-D613.
30. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T Genome Res. 2003 Nov; 13(11):2498-504.*
31. *ANNOVAR Documentation*. Utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes, wANNOVAR supports only human genome annotation 2010-2018; Available from: https://doc-openbio.readthedocs.io/projects/annovar/en/latest/.
32. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.
33. Information, N.C.f.B. *Clinical significance on ClinVar submitted records*. 2020-03; Available from: https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/.
34. Belsare, S., et al., *Evaluating the quality of the 1000 genomes project data.* BMC Genomics, 2019. **20**(1): p. 620.
35. *New ACMG Guidelines*. September 30, 2015; Available from: https://www.genedx.com/whats-new/new-acmg-guidelines/.
36. Garton, Fleur C et al. *The Effect of ACTN3 Gene Doping on Skeletal Muscle Performance*. American journal of human genetics vol. 102,5 (2018): 845-857.
37. Smith, Lucas R et al. *Systems analysis of biological networks in skeletal muscle function.* Wiley interdisciplinary reviews. Systems biology and medicine vol. 5,1 (2013): 55-71.
38. Wang, Y., et al., *Dystrophin is a tumor suppressor in human cancers with myogenic programs.* Nat Genet, 2014. **46**(6): p. 601-6.
39. Sgambato, A., et al., *Dystroglycan expression is frequently reduced in human breast and colon cancers and is associated with tumor progression.* Am J Pathol, 2003. **162**(3): p. 849-60.
40. Hosur, V., et al., *Dystrophin and dysferlin double mutant mice: a novel model for rhabdomyosarcoma.* Cancer Genet, 2012. **205**(5): p. 232-41.
41. Körner, H., et al., *Digital karyotyping reveals frequent inactivation of the dystrophin/DMD gene in malignant melanoma.* Cell Cycle, 2007. **6**(2): p. 189-98.
42. *Kaplan-Meier Plotter*. 2009-2020; Available from: www.kmplot.com/analysis.

43. Boscolo Sesillo, F., D. Fox, and A. Sacco, *Muscle Stem Cells Give Rise to Rhabdomyosarcomas in a Severe Mouse Model of Duchenne Muscular Dystrophy.* Cell reports, 2019. **26**(3): p. 689-701.e6.
44. Luce, L.N., et al., *Non-myogenic tumors display altered expression of dystrophin (DMD) and a high frequency of genetic alterations.* Oncotarget, 2017. **8**(1): p. 145-155.
45. Rania Horaitis, N.b.J.T.D.D. *Codons and amino acids*. October, 2009; Available from: https://www.hgvs.org/mutnomen/codon.html.