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

Hubert Chen

West Windsor-Plainsboro South High School, NJ

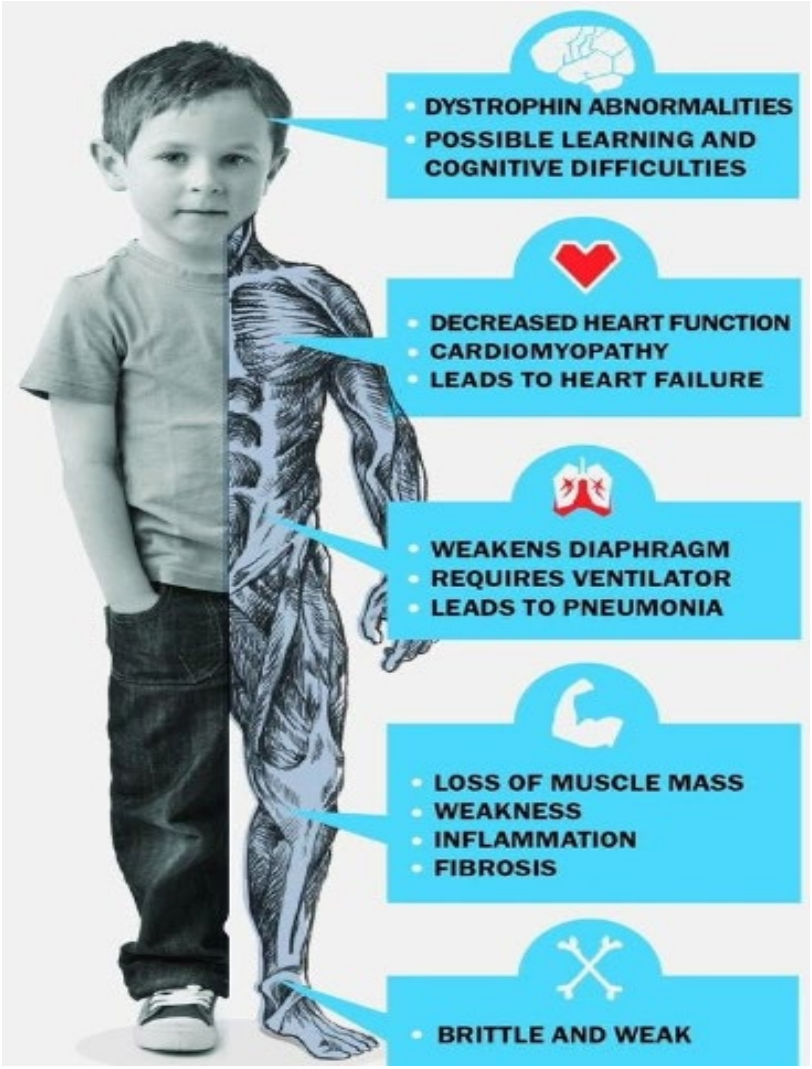
Mentor: Pingzhang Wang, Betty Wang

3/2/2020

Duchenne muscular dystrophy (DMD) **disease overview**

- DMD is a rare, severe, progressive genetic disorder causing disability and premature death
- Mutations in DMD gene, encoding dystrophin protein, lead to DMD.
- DMD primarily affects boys. The prevalence is approximately 1 in 3500 to 5000 male births worldwide.
- DMD symptom onset usually between ages 3 and 5 years.
- Phenotypic variations in DMD may also occur in patients with same primary mutation due to secondary genetic modifiers.

Figure 1: Duchenne's Impact on the Body

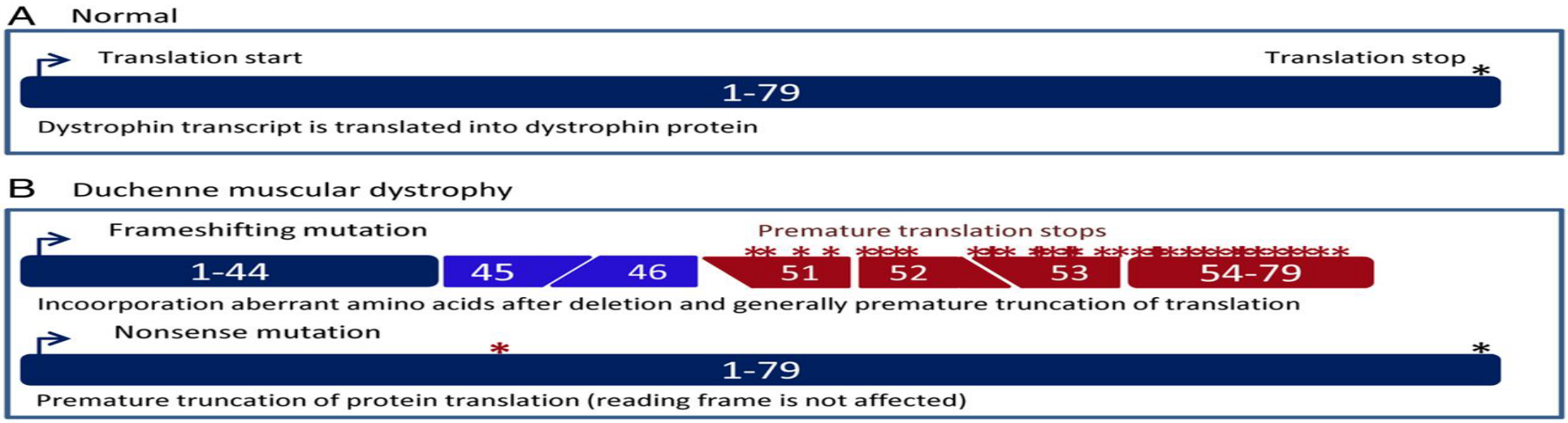


Reference source:CureDuchenne

Duchenne muscular dystrophy (DMD) **gene overview**

- DMD: One of **the largest known human gene**, spanning 2.4 Mb genomic sequence.
- DMD gene consists: 79 exons encoding a 14,000 bp messenger RNA transcript.
- DMD patients: Protein translation is stopped prematurely.
  - ➢ Frame-shifting mutations (e.g. deletion of exons 47–50, Figure 2-A, top panel)
    - Lead to inclusion of aberrant amino acids
    - Generally premature truncation of translation.
  - ➢ Alternatively, a point mutation (nonsense mutation)
    - Can change an amino acid codon into a stop codon (Figure 2-B, bottom panel)

Figure 2: Schematic depiction of dystrophin transcripts in healthy and DMD

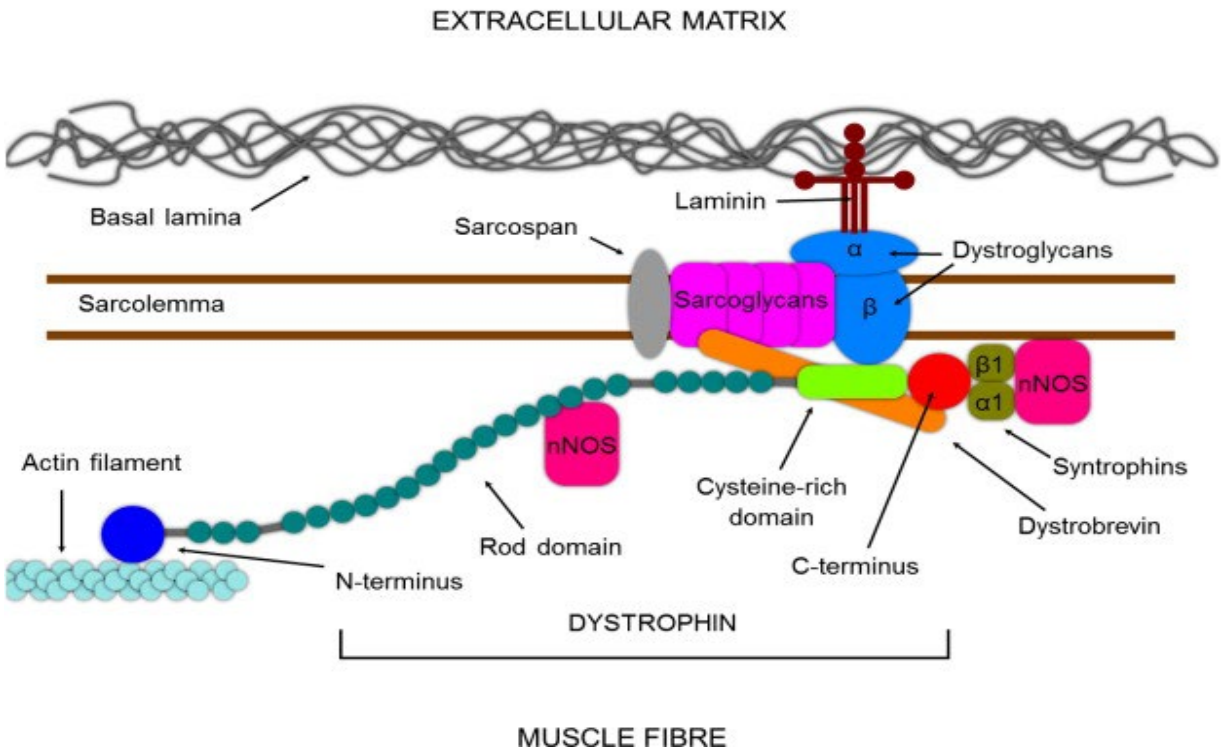


Reference source: Annemieke Aartsma-Rus et al, J Med Genet, 2016 Mar, 53(3):145-51

Dystrophin protein and dystroglycan complex overview

- Full length of dystrophin protein consists of 3,685 amino acids with 427 kDa.
- It consists of four major functional domains: actin-binding Nterminal domain (encoded by exon 1-8), central rod domain (encoded by exon 8-61), cysteine-rich domain (encoded by exon 62-69) and C-terminal domain (encoded by exon 69-79).
- Cysteine-rich domain together with C-terminal domain interact with different proteins including β-dystroglycan, syntrophin and dystrobrevin to make up dystrophin-glycoprotein complex.

Figure 3: Dystrophin and dystrophin-associated glycoprotein complex

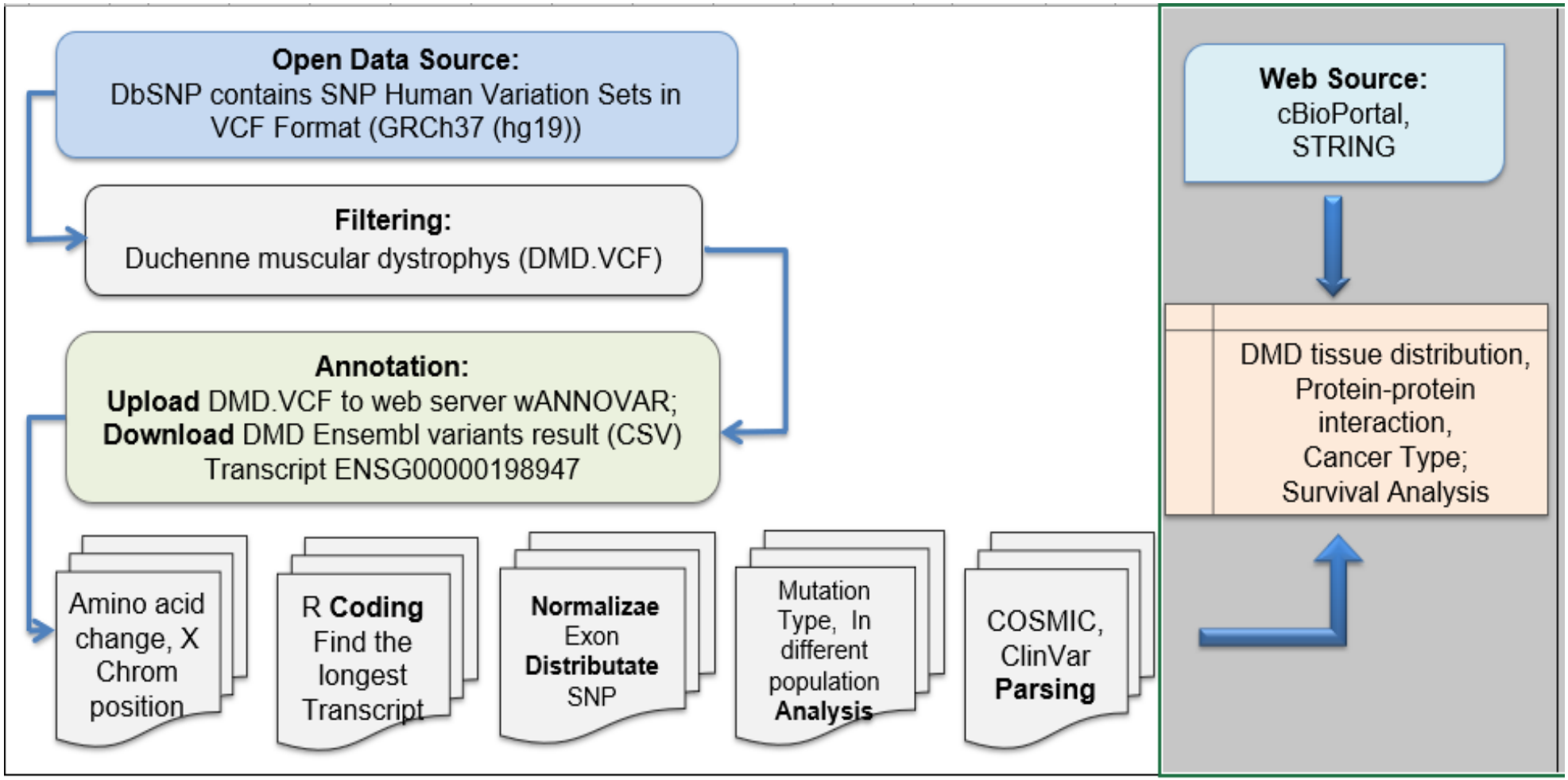


Reference source: ResearchGate.net

Research objectives and workflows

- Carry out a systematic analysis of the DMD genetic variants via dbSNP database
- Explore protein-protein interactions for genetic modifiers identified in DMD patients
- Investigate potential relationships of genetic alternations in the DMD gene with cancer

Figure 4: Research workflows

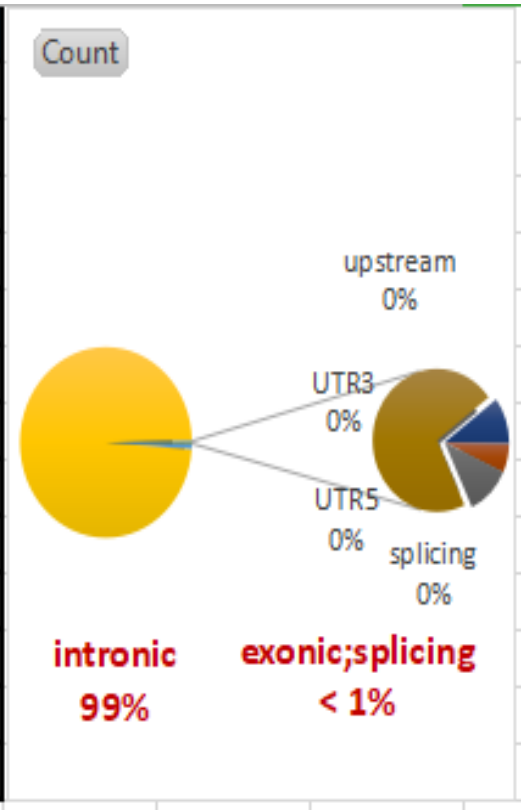


Research materials and methods

- Extract DMD genetic variants via dbSNP Database with variant call format (VCF)
- Functional annotation with wANNOVAR: Variant prioritization (Figure 5)
- Retrieve the longest transcript ENST0000035703 (Figure 6)

Figure 5: Genomic catalog in 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*



- Focus on variants in exonic (coding) region can alter the protein function

Figure 6: R coding for the longest transcript

	Amino Acid Change info in Transcript
Exonic	ENSG00000198947:ENST00000378705:exon2:c.106_113del:p.R36Gfs*2,ENSG00000198947:ENST0000361471:exon6:c.532_539del:p.R178Gfs*1,ENSG00000198947:ENST00000378702:exon6:c.532_539del:p.R178Gfs*,ENSG00000198947:ENST0000357033:exon67:c.9736_9743del:p.R3246Gfs*1,ENSG00000198947:ENST00000378677:exon67:c.9724_9731del:p.R3242Gfs*1

```
42      DMD_query.output.exome_summary.csv")
43  head(DMD)
44  library(stringr)
45  location<-str_locate(DMD$AAChange.ensGene,
46                       "ENST00000357033")
47  location
48  startpos<-location[,1] startpos
49  endpos<-location[,2] endpos
50  str_sub(DMD$AAChange.ensGene,startpos,endp
51  DMD_ENST_Output<-str_sub
52  (DMD$AAChange.ensGene,startpos, endpos+25)
53  write.csv(DMD_ENST_putput,
54            file = "DMD_ENST_Output.csv")
```

Research materials and methods

- Protein–protein interactions (**PPI**) map for genetic modifiers identified in DMD patients was constructed using STRING v11. Subsequently analyzed using Cytoscape 3.8.1 plugin Network Analyzer.

- Genetic alternations in the DMD gene with cancer was examined by using cBioPortal.

  - Data from 25 published TCGA cancer studies and 4 pediatric cancer studies that included a minimum of 100 samples. One study that reported 43 rabdomyosarcomas cases has also been included.

  - 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).

  - Kaplan-Meier curves were stratified by genotype and comparisons were tested using the Log-rank test.

Variants type and frequency of amnio acid change in the DMD gene

Table 1: Examples of DMD gene mutation

frameshift deletion		frameshift insertion		frameshift substitution		nonframeshift deletion		nonframeshift insertion		nonframeshift substitution		stopgain		nonsynonymous		synonymous	
Ref	Alt	Ref	Alt	Ref	Alt	Ref	Alt	Ref	Alt	Ref	Alt	Ref	Alt	Ref	Alt	Ref	Alt
A	-	-	A	GGT	AAAC	CTT	-	-	TTATACGG	GCC	AAT	A	T	A	C	G	T
AAAGACTTC	-	-	AAAC	TCCAAAG	CC	ACTGAT	-	-	TGA			G	A	T	C	T	C
AACGGGACT	-	-	ACCATGTGAG	TT	A	AGG	-	-	ATC			G	C	G	A	C	T
AACTGTCT	-	-	AGAC			GTT	-	-	ACA			-	TTAC	A	G	G	A
AG	-	-	AT			AGA	-					T	A	A	C	T	C
ATAA	-	-	C			GGACGA	-					C	T	T	C	T	G

** *Ref: Original nucleotide(s) present before mutation*
Alt: Alternative nucleotide(s) present after mutation

Figure 7: Frequency of amnio acid change in DMD gene

- Insertion, deletion, substitution that cause frameshift changes in protein coding sequence.
- The largest category: nonsynonymous, follow by synonymous and stop gain.

Vaniants Type	Count	Frequency
frameshift deletion	116	3.20%
frameshift insertion	47	1.30%
frameshift substitution	3	0.08%
nonframeshift deletion	27	0.74%
nonframeshift insertion	4	0.11%
nonframeshift substitution	1	0.03%
nonsynonymous SNV	2322	64.02%
startloss	1	0.03%
stopgain	242	6.67%
synonymous SNV	864	23.82%
Total unique samples	3627	

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	

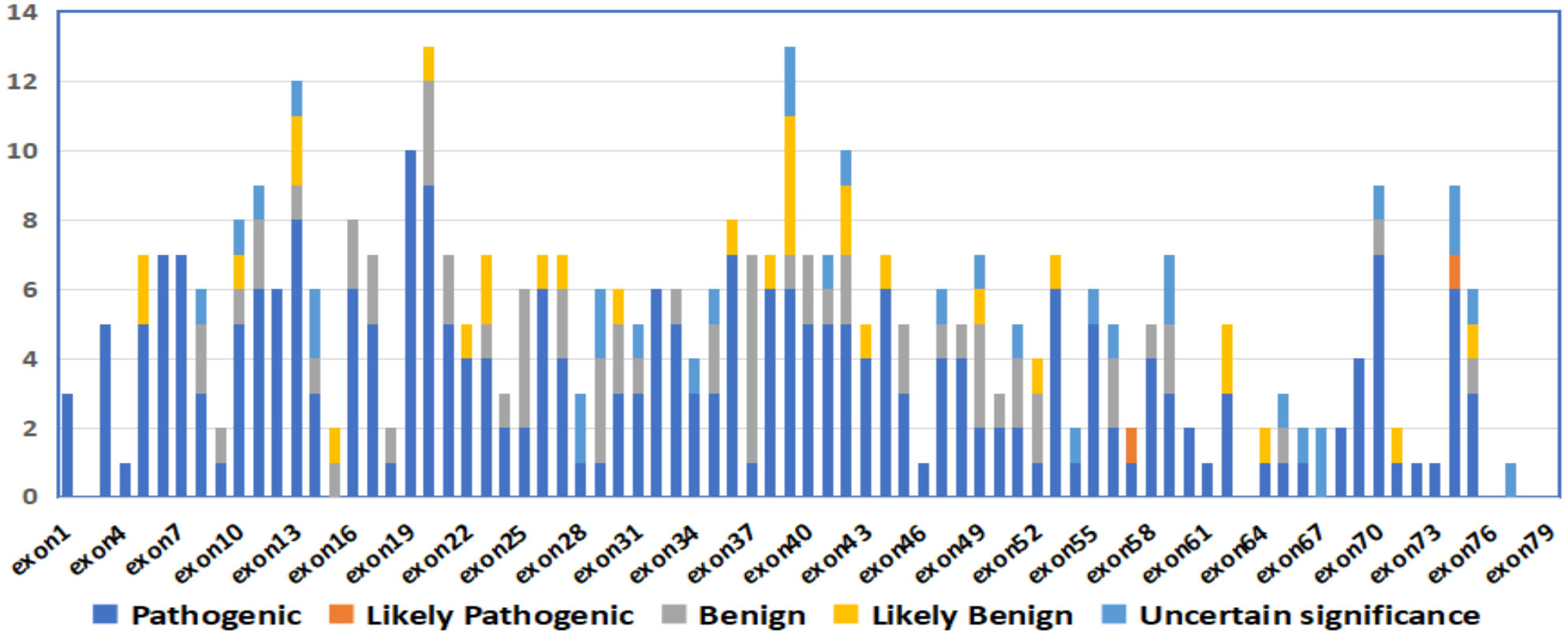
Distribution of SNPs by exonic region

- SNPs distributed across almost all exons. Exon 79 is the longest with 2703 bp in length. Exon 78 is the shortest with 32 bp
- **Normalized** exon length, then Exon 19 has most density of pathogenic SNP distribution.

Figure 8: Exon 78 and Exon 79 length example of DMD gene ENST00000357033

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 9: ClinVar classification VS Exon frequency for DMD Transcript ENST00000357033



Distribution of SNPs by ACMG-AMP classifications

- Nonsense mutation (i.e. stopgain) or frameshift mutation likely lead to more pathogenic.
- Observed a few cases with synonymous mutation (2%) also associated with pathogenicity
- Interestingly, some pathogenetic variants were also observed in healthy individuals

Figure 10-a: Partial 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	.	.	.	.	.	.

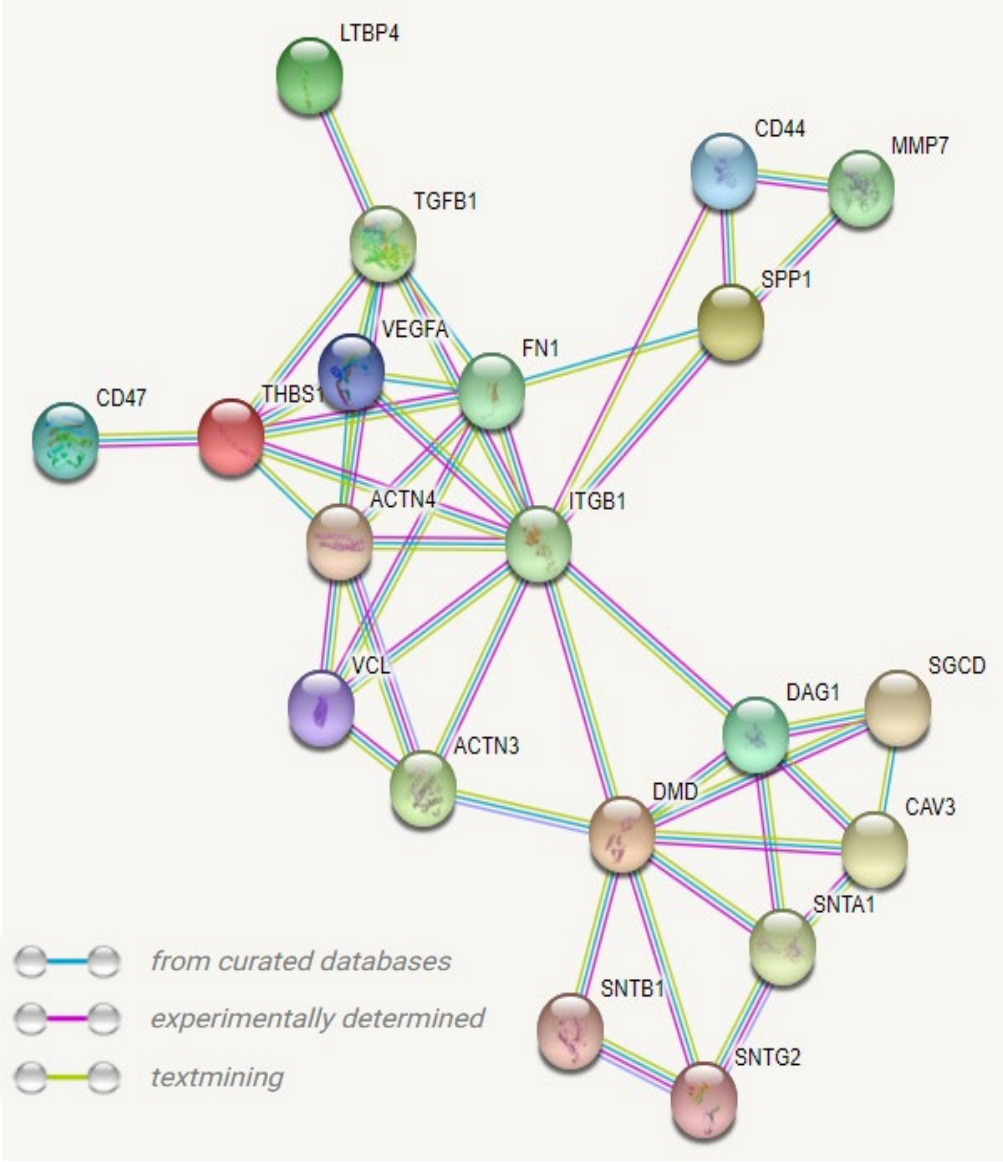
Figure 10-b: SNPs by ACMG-AMP Classification Example in DMD

	Stopgain	Frameshift deletion	Nonsynonymous SNV	Frameshift Insertion	Synonymous SNV
Pathogenic	151	57	26	20	5
Likely Pathogenic	1	1	0	0	0
Benign	0	0	47	0	24
Likely Benign	0	0	16	0	16
Uncertain Significance	1	0	35	0	0

Interaction network resulting from the genetic modifiers identified in DMD patients

- Only interactions with confidence score over 0.9 were mapped to network.
- SPP1 interact with DMD through ITGB1, which has highest node degree and BC values in the network.
- 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.

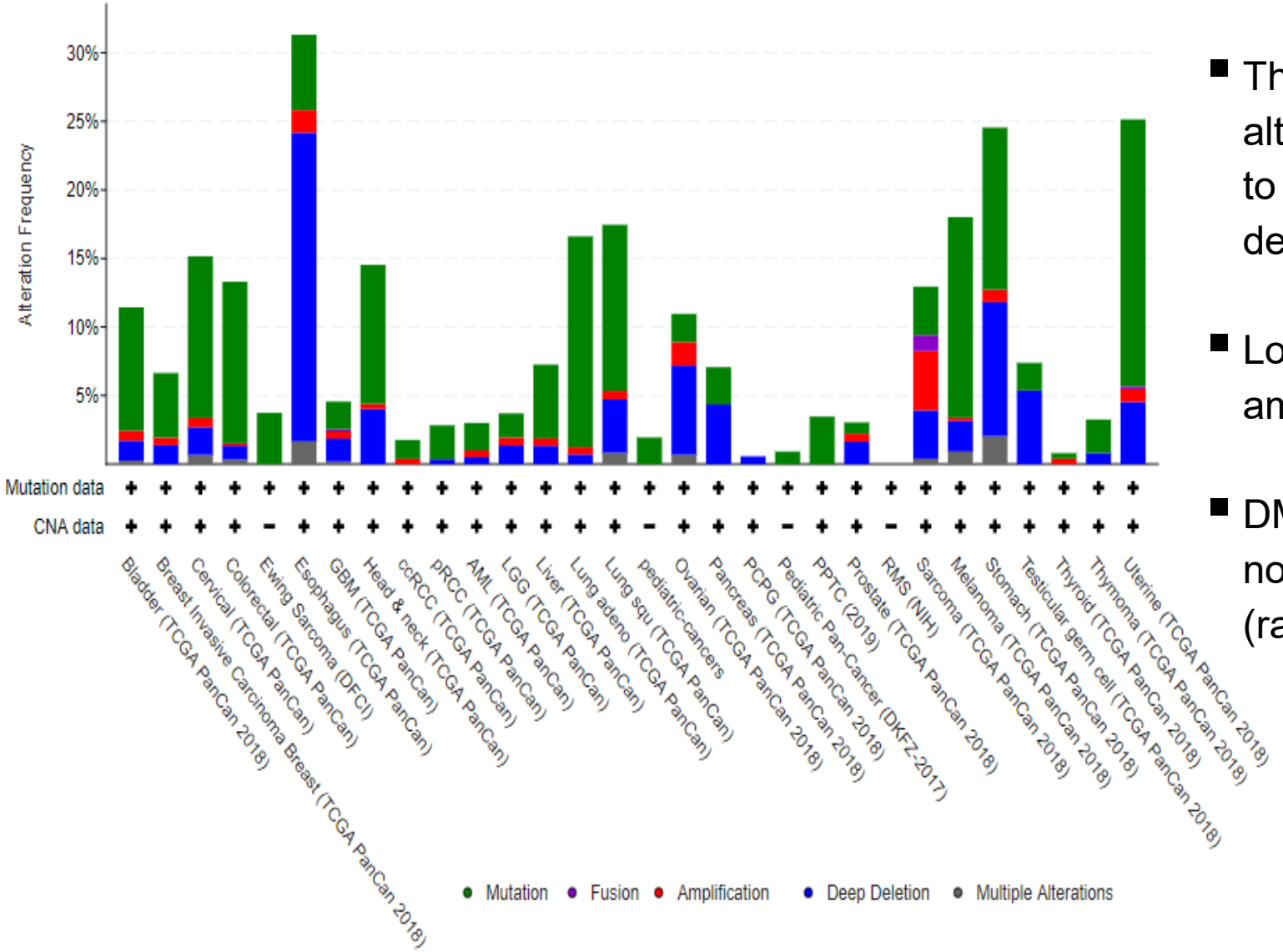
Figure 11: PPI network in DMD patients



Reference source: STRING Consortium 2000

DMD genetic alterations using cBioPortal data

Figure 12: Frequency of GMD genetic alternations in different types of tumors



- The majority of genetic alterations corresponded to mutations, deep deletions.
- Low frequency of gene amplifications.
- DMD alterations were not found in samples (rabdomyosarcomas)

Reference source: cBioPortal for Cancer Genomics

Cancer Patients with DMD alterations have poorer overall survival

Figure 13: Pooled data (30 different cancer studies)

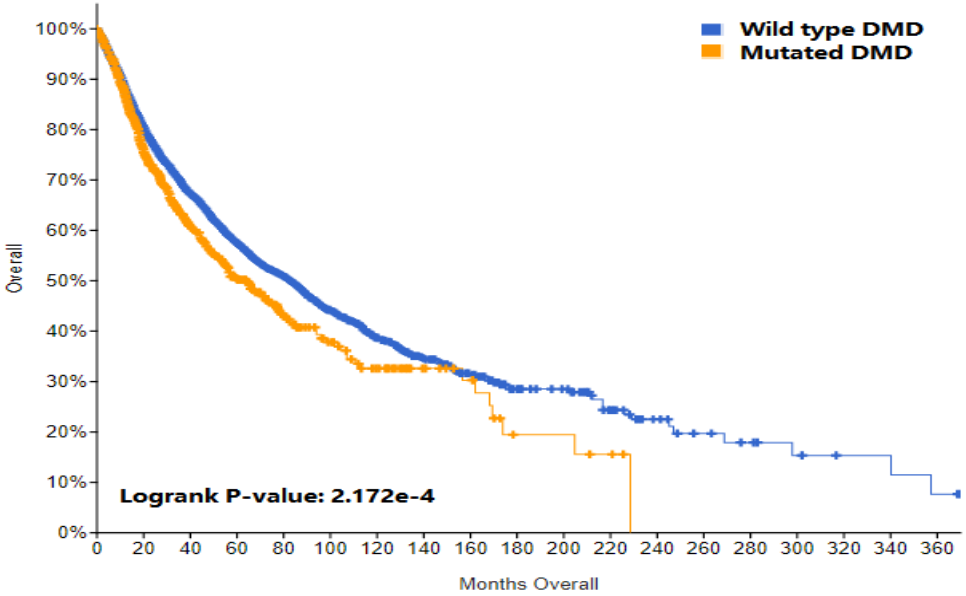
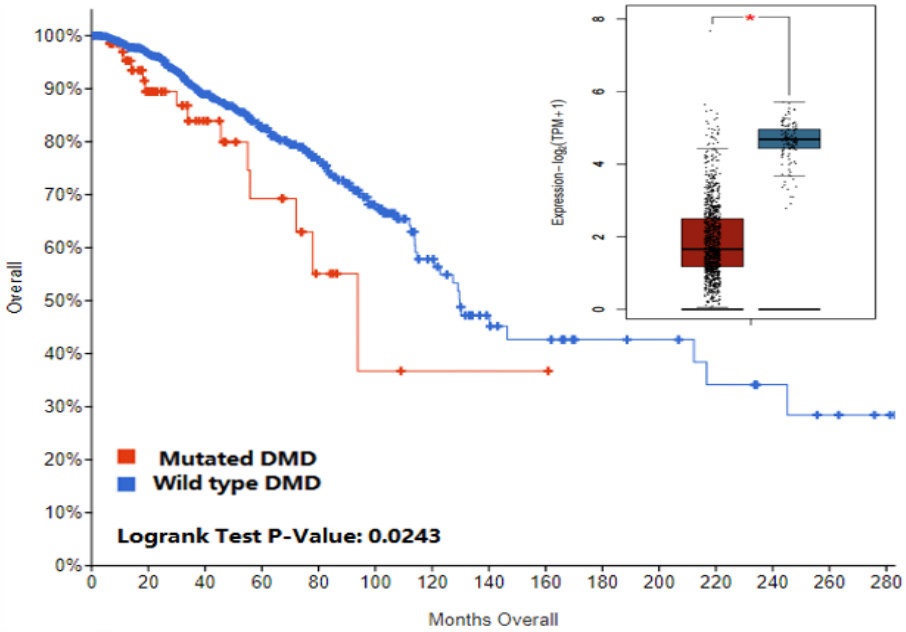
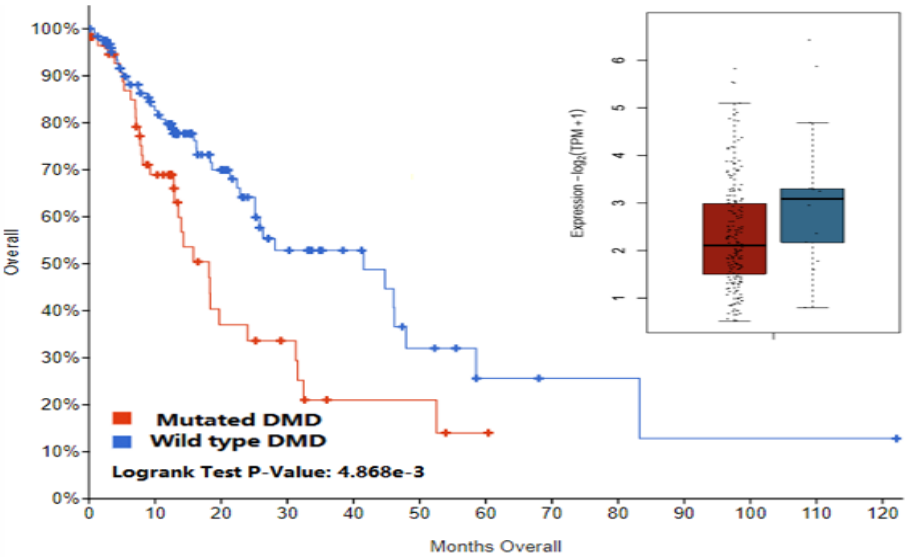


Figure 14: Invasive Breast Carcinoma



- Other tumor types e.g. ovarian carcinomas showed similar trend, but not statistically significant.
- The majority of tumor specimens had lower DMD expression compared to the normal adjacent tissue.
- The relationship between DMD genetic status and prognosis may be tumor-type specific.

Figure 15: Esophageal Carcinoma



Conclusions

- To our knowledge, this is the first data mining study with a systematic analysis of all exon variants, especially SNPs, in **the one of the largest known human gene**.

- This study examined total 3,627 exonic SNPs in the DMD gene. **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**.

- According to 1000 Genomes project, genetic variants (i.e., nonsynonymous mutation) associated with relatively **higher alteration frequency in African**. Similar frequency distributions were observed among America, Europe, East and South Asia.

- Protein 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.

- This study result also suggest DMD gene may serve as a **diagnostic and therapeutic target** for certain types of cancer.

Bibliography

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. *Chang, X. and K. Wang, wANNOVAR: annotating genetic variants for personal genomes via the web. J Med Genet, 2012. 49(7): p. 433-6.*
8. *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.*