A bioinformatic analysis of the Duchenne muscular dystrophy gene and associated gene variants across human cancers

A thesis submitted to the University of Nottingham for the degree of MRes in Bioinformatics

Lee Machado

September 2022

Supervisors: Professor Richard Eames and Professor David Brook

Abstract

The Duchenne muscular dystrophy (*DMD*) gene and its major translated protein product Dystrophin (Dp427m) have for decades been associated with musculoskeletal function, with specific mutations giving rise to dysfunction in Duchenne and Becker muscular dystrophies. Alterations in the expression of the *DMD* gene have recently been associated with the development, progression, and survival outcomes of several tumours.

A bioinformatic workflow employing an outcome-based cutpoint selection method was developed. It was implemented to provide a comprehensive approach to examine the association of *DMD* mRNA expression and survival outcomes across 33 different tumour types and used bulk RNAseq data of primary tumours from the cancer genome atlas project.

Nine of the 33 tumours had significant survival outcomes using Kaplan Meier log-rank statistics and were the focus of further downstream analysis. High *DMD* expression was significantly associated with poor survival in low grade glioma, thymoma, rectal and kidney cancer. Conversely, low expression of *DMD* was associated with poor survival in uveal melanoma, pancreatic, lung adenocarcinoma, acute myeloid leukemia, and breast carcinoma.

Univariate Cox proportional hazard modelling was used to calculate *DMD* hazard ratios. In combination with hazard ratios from other dystrophin associated glycoprotein complex genes, hierarchical clustering was used to identify clusters that may potentially be used as candidate biomarkers for different cancer types and help identify potentially common underlying causal factors in these tumours.

The expression of the individual *DMD* gene products was examined and were also significantly associated with overall survival, with specific patterns of expression likely to have differential biological effects relevant to the pathogenesis of each tumour. The smallest gene product,

Dp40 was expressed across all tumours and most tumours expressed at least one Dp71 isoform. Full length Dp427m was expressed in breast cancer, low grade glioma, lung adenocarcinoma, pancreatic adenocarcinoma, rectal cancer, and uveal melanoma. Low grade glioma had the broadest expression of different *DMD* gene products and acute myeloid leukemia was restricted to just Dp40 expression.

To explore differences between tumours expressing high or low amounts of total *DMD* RNA, differential gene expression and preliminary pathway analysis identified dysregulated genes with gene ontology biological terms that related to motility and adhesion which is concordant with dystrophin's known role as a structural/scaffold protein that facilitates cellular interaction of the actin cytoskeleton with the extracellular matrix. However, in some cancers novel terms relating to ion homeostasis (pancreatic and rectal) and chemical/sensory perception (lung) were identified, and the biological significance of this is currently unclear.

Future work will require confirmation of dystrophin protein expression in these tumours with follow-up functional experiments to demonstrate that dysregulated dystrophin is a contributor to individual hallmarks of cancer. *DMD* gene or protein product expression may have potential utility as an independent prognostic marker which can further stratify patients to identify those with risk of poor survival. This knowledge may ultimately improve risk stratification, patient management and aid our understating of the role dystrophin in these cancers.

Acknowledgements

I'd like to express my love and thanks to my wife Rachel for supporting me through my PhD and now again during an MRes degree some 20 years later. Thanks to my children Ella and Josh for their patience while I have locked myself away to draft this thesis and get my R scripts to work. A big thanks to my employer (University of Northampton) who agreed to a) promote me to professor and b) support me during this apprenticeship and to my colleagues who have supported and valued my attempts to upskill myself as a researcher and expand the range of bioinformatic support available to our students on our Life Science programs. Thanks to current and recent line managers John Sinclair and Dr Peter Jones in this regard. Many thanks to Nottingham University Professors Richard Eames and David Brook for taking the time to supervise me during this project. I imagine supervising an academic during a research project is a huge headache and I greatly appreciate their patience and input into the project. Thanks to my course mates on the apprenticeship programme and for the online coffees and emotional and practical support. A huge thanks to Dr Adam Blanchard, our programme lead, who kept us all on track and helped expand my professional network over the last few years. My huge thanks to Professor Karen Anthony and members of our joint research team for introducing me to the wonderful world of *DMD* biology and for your collegiate support and camaraderie. Finally, thanks to the patients across multiple clinical sites who consented for clinical samples to be used as part of The Cancer Genome Atlas project making their genomic and clinical data available for researchers across the globe in the hope of improving survival outcomes of all cancer patients.

Publications

Jones, L., Naidoo, M., Machado, L.R., Anthony, K. (2021) The Duchenne muscular dystrophy gene and cancer. *Cellular Oncology*. **44**(1), 19–32.

Khalsan, M., Machado, L.R., Al-Shamery, E.S., Ajit, S., Anthony, K., Mu, M., Agyeman, M.O. (2022) A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction. *IEEE Access*. **10**, 27522–27534.

Naidoo, M., Jones, L., Conboy, B., Hamarneh, W., D'Souza, D., Anthony, K., Machado, L.R. (2022) Duchenne muscular dystrophy gene expression is an independent prognostic marker for IDH mutant low-grade glioma. *Scientific reports*. **12**(1), 19–32.

Contents

Chapter 1 14

1. Introduction 14

1.1 The biology of *DMD* 14

1.1.1 *DMD* gene structure and encoded dystrophin products 14

1.1.2 Dystrophin as a component of the Dystrophin associated protein complex (DAPC) 16

1.1.3 Duchenne and Becker muscular dystrophies: The canonical diseases associated with *DMD* mutation 18

1.2 A putative role for Dystrophin in human cancers 20

1.2.1 The hallmarks of cancer 20

1.2.2 Overview of *DMD* in different tumour types 22

1.3 The Cancer Genome Atlas (TCGA) as a repository for mining cancer genomics data 29

1.4 Survival analysis and Cox proportional hazard modelling 31

1.5 Aims and objectives 34

Chapter 2 35

2. Materials and methods 35

2.1 Clinical sample data and ethical approval 35

2.2 Overview of workflow for *DMD* and gene variant survival analysis 36

2.2.1 Importing normalised gene expression and associated clinical data from TCGA 37

2.2.2 Dichotomisation of patient tumours into high and low expressers 40

2.2.3 Kaplan-Meier survival analysis *DMD* high and low expressing tumour groups 40

2.2.4 Univariate hazard modelling 40

2.2.5 Cluster analysis 40

2.3 Isoform expression analysis 41

2.4 Differential gene expression and pathway analysis 42

2.5 Software, code, and data availability 43

Chapter 3 44

3. A global analysis of *DMD* gene expression across TCGA tumours. 44

3.1 Introduction 44

3.2 Results 44

3.2.1 Estimating cutpoint values with Maxstat 44

3.2.1 Kaplan Meier survival analysis of TCGA cancers 46

3.2.2 Univariate Cox model analysis 48

3.2.3 Univariate DAPC gene hazard modelling and cluster analysis 50

3.2.4 Variant specific expression analysis across TCGA tumours 54

3.2.5 Variant specific survival analysis across TCGA cancers 57
3.2.6 iDEP pathway analysis 68
Chapter 4 79
4. Discussion 79
4.1 *DMD* is associated with survival outcomes in cancer patients 79
4.2 Is *DMD* a driver gene in cancer? 87
4.3 Do Duchenne and Becker patients have an increased risk of cancer? 88
4.4 Future work and limitations 90
4.5 Conclusions 93
5. References 94
6. Appendices 110

List of Figures

Figure 1.1 Structure of *DMD* gene. 15
Figure 1.2 Dp71 isoforms and preferred nomenclature. 16
Figure 1.3 Dystrophin as a component of the Dystrophin-associated protein complex (DAPC). 17
Figure 1.4 Dystrophin gene and protein. 19
Figure 1.5 Expression of dystrophin. 20
Figure 1.6 Updated Hallmarks of cancer. 21
Figure 1.7 cBioportal was used to rank *DMD* alteration frequencies across the TCGA PanCancer Atlas studies 22
Figure 1.8 Overview of TCGA cases including clinical and molecular data. 29
Figure 1.9 mRNA-Seq processing and data comparison in TCGA Legacy and the GDC 30
Figure 1.10 Interpretation of survival curves. 32
Figure 1.11 The log-rank chi-square statistic comparing survival curves 32
Figure 2.1 Bioinformatic workflow 37
Figure 2.2 Bioinformatic pipeline used by GDC to generate gene expression data 38
Figure 2.3 iDEP: Integrated Differential Expression and Pathway analysis 43
Figure 3.1 Estimated cutpoints (dashed vertical line) for total mRNA expression in selected TCGA cancers 45
Figure 3.2 *DMD* expression is significantly associated with overall survival in nine specific tumour types 47
Figure 3.3 Hazard ratios in selected TCGA tumours 49
Figure 3.4 Association of DAPC gene expression with hazard ratios in selected TCGA tumours 52
Figure 3.5 Cluster dendrograms based on hazard ratios 53
Figure 3.6 Expression of individual *DMD* gene products in selected TCGA cancers. 55
Figure 3.7 Cluster dendrograms based on Dp gene expression 56
Figure 3.8 The expression of Dp71ab and Dp71b gene products are significantly associated with BRCA survival outcomes 57
Figure 3.9 The expression of Dp71ab and Dp71b products are significantly associated with KIRP survival outcomes 58
Figure 3.10 The expression of the Dp40 gene product was not significantly associated with LAML survival outcomes 59
Figure 3.11 The expression of Dp40, Dp71ab, Dp71, Dp116 and Dp140 gene products are significantly associated with LGG survival outcomes 60
Figure 3.12 The expression of Dp71ab and Dp427m gene products are significantly associated with LUAD survival outcomes 61
Figure 3.13 The expression of Dp40, Dp71b and Dp427m gene products are significantly associated with PAAD survival outcomes 62
Figure 3.14 The expression of the Dp40, Dp71ab and Dp427m gene products are significantly associated with READ survival outcomes 63
Figure 3.15 The expression of the Dp71ab gene product was significantly associated with THYM survival outcomes 64
Figure 3.16 The expression of Dp40, Dp71ab, Dp260-1 and Dp427m gene products are significantly associated with UVM survival outcomes 65
Figure 3.17 Hazard ratios of TCGA tumours expressing specific DMD gene products 67
Figure 3.18 Exploratory analysis of the DEGs with high versus low *DMD* expression in UVM 69

Figure 3.19 Exploratory analysis of the DEGs with high versus low *DMD* expression in THYM 71
Figure 3.20 Exploratory analysis of the DEGs with high versus low *DMD* expression in READ 72
Figure 3.21 Exploratory analysis of the DEGs with high versus low *DMD* expression in PAAD 73
Figure 3.22 Exploratory analysis of the DEGs with high versus low *DMD* expression in LUAD 74
Figure 3.23 Exploratory analysis of the DEGs with high versus low *DMD* expression in LGG 75
Figure 3.24 Exploratory analysis of the DEGs with high versus low *DMD* expression in LAML 76
Figure 3.25 Exploratory analysis of the DEGs with high versus low *DMD* expression in KIRP 77
Figure 3.26 Exploratory analysis of the DEGs with high versus low *DMD* expression in BRCA 78
Figure 4.1 Composition of the DAPC at different tissue sites 82
Figure 4.2 Proposed model of *DMD* driven cancer development 85

List of tables

Table 1.1 Summary of evidence for the role of the *DMD* gene in cancer (Jones *et al*., 2021) 23
Table 2.1 Definition of TCGA cancer types abbreviations. 35
Table 2.2 Details of major *DMD* gene products including ids, genome and coding sequence start and end positions and exons counts..... 42
Table 3.1 Genes and encoded protein products in the Dystrophin associated protein complex 50
Table 4.1 Published case reports of cancer in individuals with *DMD*. 89

Abbreviations

Abbreviation	Definition
ABD	Actin binding domain
ACC	Adrenocortical carcinoma
AFF2	AF4/FMR2 Family Member 2
ALL	Acute lymphoblastic leukaemia
API	Application Programming Interface
ARMS	Amplification-refractory mutation system
BAM	Binary alignment map
BLCA	Bladder urothelial carcinoma
BMD	Becker muscular Dystrophy
BRCA	Breast invasive cancer
CCL20	Chemokine ligand twenty
CDKN2A	Cyclin-dependent kinase inhibitor 2A
CESC	Cervical squamous cell carcinoma
CGGA	Chinese Glioma Genome Atlas
CHOL	Cholangiocarcinoma
CI	Confidence interval
CLL	Chronic lymphocytic leukaemia
CNA	Copy number alteration
CNS	Central nervous system
CNV	Copy number variation
COAD	Colorectal adenocarcinoma
CRAN	Comprehensive R archive network
CRCS	Continuous Representation of Codon Switches
CRM1	chromosomal region maintenance one
DAG1	Dystroglycan 1
DAPC	Dystrophin-Associated Protein Complex
DDX53	DEAD-Box Helicase 53
DEG	Differentially expressed gene
DEP	Differential Expression and Pathway analysis
DG	Dystroglycan
DGC	Dystrophin-associated glycoprotein complex
DLBC	Diffuse large B cell lymphoma
DMD	Duchenne mdystrophy
DNA	Deoxyribonucleic acid
DSS	Disease specific survival
DTNA	Dystrobrevin Alpha
DTNB	Dystrobrevin beta
EBV	Epstein Barr virus
ECM	Extracellular matrix
ESCA	Esophageal carcinoma
EXP	Expression
FDR	False discovery rate

FPKM	Fragments per kilobase per million reads
FRAXC	Fragile Site, Aphidicolin Type, Common, Fra(X)(Q22.1) C
GBM	Glioblastoma multiform
GDC	Genomic data commons
GEO	Gene Expression Omnibus
GIST	Gastrointestinal stromal tumour
GO	Gene ontology
HNSC	Head and neck squamous cell carcinoma
HOX	Homeobox gene
HR	Hazard ratio
ID	Inhibitor of DNA binding
IDH	Isocitrate Dehydrogenase
IL1RAPL1	Interleukin 1 Receptor Accessory Protein Like 1
KICH	kidney chromophobe
KIRC	kidney renal clear cell carcinoma
KIRP	kidney renal papillary cell carcinoma
LAML	Acute myeloid leukaemia
LGG	Low grade glioma
LIHC	Liver hepatocellular carcinoma
LMS	Leiomyosarcoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MAGE	Melanoma-associated antigen
MDM2	Mouse double minute two homolog
MESO	Mesothelioma
METH	Methylation
MLPA	Multiplex-ligation dependent probe amplification
MMP2	Matrix metalloproteinase-2
NCI	National Cancer Institute
NF1	Neurofibromatosis type 1
NGS	Next generation sequencing
NMJ	Neuromuscular Junction
NOS	Nitrogen oxide synthase
NOS1	Nitrogen oxide synthase one
NPC	Nasopharyngeal carcinoma
OFD1	Orofaciodigital syndrome type 1
ORF	Open reading frame
OS	Overall survival
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and paraganglioma
PCR	Polymerase chain reaction
PFI	Progression free interval
PPI	Protein-protein interaction
PRAD	Prostate adenocarcinoma
READ	Rectal adenocarcinoma

RM	Reads mapped
RMA	Rhabdomyosarcoma
RMS	Rhabdomyosarcoma
RNA	Ribonucleic acid
RSEM	RNA-Seq by Expectation-Maximization
RSK4	Ribosomal Protein S6 kinase four
SARC	Sarcoma
SEQ	Sequencing
SG	Sarcoglycan
SGCA	Sarcoglycan Alpha
SGCB	Sarcoglycan beta
SGCD	Sarcoglycan delta
SGCE	Sarcoglycan epsilon
SGCG	Sarcoglycan gamma
SGCZ	Sarcoglycan zeta
SKCM	Skin cutaneous Melanoma
SNAP25	Synaptosome Associated Protein 25
SNP	Single nucleotide polymorphism
SNTA1	Syntrophin alpha one
SNTB1	Syntrophin beta one
SNTB2	Syntrophin beta two
SNV	Single nucleotide variant
SSPN	Sarcospan
STAD	Stomach adenocarcinoma
STAR	Spliced transcripts alignment to a reference
STRING	Search tool for the retrieval of interacting genes /proteins
STS	Soft tissue sarcoma
TCGA	The cancer genome Atlas
TGCT	Testicular germ cell tumours
THCA	Thyroid carcinoma
THYM	Thymoma
TP53	Tumour protein fifty-three
TPM	Transcript per million
UCEC	Uterine corpus endometrial carcinoma
UCS	Uterine carcinosarcoma
UCSC	University of California Santa Cruz
UPGMA	Unweighted pair group method with arithmetic mean
UQ	Upper quartile
UVM	Uveal Melanoma
WDR44	WD repeat domain forty-four
WHO	World Health Organisation
WW domain	Tryptophan-Tryptophan
ZZ domain	Zinc-Zinc domain

Chapter 1

1. Introduction

1.1 The biology of *DMD*

1.1.1 *DMD* gene structure and encoded dystrophin products

The Duchene muscular dystrophy gene (*DMD)* is named after the clinical condition of the same name and is one of the largest genes in the human genome comprised of seventy-nine exons spanning 2Mb on the short arm of chromosome X (ChrX (p21.2-p21.1). It resides within a known a fragile site and encodes a large 427KDa protein with an N-terminal binding domain and multiple spectrin repeats (Jones *et al.*, 2021). As part of a dystrophin associated protein complex (DAPC), dystrophin bridges the inner cytoskeleton to the extracellular matrix. The locus is a known site where point mutations and larger copy number alterations (deletions or duplications) contribute to disease causing Duchene and Becker muscular dystrophies (Muntoni *et al.*, 2003). In addition, mutation in *DMD* contributes to cardiomyopathy. With seven alternate promoters and alternative splicing events, a number of dystrophin gene variants (Figure 1.1) and isoforms (i.e. Dp71; Figure 1.2) are produced that have distinct tissue localisation and function which are incompletely characterised. The complexity is reflected in examples such as Dp71 and Dp40, which share the same promoter and first exon but Dp40 makes use of an alternative polyadenylation site.

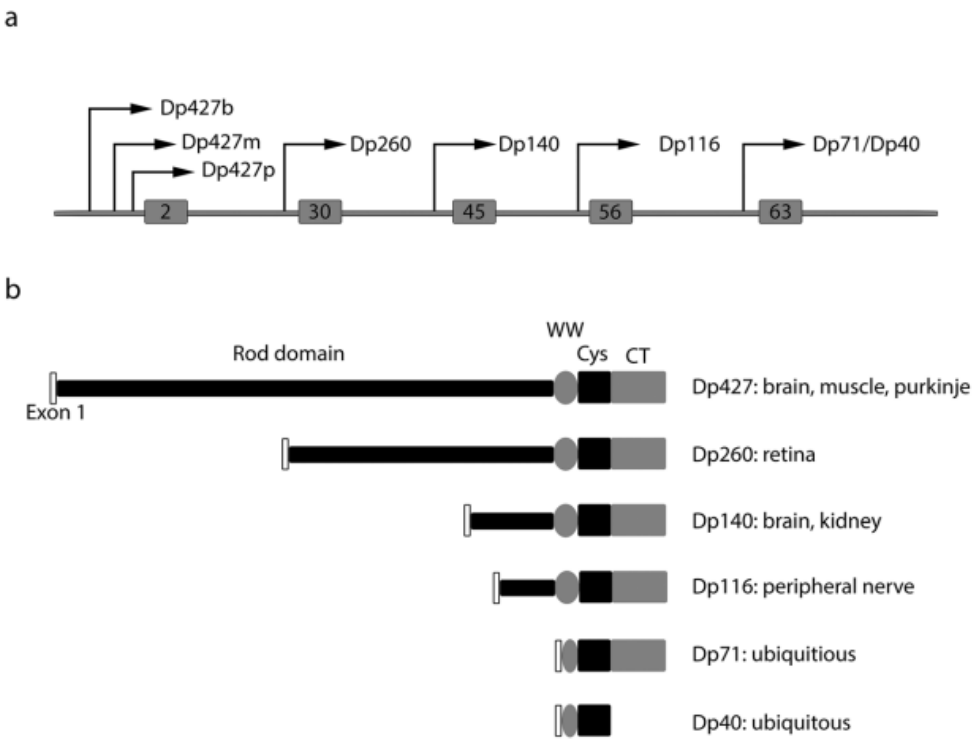


Figure 1.1 Structure of *DMD* gene.
(a) location of independent promoters. (b) Differential expression and domain structure of the different gene products. WW:WW domain; Cys: cysteine rich domain; CT: C-terminal domain (Jones *et al.*, 2021).

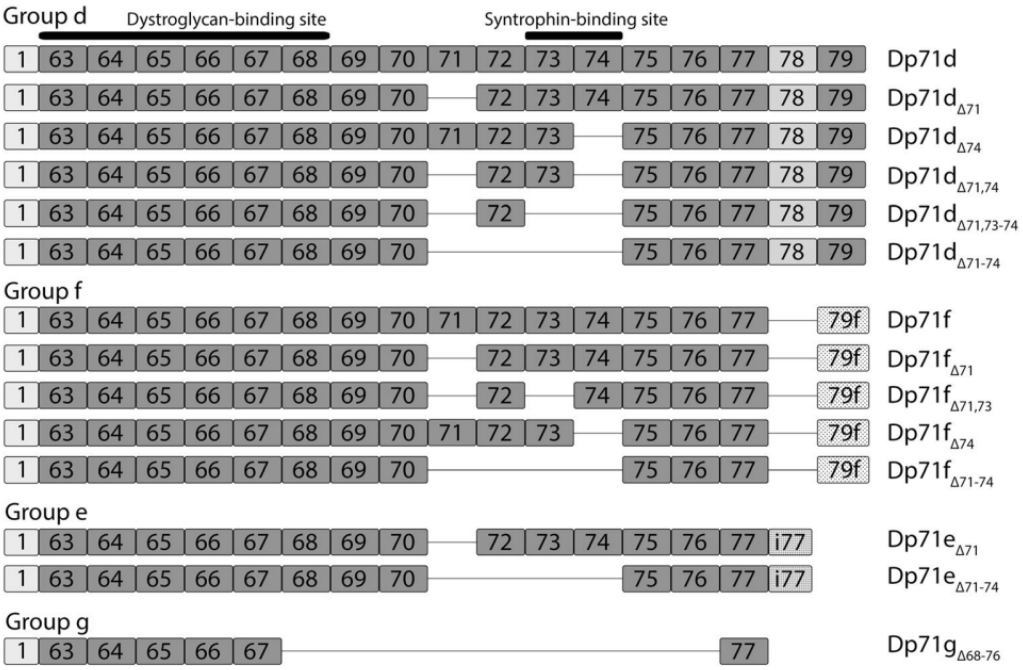


Figure 1.2 Dp71 isoforms and preferred nomenclature.
Dp71 splice isoforms are grouped according to their C-terminus. Group d contains exon 78 and 79. Group F lacks exon 78 and has an alternative exon 79 (79f). Group E contains part of intron 77 (i77) and lacks exon 78 and 79. Group G has an isoform with a stop codon in exon 77. Alternative names of major Dp71 isoforms: Dp71d = Dp71; Dp71d$_{\Delta71}$ = Dp71a; Dp71f = Dp71b; Dp71; Dp71f$_{\Delta71}$ = Dp71ab. Their differential C-termini are illustrated and the location of dystroglycan and syntrophin- binding sites are indicated (Naidoo and Anthony, 2020).

1.1.2 Dystrophin as a component of the Dystrophin associated protein complex (DAPC)

Dystrophin is associated with a protein complex (Figure 1.3) that has unique structural and functional roles that depend on tissue localisation. The canonical role of the DAPC is to stabilise the plasma membrane of striated muscle cells by linking it to the basal lamina by interacting through ECM interactions. If this does not correctly function, the result is a collection of inherited diseases characterised by degeneration of muscle fibres and muscle weakness (discussed in next section). Additional functions also include regulation of cation and water channels as well as kinases and nNOS (Pilgram *et al.*, 2010). DAPC proteins can reside either as extracellular, transmembrane or cytoplasm proteins. α-dystroglycan is located on the surface of the sarcolemma, is heavily glycosylated and interacts with laminin-

2 linking the plasma membrane to the extracellular matrix. α-dystroglycan is associated with β-dystroglycan and together forms an interaction with dystrophin. A Sarcoglycan sub-complex is tightly linked to dystroglycan (through interaction with Sarcospan) and consists of four transmembrane proteins (α, β, γ and δ). The sarcoglycan-sarcospan sub-complex is also involved in signal transduction and mechanoprotection. Cytoplasmically, dystrophin interacts with both β-dystroglycan and the actin cytoskeleton. Dystrophin also interacts with syntrophins, dystrobrevins, and nNOS (a producer of nitric oxide) which are recruited to the C-terminus of dystrophin and mediate signal transduction pathways that (for example) override sympathetic vasoconstriction and prevent functional ischemia in contracting muscles (Percival, 2018).

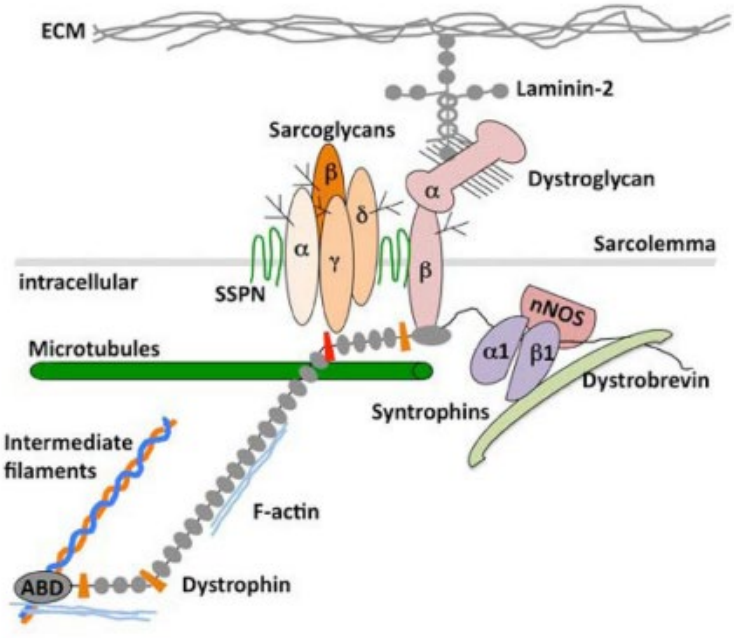


Figure 1.3 Dystrophin as a component of the Dystrophin-associated protein complex (DAPC).

The DAPC facilitates the stabilisation of muscle fibres by connecting the intracellular actin cytoskeleton to the extracellular matrix. ABD, actin binding domain; SSPN, Sarcospan; ECM, extracellular matrix; nNOS, neuronal nitric oxide synthase (taken from (Gao and McNally, 2015).

The full-length Dystrophin protein (427KDa, 3,684 amino acids) has four functional domains (Figure 1.4). An amino-terminal actin-binding domain, a central rod domain with twenty-four spectrin-like repeats interspersed with 4 hinge regions, a cysteine-rich and WW containing dystroglycan-binding domain, and finally a carboxy-terminal domain facilitating interaction with the sarcolemma via dystrobrevin and syntrophin binding sites. The N and C termini are critical for function but the loss of spectrin repeats in the rod domain are somewhat tolerated (Kahana and Gratzer, 1995) and are the basis of dystrophin restoration therapies for DMD (Cirak *et al.*, 2012).

1.1.3 Duchenne and Becker muscular dystrophies: The canonical diseases associated with *DMD* mutation

DMD mutations occur as germline (in two thirds of cases) or as sporadic copy number alterations. 60-70% of cases harbour deletions and 10% have duplications within the *DMD* gene. There are also point mutations (20-35% of cases) that collectively give rise to Duchenne muscular dystrophy (*DMD*) and the milder and related disorder Becker muscular dystrophy (BMD). These monogenic disorders, are rare and have an estimated combined prevalence of 1 in 7250 males (Romitti *et al.*, 2015). Clinically, from 4-5 years of age in *DMD*, there is progressive weakness and degeneration of skeletal muscle. Patients struggle to walk and are typically wheelchair bound by their early teens. Although better management has ameliorated life expectancy, patients require more clinical intervention (ventilation) in their mid-late teens and often die with associated cardiomyopathy by 30 years of age. In contrast, BMD occurs later (average age of 12 years old), and ambulation loss, if it occurs, does not manifest until post-twenties with overall survival times that are longer than *DMD* and in some cases can even approach normal life expectancy.

Typically, *DMD* deletions (60-65% of cases) account for the largest proportion of dystrophin mutations and more rarely duplications (5-15%) are observed. Although these CNAs can occur anywhere within the *DMD* locus, two deletion hotspots cluster, either within the central region (between exons 43 and 53) of the gene or towards the 5' end (exons 6 and 7) (Figure 1.4). The consequence of these and other mutational events may be a truncated but partly functional protein (i.e. due to an in-frame mutation in BMD) or a truncated but unstable protein (due to a frameshift mutation or nonsense point mutation in DMD).

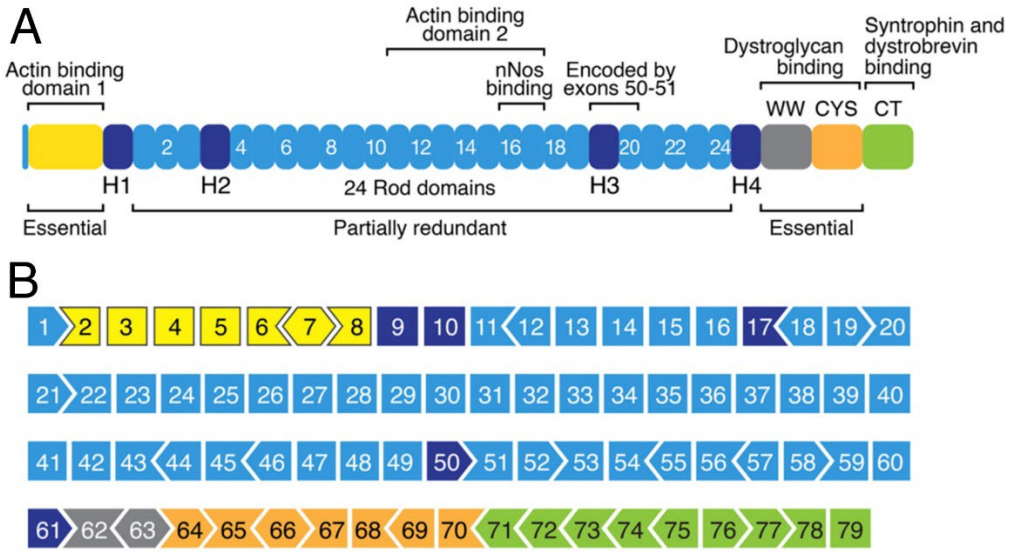


Figure 1.4 Dystrophin gene and protein.
(A) The dystrophin protein has functional domains including a hinge (H), WW, cysteine-rich (CR), and carboxyl-terminal (CT) domain. (B) Options for alternative splicing of 79 exons of the human dystrophin gene. The colours correspond to functional domains of the protein in A. Exon shapes highlight whether splicing between adjacent exons maintains a contiguous ORF (when the shapes fit together) (Olson, 2021)

The phenotype of patients with *DMD* mutations can be complex and immunohistochemistry of muscle biopsy material using anti-dystrophin antibodies can be informative and diagnostic. In some cases interesting phenotypes occur intermediate between Becker and Duchenne. In others, mosaicism leads to revertant fibres (Figure 1.5) (Muntoni *et al.*, 2003). For female carriers, even asymptomatic individuals, there can be high risk of cardiomyopathy with a

spectrum of severity due to random X inactivation. However, overt DMD in females is very rare but does happen (Abbadi *et al.*, 1994).

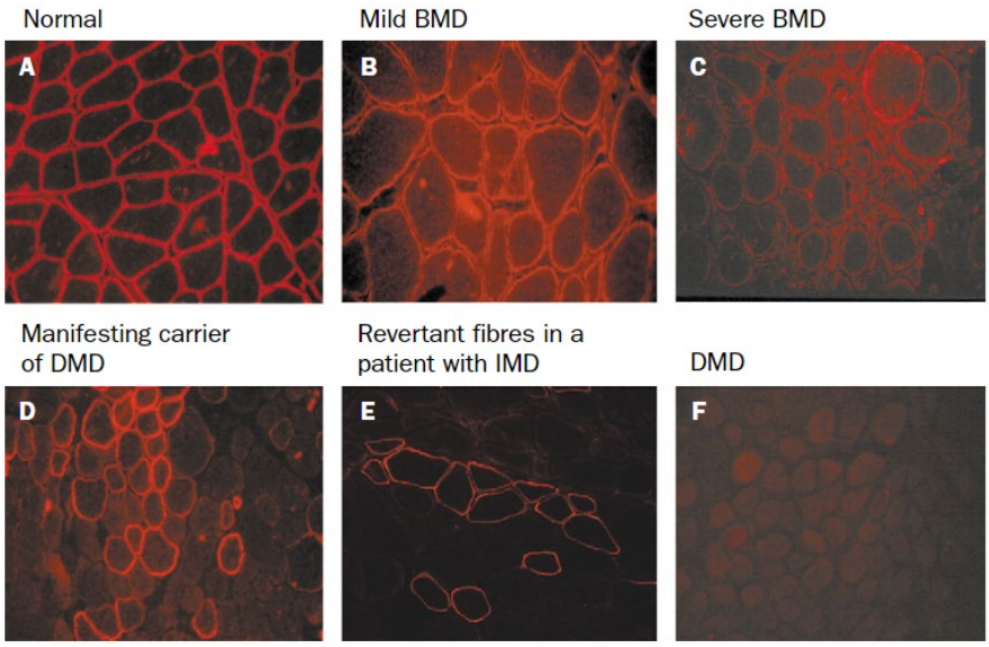


Figure 1.5 Expression of dystrophin
Dystrophin protein expression in (a) Normal muscle, (b) mild BMD, (c) severe BMD, (d) a manifesting *DMD* carrier, (e) a patient with an intermediate BMD/DMD phenotype and (f) DMD (Muntoni *et al.*, 2003).

1.2 A putative role for Dystrophin in human cancers

1.2.1 The hallmarks of cancer

Recently, the Hanahan and Weinberg hallmarks of cancer have been updated (Hanahan, 2022) where the original hallmarks defined capabilities acquired by cells as they underwent neoplastic progression (Figure 1.6). The acquired capabilities proposed in 2000 were consolidated where provisional hallmarks have now been sufficiently validated and are considered part of the core set. In this recent review by Hanahan, two new proposed hallmarks and two enabling characteristics have been described.

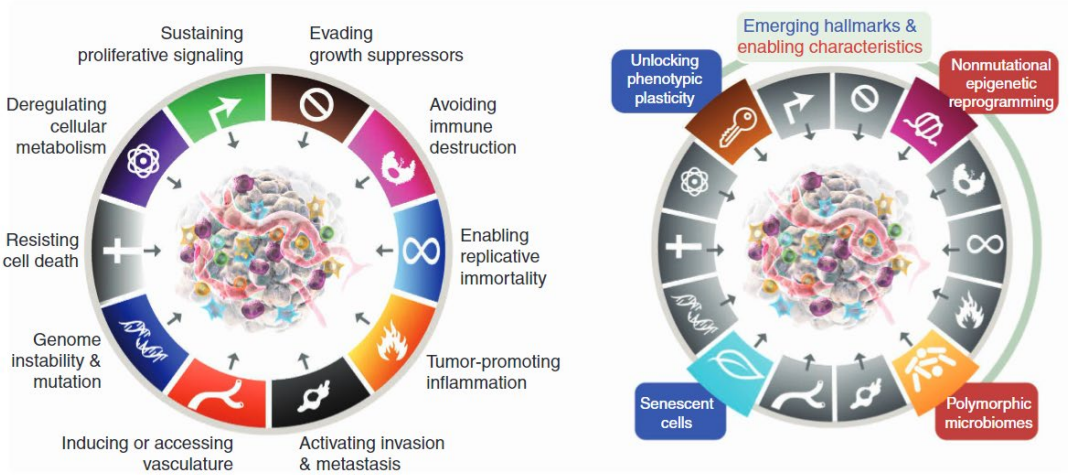


Figure 1.6 Updated Hallmarks of cancer.
Left Hallmarks of cancer embodying eight hallmark capabilities and two enabling characteristics (Inducing or accessing vasculature and tumour promoting inflammation). **Right** Additional proposed emerging Hallmarks and enabling characteristics (Hanahan, 2022).

The two new emerging hallmarks are unlocking phenotypic plasticity (i.e. to evade or escape from the state of terminal differentiation) and (tumour promoting) senescent cells. Enabling characteristics include non-mutational epigenetic reprogramming and polymorphic microbiomes. In the context of this expanded cancer model, this thesis will explore consequences of the new enabling characteristics where alterations in gene expression may be driven either by mutation and general genome instability or these new enabling characteristics, which include non-mutation or epigenetic reprogramming (Skrypek *et al.*, 2017). This may occur because of mechanisms within the tumour microenvironment that results in epigenetic reprogramming, for example hypoxia in tumours may alter the activity of enzymes involved in epigenetic reprogramming (Thienpont *et al.*, 2016). Another example may be epigenetic regulatory heterogeneity or epigenetic regulation of stromal cell types that populate the tumour microenvironment (Lu *et al.*, 2020).

1.2.2 Overview of *DMD* in different tumour types

There have been a number of studies suggesting *DMD* mutation is associated with tumorigenesis across different cancers which we reviewed recently (Jones *et al.*, 2021) and (Figure 1.7).

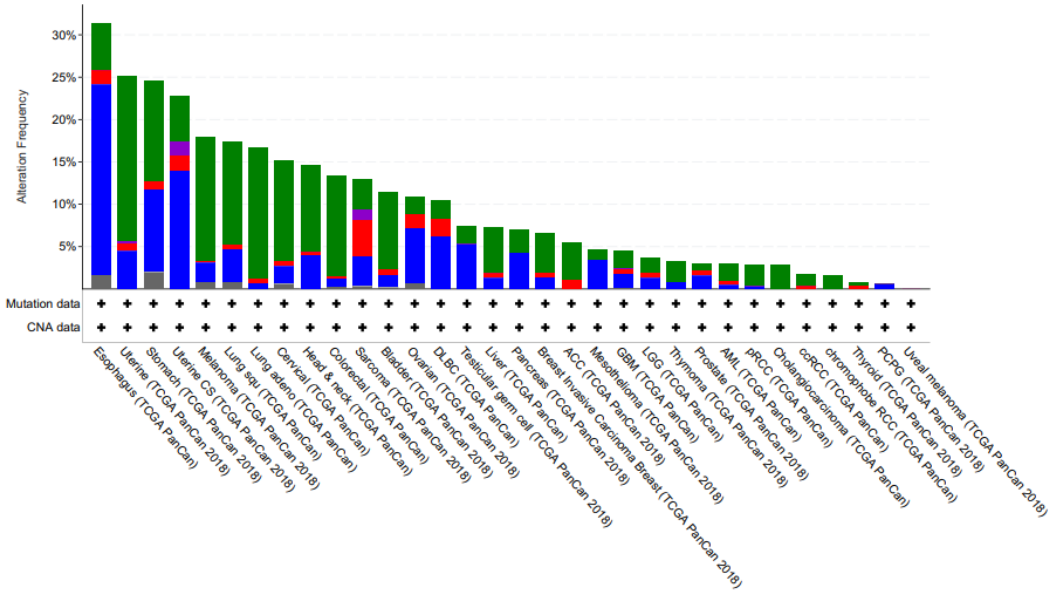


Figure 1.7 cBioportal was used to rank *DMD* alteration frequencies across the TCGA PanCancer Atlas studies

10,953 patients were studies across thirty-three studies. Alteration frequencies consist of mutations (green), fusions (purple), amplifications (red), deep deletions (blue) and multiple alterations (grey). Only TCGA studies with both mutation data and copy number alteration (CNA) data are shown (Jones *et al.*, 2021).

Examples are found across sarcomas, central nervous system tumours, melanomas, and haematological malignancies. There are also several carcinoma's that are implicated. Below (and Table 1.1) are detailed some of the studies that report these interesting associations.

Table 1.1 Summary of evidence for the role of the *DMD* gene in cancer (Jones *et al.*, 2021)

Tumour type	*DMD* mutation(s)	*DMD* expression	Oncogenic or tumour suppressive?	Strength of evidence	Ref(s)
STS	5' deletions affecting only Dp427. Intron retention.	Dp427 absent or severely reduced, Dp71 maintained	Tumour suppressive role for Dp427	+++	[18, 37, 38, 41]
Olfactory neuroblastoma	5' deletions	Predicted to affect Dp427 expression and maintain Dp71	nd	++	[22]
Meningioma	5' deletions, 2nd most frequently altered gene	Reduced Dp427 expression	Tumour suppressor role for Dp427	+++	[46, 47]
Malignant melanoma	Tumour specific deletions and polymorphisms e.g. IF Δ3–29, IF Δ17–30, OOF Δ42–43	Reduced Dp427m expression, Dp71 maintained, Dp116 frequently absent	Tumour suppressor role for Dp427	+	[21, 43]
Lymphoma	nd	*DMD* downregulated in tumour vs. progenitor cells. *DMD* upregulated in EBV-positive vs. EBV-negative tumours	nd	++	[19, 43]
Lymphocytic leukaemia	nd	*DMD* upregulated	Oncogenic	++(+)	[20, 43, 57–59]
Lung adenocarcinoma	nd	*DMD* downregulated	Oncogenic role for Dp71	++	[24, 43]
Gastric adenocarcinoma	nd	*DMD* downregulated, attributed to Dp71	Tumour suppressive role for Dp71	++	[25]
Nasopharyngeal carcinoma	Intronic SNP rs5927056	nd	nd, SNP associated with reduced risk	++	[26]
Oropharyngeal squamous cell carcinoma	nd	*DMD* downregulated	Tumour suppressive	++(+)	[61, 62]
Renal cell carcinoma	nd	*DMD* upregulated	nd	+	[43]
Other carcincomas including prostate cancer, pancreatic ductal adenocarcinoma, colon, breast and uterine cancer	nd	*DMD* downregulated (nd for uterine cancer)	Tumour suppressive	+(+)	[43]

Only characterised mutations are noted. Where we refer to *DMD* we cannot attribute the effect to specific gene product(s). IF: in-frame; OOF: out-of-frame; Δ: denotes exon deletions; nd: not determined; +: *in-vitro* or *in-silico* evidence only; ++: primary tumour cells or *in-vivo* evidence, tissue gene expression; +++: survival and clinicopathological associations, xenograph models. Parentheses indicate evidence partially met

1.2.2.1 Sarcomas

Sarcomas have been of particular interest in evaluating whether *DMD* is associated with cancer. In part, this is because of the role of dystrophin in the normal biology of muscle. *mdx* mice have a naturally occurring mutation (nonsense point mutation in exon 23) that abrogates Dp427 expression. A consequence is that aged *mdx* mice develop alveolar rhabdomyosarcoma (RMS) like tumours (Chamberlain *et al.*, 2007). The occurrence of RMS in these mice is thought to be, in part, due to continual degeneration and regeneration of myofibres throughout the life of the animals. The result is increased satellite cell proliferation and their constitutive activation increasing the chance of further mutation that may affect cellular differentiation (Chamberlain *et al.*, 2007). RMS spontaneously develops in nine percent of *mdx* mice aged over one year, compared with control animals (Fernandez *et al.*, 2010). Upon further characterization of the RMS tumours, they were found to be more consistent with the embryonal rather than alveolar type. These tumours had mutations in genes orthologous with human genes including *TP53* and mouse double minute 2 (*MDM2)*, a

negative regulator of p53. In support for the evidence of *DMD* in sarcomas, abnormalities in the dystrophin associated protein complex (particularly dystroglycans) have been linked to tumorigenesis (Brennan *et al.*, 2004; Sgambato and Brancaccio, 2005). Interestingly, the incidence of tumours in *mdx* mice has been as high as 40% in one study (Schmidt *et al.*, 2011). The reasons for this were unclear but may include housing or environmental factors. Alternatively, the genetic background of the mice may be pertinent which is why researchers compared *DMD* deficiency in mice on different backgrounds (Schmidt *et al.*, 2011). For example, both *mdx* mice and *mdx-3Cv* mice (Cox *et al.*, 1993) both develop tumours. However, *mdx-3Cv* mice have reduced incidence (and develop tumours at approx. 660 days compared with 540 days: mean age of onset). Both animals were on a *C57BL/6* backgrounds. However *mdx* animals lack only Dp427 whilst mdx-*3Cv* mice have low Dp427 expression but lack all C-terminal products. Therefore, strain specific differences or the levels of the different *DMD* gene products may account for differences in the incidence of cancer in these animals. *Mdx* mice also exhibit genomic instability with recurrent amplification of Jun and *Met*. *CDKN2A* and *NF1* are frequently lost and copy number gains frequently occur involving chromosomes 8 and 15. In summary, mouse models recapitulating Duchenne muscular dystrophy harbour genomic instability with increased risk of specific sarcomas.

In humans, a study by Wang *et al* identified Dp427 as having tumour suppressor activity in myogenic tumours (Wang *et al.*, 2014) where intragenic deletions drive progression to high grade sarcoma. The authors used SNP arrays and identified deletions in 63% of myogenic cancers. This included gastrointestinal stroma tumours (GIST), leiomyosarcomas (LMS) and Rhabdomyosarcomas (RMA). Employing Multiplex Ligation Dependent Probe Amplification (MLPA), 43% of high-grade myogenic tumours had copy number alterations within the *DMD* gene. These mutations typically occurred and involved exons 1-3 and but rarely extended

beyond exon 62. Consequently, Dp427 was typically absent in these tumours. However, the smaller dystrophin gene product Dp71 was maintained in these patients implicating a role for Dp71 in the pathology of these myogenic tumours. Functionally, Dp71 knockdown reduced cancer cell growth in cell line models. Restoring Dp427 expression resulted in inhibition of migration, invadeapodia formation and invasiveness. Therefore, the lack of full length Dp427 but the presence of the shorter Dp71 gene variant may be an oncogenic driver event in the tumorigenesis of these myogenic cancers. Array comparative genomic hybridization studies confirm *DMD* deletions in 16.5% of all tumours examined (Mauduit *et al.*, 2019). This occurred in 16.5% of sarcomas with structurally complex genomic profiles (including LMS), 21.6% of synovial sarcomas and 14.2% of GIST cases. In summary, these data support a role for *DMD* dysregulation in the pathogenesis of myogenic tumours.

1.2.2.1 Central nervous system tumours

RNAseq and microarray analysis of data from the CBioPortal repository and GEO database respectively allowed *DMD* expression and mutation to be identified in non-myogenic tumours including those of the central nervous system where *DMD* was overexpressed compared to healthy tissues (Luce *et al.*, 2017). This included ependymoma and astrocytoma. Conversely, medulloblastoma was underexpressed compared with matched control tissue. Recent work from our group has identified a novel association of *DMD* expression with low grade glioma using bulk RNAseq data (discussed further in Chapter 4). High *DMD* expression was significantly associated with poor survival outcomes in low grade glioma (LGG) with a difference in survival of over seven years ($p=<0.01$). *DMD* remains significant in a multivariate model and may represent an independent prognostic marker for low grade glioma. This association of *DMD* with survival was only apparent in IDH mutant cases where non-1p/19q deleted patients could be further stratified into high and low *DMD* groups. This work identifies

DMD expression as an independent prognostic marker potentially further stratifying IDH mutant cases to identify those at increased risk of poor survival. For neuroblastoma, SNP analysis as well as whole genome and exome sequencing was employed on 14 cases to identify somatic *DMD* deletions which occurred in 86% of the tumours (Gallia *et al.*, 2018). These olfactory neuroblastoma cases had mutations localised to the 5' end of the gene with predicted Dp71 retention. In high grade meningioma patients (n=55) 32% had *DMD* deletions or silenced expression (Juratli *et al.*, 2018). Twenty-one percent of deletions were in the 5' region (between exons 2-30 or entire deletions of *DMD*). These cases had a loss or reduction of full-length dystrophin resulting in reduced density of cytoskeletal components in the tumour. When comparing patients with *DMD* alterations, those patients had shorter progression free survival and overall survival compared to patients without *DMD* alterations. These alterations were more common in high grade meningiomas compared with low grade (grade I and II) meningiomas. In studying mutation patterns in these tumours, it was found that *DMD* is the second most frequently altered gene (Paramasivam *et al.*, 2019).

Glioblastoma has a particularly poor survival outcome for patients and McAvoy *et al* demonstrated that *DMD* expression was reduced in brain tumour cell lines and xenograft models of GBM (McAvoy *et al.*, 2007). GBM lines have been examined for expression of dystrophin isoforms and to date there have been six Dp71 isoforms identified in the U251-MG cell line (Rani *et al.*, 2019). Ruggieri *et al* explored Dp71 and its role in glioblastoma and meningioma and the authors found Dp71d was decreased in a GBM cell line and biopsy material compared with a control cell line (Ruggieri *et al.*, 2019). Dp71 expression resulted in reduced proliferation (Ki-67 staining) suggesting in the context of GBM Dp71 is associated with reduced proliferation.

1.2.2.2 Melanomas

In melanoma dysregulation of *DMD* may be involved in the pathogenesis of this disease (Körner *et al.*, 2007). PCR analysis of cell lines revealed *DMD* deletions in three cell lines towards the 5' region of *DMD*. These mutations were not in the classic hotspot regions that occurs in patients with dystrophinopathies suggesting these mutations are tumour specific. Dp427 was highly expressed in these cell lines. However, in an expanded panel of 55 melanoma cell lines, full length *DMD* expression was low or absent in 87% of them. Immunoblotting of cell lines show Dp71 expression remains even in the absence of Dp427. Dp427m when knocked down in melanoma cell lines and cells had reduced spheroid formation and enhanced invasion and migration.

1.2.2.3 Haematological malignancies

Baumforth *et al* used microarray analysis to show that *DMD* was downregulated eightfold in primary Hodgkin's lymphoma (in the nodular sclerosing subtype) compared with germinal centre B cells (Baumforth *et al.*, 2008). In this work EBV expression drove *DMD* upregulation compared with EBV negative cell lines. This suggests EBV driven upregulation of *DMD* in lymphoma may be important in the pathogenesis of this disease although it requires further investigation. To note, a case report identified in Becker muscular dystrophy patients an association with Hodgkin's (Cereda *et al.*, 2004) and non-Hodgkin's lymphoma (Uotani *et al.*, 2001).

In acute lymphoblastic leukaemia (ALL) only one case study has observed an association between *DMD* and ALL (Svarch *et al.*, 1988). In a study of 134 chronic lymphocytic leukaemia (CLL) patients, it was reported that high *DMD* expression in these tumours was associated with reduced cell doubling time and was predictive of patient survival where median overall

survival for patients with high *DMD* expression was 90 months compared with a median that was not reached in patients with low *DMD* expression (Nikitin *et al.*, 2007).

1.2.2.4 Carcinoma's

The association of *DMD* with various carcinomas has been reported with most evidence being accumulated for lung adenocarcinoma, gastric carcinoma and carcinomas of the head and neck. In Lung adenocarcinoma Tan *et al* explored the use of cell line shRNA mediated knockdown of Dp71 in A549 cells which led to reduced growth, migration an invasion capacity compared with controls functional assays (Sichuang Tan *et al.*, 2016). These knockdown cells were more chemosensitive to cisplatin mediated apoptosis and had enhanced caspase activity. Transplanting the cells into a nude xenograft model with Dp71 depletion, led to reduced tumour growth, compared with controls and reduced expression of Lamin B1, Bcl-2 and MMP2 proteins.

In gastric carcinoma, Dp71 may play a tumour suppressor role as immunohistochemistry was used to show that cancer cell differentiation ($p=0.001$) and lymphovascular invasion ($p=0.041$) were associated with downregulation of the Dp71 (Sipin Tan *et al.*, 2016). Patients with high Dp71 expression had a favourable overall survival outcome compared with patients with low Dp71 expression suggesting Dp71 may act as a tumour suppressor in this context. The authors also overexpressed Dp71d and Dp71f in gastric cell lines and this inhibited proliferation compared with controls cells. Using pull-down experiments, Dp71 interacted with Lamin B1 in normal gastric epithelial cells suggesting a significant role for Dp71 in proliferation and Lamin-B complex formation.

In the EBV associated tumour nasopharyngeal carcinoma an X-chromosome wide SNP based study identified a strong signal within the *DMD* gene (intronic SNP rs5927056) validated in replication cohorts (Zuo *et al.*, 2019). This intronic SNP was associated with reduced risk of NPC. Finally, in other tumour types *DMD* has been associated with neoplasia. This includes breast and uterine cancer patients where *DMD* alterations had significantly poorer survival outcomes (Luce *et al.*, 2017).

1.3 The Cancer Genome Atlas (TCGA) as a repository for mining cancer genomics data

The Cancer Genome Atlas (TCGA) repository represents one of the most valuable publically available data sources for cancer scientists comprised of sequence, expression, single nucleotide, copy number, and methylation data from 11K cancers across 33 major types (Figure 1.8). Accompanying this molecular data, is extensive clinical data for each cancer type including survival data which can be joined together to provide a powerful tool for cancer interrogation. A series of landmark papers have been published by the TCGA research network and continue to be updated (National Cancer Institute, 2022).

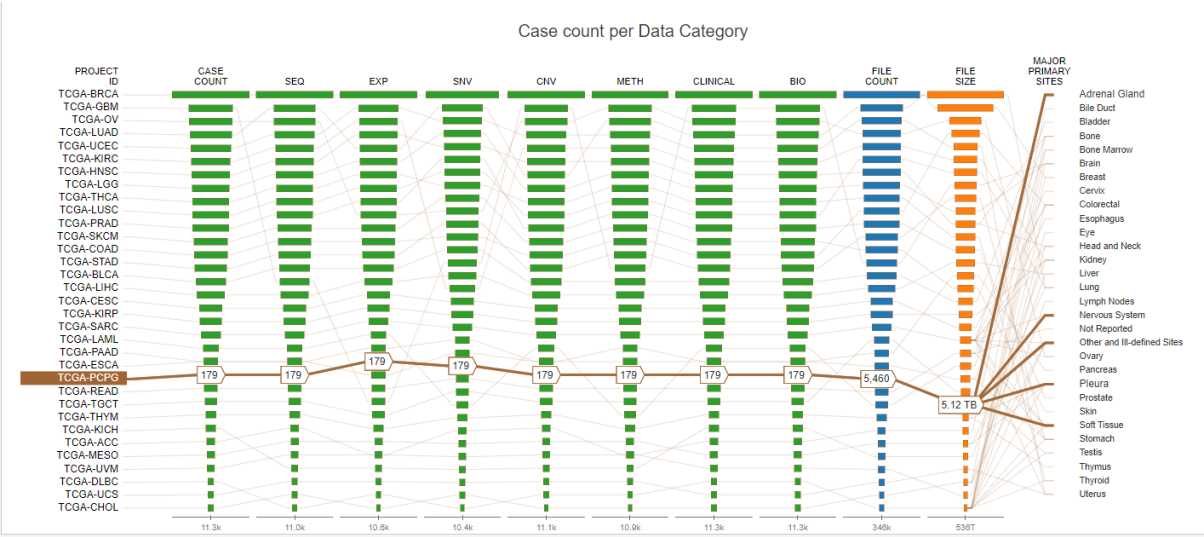


Figure 1.8 Overview of TCGA cases including clinical and molecular data.

Sequence (SEQ), RNAseq and microarray expression (EXP), single nucleotide variation (SNV), Copy number variation (CNV), methylation (METH) clinical and Biosample (BIO) sample are available across 11.3K cases (33 tumour types). Expression data was available from 10.6K cases.

Recently, the pipelines for TCGA bioinformatic analysis have been harmonised across clinical studies at the Genomic Data Commons (GDC) (Figure 1.9). Raw sequence FASTQ or BAM files originally mapped to the GRCh37/hg19 (legacy) assembly were re-mapped to an updated GRCh38/hg38 reference genome assembly (Gao *et al.*, 2019). mRNA expression data from these studies was originally derived from polyA+ RNA that was sequenced using Illumina NGS instruments and sequencing chemistry kits that evolved over time. As discussed by others (Gao *et al.*, 2019), the bioinformatic workflow for generating GRCh38/hg38 RNAseq data at the GDC is considerably different from that used to generate the earlier GRCh37/hg19 RNAseq data in TCGA. In addition, recent release updates have also occurred (March 2022) at GDC with differences in alignment, expression quantification, normalisation, and reference assemblies, potentially giving different results in abundance estimates.

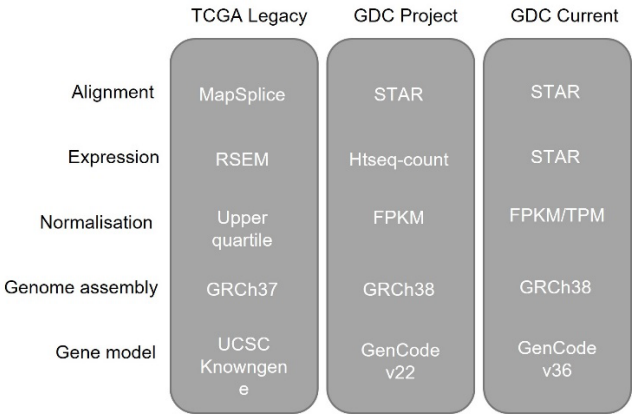


Figure 1.9 mRNA-Seq processing and data comparison in TCGA Legacy and the GDC
Three bioinformatic pipelines have been used to derive gene or isoform level abundance estimates for this project. Legacy isoform level data (**Left**) is derived from the TCGA Legacy (GRCh37/hg19) pipeline with output files deposited in the GDC legacy archive. The GDC project pipeline (**Middle**) was used to obtain total *DMD* gene expression data. During the project, the GDC project pipeline was superseded by the GDC current pipeline (hg38) (**Right**). All aspects of sample processing differ including computational methods, the reference genome, and the reference transcriptome (adapted from (Gao *et al.*, 2019).

1.4 Survival analysis and Cox proportional hazard modelling

As much of the analysis presented in this thesis is related to patient survival data (as an outcome measure), a basic overview of survival analysis is warranted. For further information on these topics, the reader is referred to some excellent articles (Clark *et al.*, 2003; Bradburn *et al.*, 2003b; Bradburn *et al.*, 2003a).

Survival analysis considers the time between a starting point (e.g. either the date of diagnosis or start of treatment) and the event of interest (e.g. death). A caveat is that for some patients the event (death) may not have occurred by the end of the study period and so their 'time-to-event' cannot be determined. Therefore specific methods are required to deal with this. The Kaplan-Meier method can be used to estimate the survival probability (Kaplan and Meier, 1958).

Visually, this function can be presented as Kaplan-Meier survival curve (a plot of the survival probability against time), and it indicates the probability of the event (for example, survival) at specific time points. Summarising the data in this way also allows facile estimation of median survival times. For example, in a hypothetical cohort (Figure 1.10), at 0.2 months, all patients are alive, although one patient has undergone censoring which can occur for the following reasons: (a) the patient does not experience the event of interest (death) for the duration of the study (b) the patient was lost to follow-up during the study (c) a different event occurs preventing further follow-up. When an individual patient event occurs beyond the study period, this censoring is described as right censoring. 50% of the patients had died at 0.7 years and only 22% of patients were alive at 1 year (Figure 1.10).

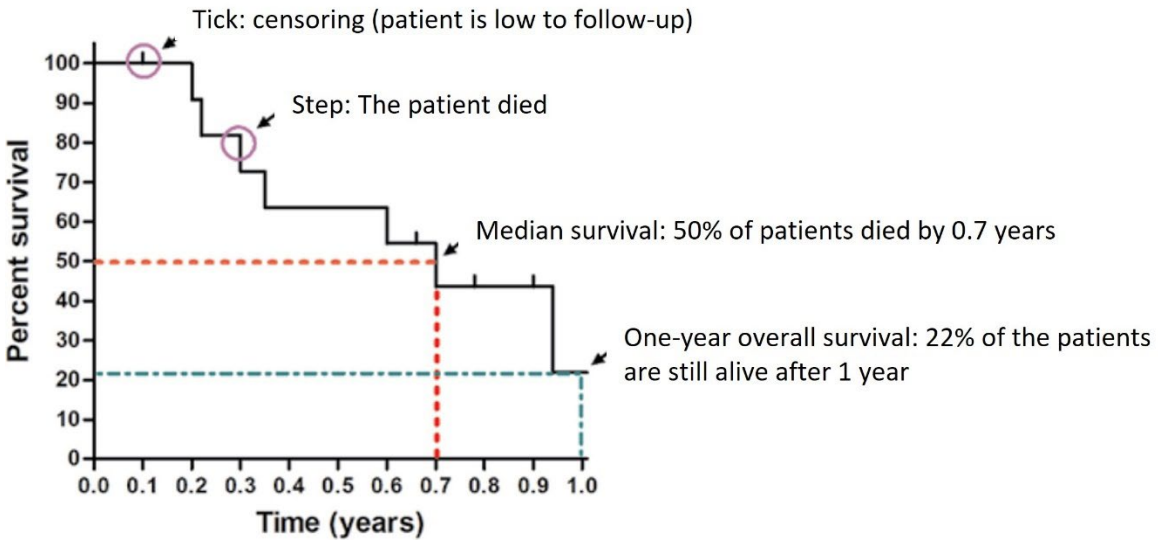


Figure 1.10 Interpretation of survival curves
Survival curve of patients over 1 year. 4 patients are censored with an unknown time to event for them. Median survival of the cohort is 0.7 years determined by extrapolation from 50% survival. At one year only 22 of patients are still alive. Modified from Van den Reek (van den Reek *et al.*, 2015)

If two or more survival curves are presented, they can be compared using the log-rank test, which is a non-parametric test that is widely used to compare survival curves. Each curve represents a group of patients (e.g. placebo vs. control) and the method calculates the expected number of events (if the Null hypothesis were true), since the previous event and compares it with observed number in each group. This is done for each event time, and for each group and sums them calculating the following chi-square test statistic for which an associated p-value can be computed:

$$X^2 = \sum_{i=1}^{g} \frac{(O_i - E_i)^2}{E_i}.$$

Figure 1.11 The log-rank chi-square statistic comparing survival curves
O_i represents observed events in group i, E_i is the expected events in group i and g is the number of groups. P- values are computed from the chi-square distribution.

When comparing only two groups, the log rank test (Peto *et al.*, 1977) can be used to test whether there is a difference between the survival times of separate groups, but it does not allow other explanatory variables to be considered. To achieve this and quantify the contribution of multiple covariates, Cox modelling is a helpful semi-parametric approach that can be used to fit univariate and multivariate regression models that have survival outcomes. The regression uses:

$$h(t)=h_0(t) \times \exp\{\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p\}$$

The hazard function h(t) is determined by p covariates (x1, x2, …, xp). Their impact is determined by the size of their coefficients (β1, β2, …, βp). The term h0 represents the baseline hazard when the value of all covariates x_i are equal to zero. The *(t)* in *h(t)* indicates that the hazard varies over time (but is proportional).

This equation can then be rearranged first dividing by the baseline hazard and then taking the natural logarithm of both sides. The following form is obtained, with example covariates substituted in:

1. *ln(h(t)/h0(t)) = (Sex[Male] * β1) +(Smoking[Yes]) * β2)+ (Chemotherapy[Yes]) * β3)+….*

If the parameter estimates are exponentiated for the given predictor variable, the quantities $\exp(\beta_p)$ are obtained:

2. $(h(t)/h0(t)) = \exp(\beta 1)^{(Sex[Male]} * \exp(\beta 2)^{(SmokingYes]} * \exp(\beta 3)^{(Chemotherapy[Yes]} *…$

These are the hazard ratios (i.e. $\exp(\beta 1)^{(Sex[Male]}$). A hazard ratio of 1 indicates no difference in survival between the groups. Greater than 1 means an increase in the event probability (death). Less than 1 represents a reduction in the hazard and chance of death.

Hazard ratios represent the multiplicative effect that a given covariate (e.g. patient sex) has on the outcome. If a particular covariate has a hazard ratio of 2, then an increase of 1 in that covariates value will double the hazard rate across all time points. Cox proportional hazard modelling assumes there is non-informative censoring, and the hazards are proportional.

With suitable background provided on the biology of *DMD*, it's putative association with cancer and the nature/provenance of the gene expression and clinical data analysed during this project, the aim and objectives of the project can be defined.

1.5 Aims and objectives

Aim

Identify how *DMD* and *DMD* gene variants are associated with survival outcomes for individuals with cancer.

Objectives

1. To identify whether total *DMD* gene expression is associated with survival outcomes using Kaplan-Meier survival analysis across 33 tumour types.
2. To model hazard ratios of associated DAPC genes across TCGA cancers
3. To identify expression patterns of specific *DMD* gene variants in cancers and determine their association with survival outcomes.
4. To identify transcriptome wide differentially expressed genes in *DMD* high vs. low expressing cases to identify biological pathways dysregulated in specific cancer types by *DMD*.

Chapter 2

2. Materials and methods

2.1 Clinical sample data and ethical approval

Project approval was obtained from the University of Nottingham project approval committee. All data were downloaded from the Cancer Genome Atlas (TCGA); therefore no ethical approval was needed. Full details of ethics and policies associated with TCGA is described elsewhere (NCI, n.d.). Genomic analysis was done on a data set of molecular and clinical information from over 10,000 tumours representing 33 types of cancer (Weinstein *et al.*, 2013). Normalised RNA-seq gene expression and clinical data from these tumours were collected and analysed using a bespoke bioinformatic workflow (Figure 2.1). The acronyms for each cancer type analysed in this study are detailed and used throughout (Table 2.1).

Table 2.1 Definition of TCGA cancer types abbreviations.

Abbreviation	Type of cancer
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
ESCA	Oesophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute Myeloid Leukaemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma

OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular Germ Cell Tumours
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UVM	Uveal Melanoma

2.2 Overview of workflow for *DMD* and gene variant survival analysis

To analyse the association of *DMD* and variant gene expression with patient survival outcomes, RNAseq data from 33 tumour types from the genomic data Commons (GDC) was imported into Rstudio using an R TCGAbiolinks library (Colaprico *et al.*, 2016). Associated harmonised clinical data was imported into Rstudio. *DMD* RNAseq gene expression was linked to patient clinical data to perform survival analysis. Patients were split into high and low expressing *DMD* RNA expressing groups based on cutpoint selection using Maxstat (Lausen and Schumacher, 1992) . Kaplan-Meier survival analysis was used to explore overall survival outcomes in these two patient groups. Hazard ratios were calculated based on univariate analysis. Pathway analysis with iDEP identified putative functional pathways based on differential gene expression between high and low *DMD* expressing tumours.

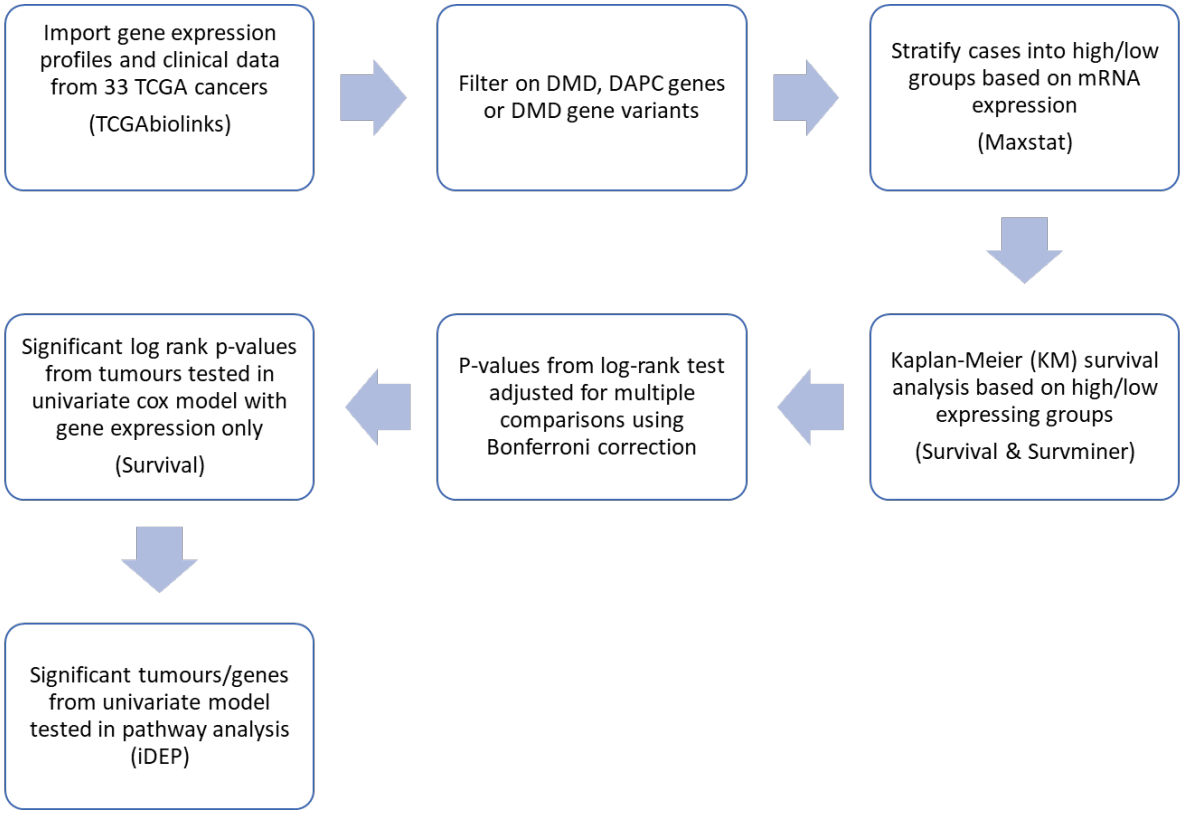


Figure 2.1 Bioinformatic workflow
Workflow for analysing *DMD*, DAPC and *DMD* gene variants and their association with overall survival outcomes in cancer and subsequent DEG analysis. Software packages and R libraries are indicated in brackets.

2.2.1 Importing normalised gene expression and associated clinical data from TCGA

Although largely concordant, it is important to recognise the origin of the types of normalised data used in this project. For isoform level data, TCGA legacy sequence data (hg19) was originally aligned using MapSplice, with translation of co-ordinates using UCSC KnownGene. Expression was quantified with RSEM, and raw counts were normalised to fixed upper quartile values (500 for isoform estimates). Upper Quartile (UQ) normalisation methods remove genes that have zero read counts across all samples and the remaining gene counts are scaled by the upper quartile of the count distribution of the sample and multiplied by the mean upper quartile across all samples (Abbas-Aghababazadeh *et al.*, 2018).

Output files for total *DMD* expression were derived from a GDC workflow that used legacy BAM files which were reformatted as FASTQs using biobambam. These were then re-aligned to the hg38 genome assembly using the STAR 2-pass approach (Dobin and Gingeras, 2015). The Gencode v22 transcriptome definition was then quantified using htseq-count procedure within samtools. Raw counts, FPKM, and upper quartile normalized FPKM estimates are provided (Figure 2.2). During the project (March 2022) GDC updated their pipeline using Gencode v36 transcriptome annotation and used STAR for both alignment and raw count production. RNA-Seq STAR-Counts output files from GDC now contain additionally not only FPKM, FPKM-UQ but also TPM normalised abundance values.

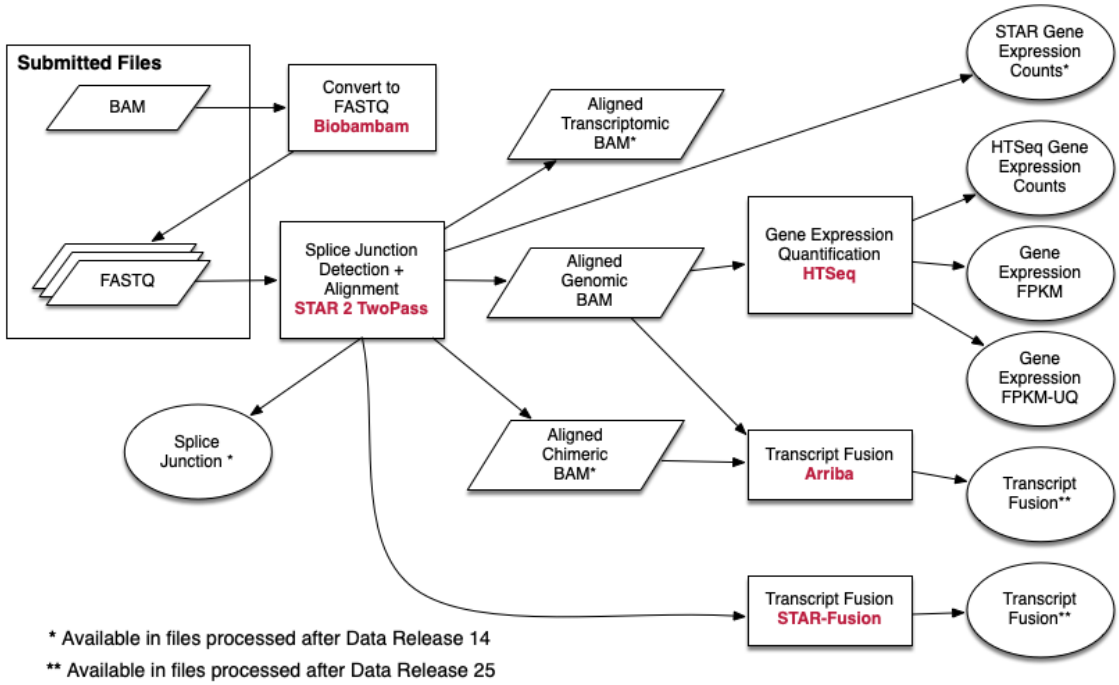


Figure 2.2 Bioinformatic pipeline used by GDC to generate gene expression data
Overview of how submitted BAM and FASTQ files are processed by the GDC to produce output files for users (GDC, 2022) https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/images/gene-expression-quantification-pipeline-v3.png)

FPKM-UQ was applied to gene-level read counts that are produced by HTSeq and generated using custom scripts. The formula used to generate FPKM-UQ values is as follows:

$FPKM = [RM_g * 10^{9}] / [RM_{75} * L]$

- RM_g: The number of reads mapped to the gene
- RM_{75}: The number of read mapped to the 75th percentile gene in the alignment.
- L: The length of the gene in base pairs

As an example of how FPKM-UQ normalisation works consider the following example:

Sample 1: Gene A

- Gene length: 3,000 bp
- 1,000 reads mapped to Gene A
- 1,000,000 reads mapped to all protein-coding regions
- 2,000 reads mapped in upper quartile in sample 1

FPKM for Gene A = $(1,000)*(10^9)/[(3,000)*(1,000,000)] = 333.33$

FPKM-UQ for Gene A = $(1,000)*(10^9)/[(3,000)*(2,000)] = 166,666.67$

For this study normalised FPKM-UQ RNA-seq data was extracted from GDC. FPKM-UQ files were available as tab delimited files with the Ensembl gene IDs in the first column and the expression values in the second. The R/Bioconductor package TCGAbiolinks (Colaprico *et al.*, 2016) version 2.24 was used employing GDCquery(), GDCdownload() and GDCprepare() functions and using *data.catagory as "Transcriptome profiling", data.type* as "Gene Expression quantification" and *workflow.type* as "HTSeq – FPKM-UQ". Later during pre-processing non-primary tumours were filtered out (see FINALscript.R code – section 2.5 for

code availability). Please note, GDC altered their pipeline for data extraction during the project but after data extraction and analysis (see updated STAR_script_edited.R code – section 2.5).

2.2.2 Dichotomisation of patient tumours into high and low expressers

High versus low *DMD* expressing patients were dichotomised using cutpoint selection using the R library Maxstat. Maxstat uses maximally selected rank statistics (smethod=LogRank) to evaluate a simple estimated cutpoint. Simulation used conditional Monte-Carlo with B = 9999 replications.

2.2.3 Kaplan-Meier survival analysis *DMD* high and low expressing tumour groups

The R library package Survival (Therneau, 2021) was used to do Kaplan-Meier survival analysis. The function survfit() was used to compute Kaplan-Meier survival estimates. Survdiff() was used to compute the log-rank test comparing the two survival curves. P-values were adjusted for multiple testing using Bonferroni correction. The function ggsurvplot(), in the R package Survminer was used to produce the survival curves for the two groups of subject.

2.2.4 Univariate hazard modelling

Tumours where *DMD* or DAPC genes with high or low expression gave significant differences in the Kaplan Meier analysis were analysed for the proportional hazard using Cox modelling. For the univariate analysis, gene expression alone was used as a covariate. The function coxph()[in the Survival package] was used to compute the Cox proportional hazards regression model in R.

2.2.5 Cluster analysis

Agglomerative hierarchical clustering analysis was performed using the hclust() function [in stats package] using a Euclidean distance matrix with a Ward D2 minimum variance linkage method which minimises the total within-cluster variance. For each step, the pair of clusters with minimum between-cluster distance are merged. Enhanced visualisation was provided in ggplot using the fviz_dend() function.

2.3 Isoform expression analysis

As the current GDC pipeline does not have isoform level data for protein coding genes, isoform expression data from the GDC legacy archive data was extracted using the R/Bioconductor package TCGAbiolinks (Colaprico *et al.*, 2016) version 2.24 using GDCquery(), GDCdownload() and GDCprepare() functions for primary tumour *samples.types* as well as using *data.type* as "Isoform expression quantification" and *file.type* as "normalized" (see isoform_GDC.R code – section 2.5 for code availability). This pipeline used MapSplice (Wang *et al*., 2010) to do the alignment and RSEM to perform the quantification (Li and Dewey, 2011). Output files contained UCSC isoform identifiers and Table 2.2 (curated from the UCSC Table browser) was used to convert them to specific *DMD* gene products for processing and survival analysis.

Table 2.2 Details of major *DMD* gene products including ids, genome and coding sequence start and end positions and exons counts

DMD variant	exonCount	UCSC ID	Ensembl ID	chrom	strand	cdsStart	cdsEnd	proteinID	RefSeqID
Dp40	13	uc011mkb.1	ENST00000378723	chrX	-	31139949	31284946	B4DSV7	NM_004019
Dp71ab	16	uc004dcp.1	ENST00000378723	chrX	-	31139949	31284946	P11532-5	NM_004018
Dp71a	17	uc004dcn.1	ENST00000378723	chrX	-	31140035	31284946	NP_004008	NM_004017
Dp71b	17	uc004dco.1	ENST00000378723	chrX	-	31139949	31284946	P11532-6	NM_004016
Dp71	18	uc004dcm.1	ENST00000378723	chrX	-	31140035	31284946	E9PDN1	NM_004015
Dp116	25	uc004dcq.1	ENST00000378707	chrX	-	31140035	31526354	NP_004005	NM_004014
Dp140bc	31	uc004dcs.1	ENST00000343523	chrX	-	31139949	31792238	NP_004014	NM_004023
Dp140c	32	uc004dcr.1	ENST00000541735	chrX	-	31140035	31792238	NP_004011	NM_004020
DpD140ab	34	uc004dcv.1	ENST00000359836	chrX	-	31139949	31792238	NP_004013	NM_004022
Dp140b	35	uc004dcu.1	ENST00000378707	chrX	-	31139949	31792238	A7E212	NM_004021
Dp140	36	uc004dct.1	ENST00000378707	chrX	-	31140035	31792238	A1L0U9	NM_004013
Dp260-1	51	uc004dcx.2	ENST00000378677	chrX	-	31140035	32430326	NP_004002	NM_004011
Dp260-2	51	uc004dcw.2	ENST00000378677	chrX	-	31140035	32430174	NP_004003	NM_004012
Dp427c	79	uc004ddb.1	ENST00000378677	chrX	-	31140035	33357382	NP_000100	NM_000109
Dp427m	79	uc004dda.1	ENST00000357033	chrX	-	31140035	33229429	P11532	NM_004006
Dp427p1	79	uc004dcy.1	ENST00000378677	chrX	-	31140035	33146282	P11532-4	NM_004009
Dp427p2	79	uc004dcz.2	ENST00000378677	chrX	-	31140035	32834745	NP_004001	NM_004010

2.4 Differential gene expression and pathway analysis

RNA-seq data is a powerful tool for transcriptome profiling of tumours. However, exploratory analysis, differential expression and subsequent pathway analysis can be complicated. iDEP connects 63 R/Bioconductor packages to provide workflows (Figure 2.3) that enable cancer biologists to leverage gene expression data into functional experiments (Ge *et al.*, 2018).

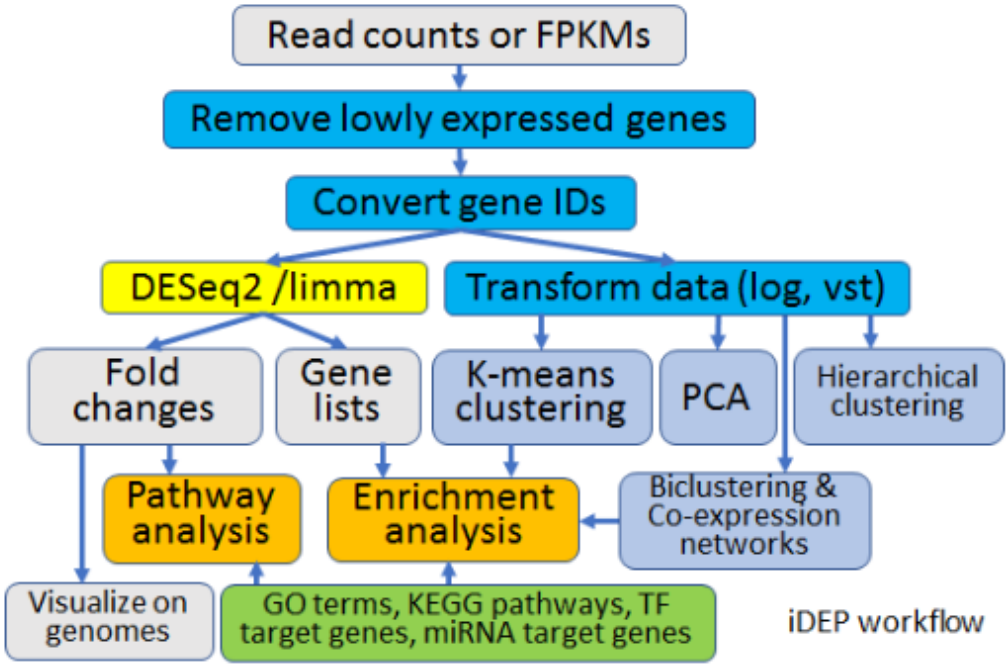


Figure 2.3 iDEP: Integrated Differential Expression and Pathway analysis
iDEP is an integrated web application for differential expression and pathway analysis of RNA-Seq data iDEP workflow and functional modules (Ge *et al.*, 2018)

DEGs were identified using the Limma package (false discovery rate (FDR) cut-off of 0.1 and a minimum fold-change of 2). Functional enrichment analysis of DEGs was performed in iDEP using gene ontology (GO) biological processes. Enrichment trees and networks were generated in iDEP. Protein-protein interaction (PPI) networks among top DEGs were retrieved via an API access to the STRING database.

2.5 Software, code, and data availability

All code for this project is available at https://github.com/lrmacha/TCGA All analyses used R Statistical Software (v4.0.3; R Core Team 2020-10-10), GraphPad Prism 8 and Microsoft Excel. Gene expression data analysed during this study are publicly available in the repository https://portal.gdc.cancer.gov/ and can be downloaded directly by using the TCGAbiolinks R package as described above.

Chapter 3

3. A global analysis of *DMD* gene expression across TCGA tumours.

3.1 Introduction

Currently the association of *DMD* gene expression with survival outcomes across cancer types is unclear. To explore this, existing libraries in R were employed to extract TCGA clinical cancer and RNA-seq genomics data. To do survival analysis, continuous gene expression data was dichotomised into high and low expressing cases. To achieve this, the R package Maxstat was employed to achieve cutpoint selection and allow recoding of cases into high vs. low *DMD* expressing groups. The CRAN 'Survival' package (Therneau, 2021) facilitated Kaplan–Meier survival analysis (including associated log-rank tests) and allowed the production of summary survival statistics.

3.2 Results

3.2.1 Estimating cutpoint values with Maxstat

The GDC data portal contains RNA-seq gene expression sets for 33 different tumour types (Table 2.1) for download and downstream analysis. Upon download, gene expression data was filtered on the gene of interest (e.g. total *DMD*) and concatenated with matched clinical data. Overall survival, status and gene expression values were used to derive cutpoint values for recoding cases into high or low *DMD* expressing cases.

To dichotomise continuous gene expression data for survival analysis, optimal cutpoint values were obtained based on the use of total *DMD* gene expression, overall survival and status and computed using maximally selected log-rank statistics implemented in Maxstat (Figure 3.1). The output of the *Maxstat.test* function provided a log-rank statistic M and p-value by conditional Monte-Carlo replication providing an estimated cutpoint value.

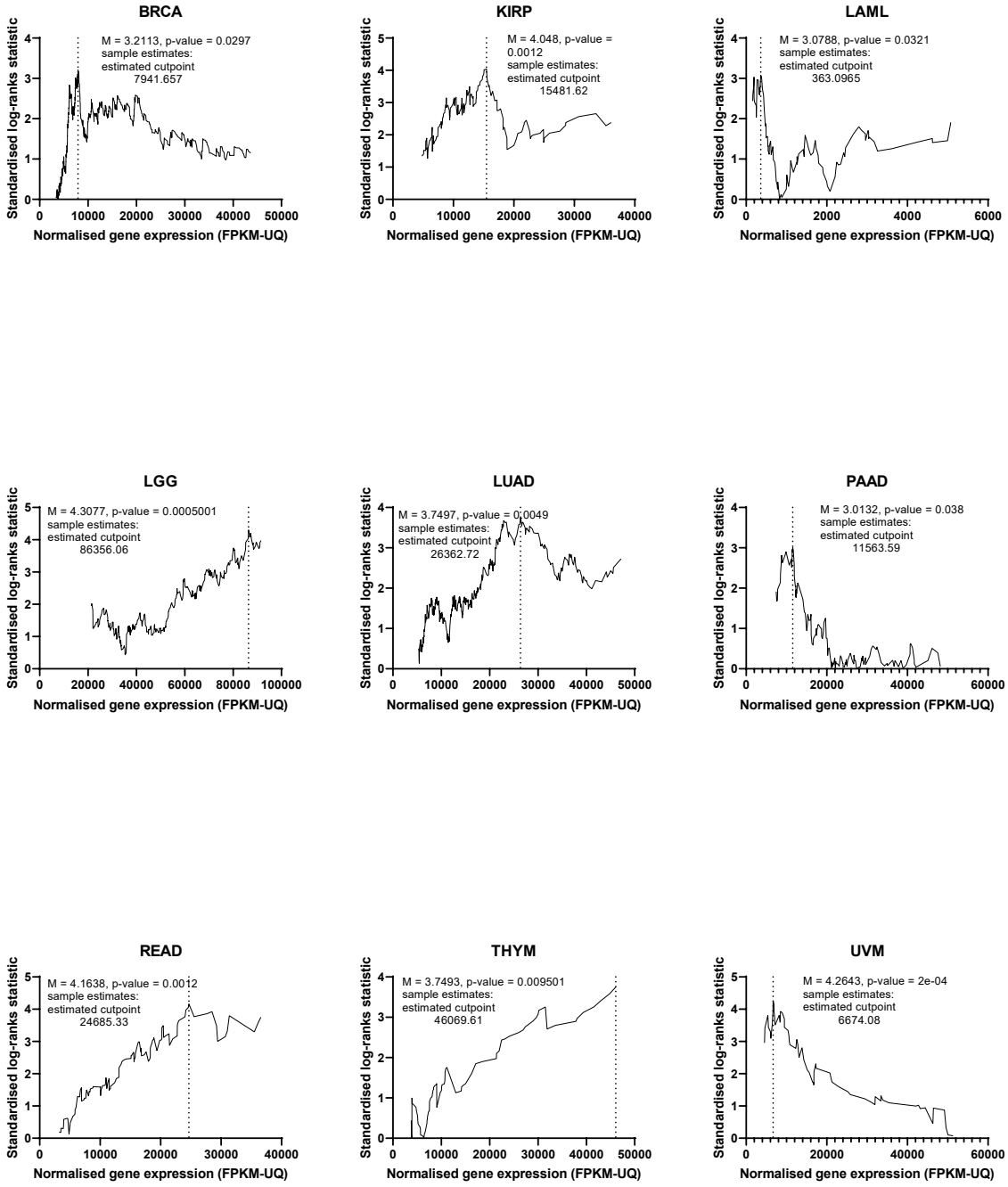


Figure 3.1 Estimated cutpoints (dashed vertical line) for total mRNA expression in selected TCGA cancers

See Table 2.1 for abbreviations. Based on standardised log-rank statistics. Exact conditional p-values were simulated via conditional Monte-Carlo. M = maximum of the log-rank statistics.

As an example, for Breast invasive carcinoma (BRCA), the estimated cutpoint was 7942 FPKM-UQ, and the maximum of the log-rank statistics is M = 3.2113. The probability that, under the null hypothesis, the maximally selected log-rank statistic is greater M = 3.2113 is less than 0.0297. Therefore, BRCA cases with *DMD* gene expression values above 7642 were recoded as high expressing cases and those below 7942 were low expressing cases.

3.2.1 Kaplan Meier survival analysis of TCGA cancers

Applying the above discussed cutpoint approach, survival analysis was done on 33 TCGA tumours dichotomising patients into high or low expressing groups (see appendices). Of the 33 tumour types examined, nine had significant differences in survival outcomes (Log-Rank test) after Bonferroni correction (Figure 3.2). These included BRCA (p=0.0021), KIRP (p<0.001), LAML (p=0.0048), LGG (p<0.0001), LUAD (p= 0.0003), PAAD (p=0.0008), READ (p<0.0001), THYM (p<0.0001) and UVM (p<0.0001). For BRCA, LAML, LUAD, PAAD and UVM patient overall survival was better in those patients with high total *DMD* tumour RNA expression. In KIRP, LGG, READ and THYM high expression of total *DMD* was associated with worse survival outcomes. As an example, LGG median survival of patients with high expression of *DMD* was 1120 days compared with patients with high tumour expression of *DMD* who lived for a median of 2875 days (2.57-fold, or a 4 years and 10 month increase in overall survival time). For KIRP (low), LUAD (high *DMD*), READ (high *DMD*), THYM (low *DMD*) and UVM (high *DMD*) median survival could not be calculated as it was greater than 50% at the last time point.

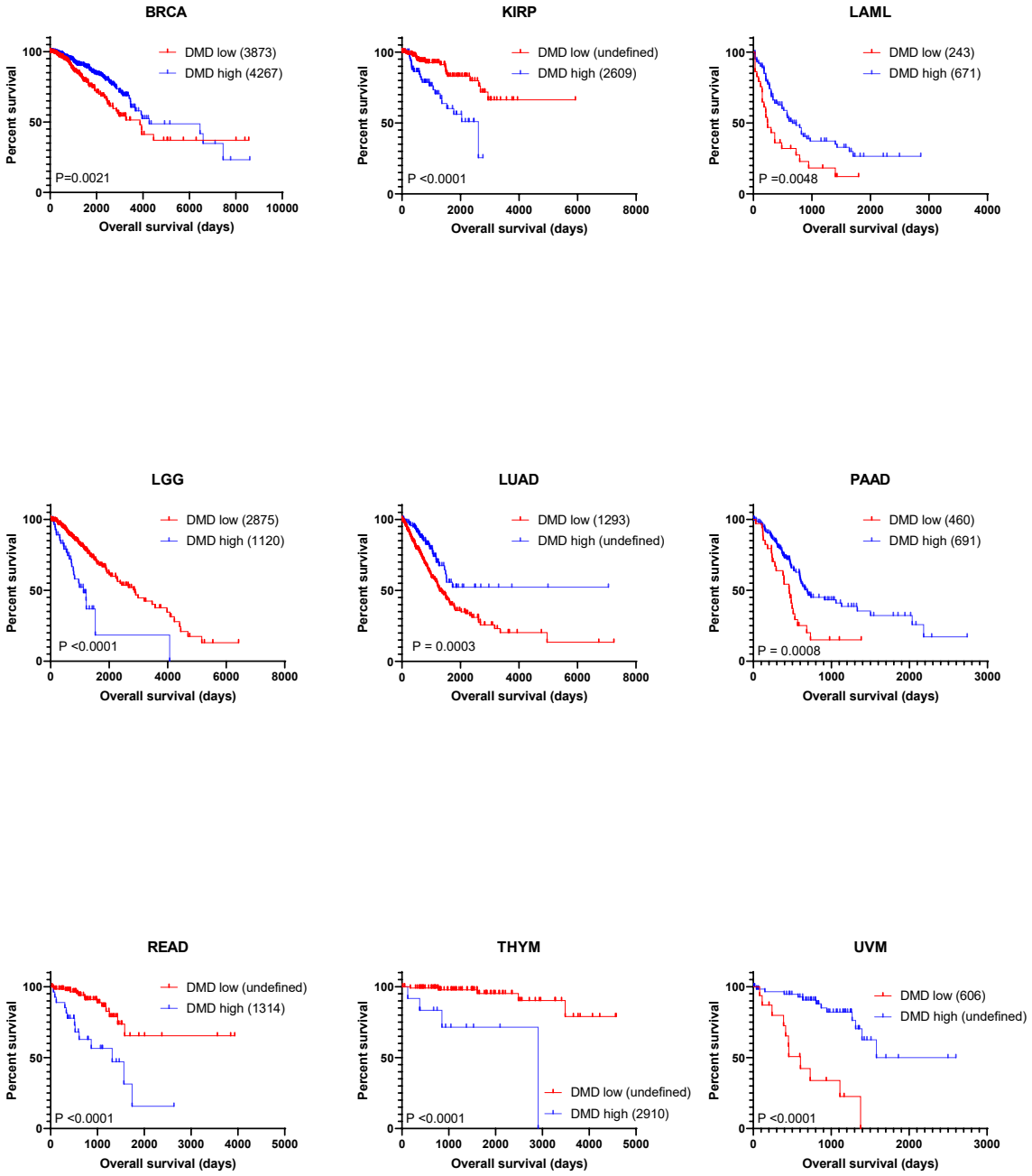


Figure 3.2 *DMD* expression is significantly associated with overall survival in nine specific tumour types

(a) TCGA RNAseq data from nine TCGA cancer cases were dichotomised into high (blue) and low (red) *DMD* expressing groups and survival analysis performed in GraphPad using the log-rank test. Numbers in brackets are median overall survival times in days.

3.2.2 Univariate Cox model analysis

Tumours significant in the Kaplan-Meier analysis were analysed in a Cox proportional hazard model (Figure 3.3) employing gene expression as the sole covariate. A hazard ratio (HR) > 1 means that high expression of that gene is associated with decreased survival. Conversely, a HR < 1 means that high expression of the gene is associated with increased survival (i.e. is protective). High expression of *DMD* in THYM (HR 11.8, 95% CI 2.9 to 47.8, p<0.001), READ (HR 4.19, 95% CI 1.96 to 8.94, p<0.001), LGG (HR 3.15, 95% CI 2.02 to 4.91, p<0.001), and KIRP (HR 3.44, 95% CI 1.87 to 6.33, p<0.001) was associated with increased risks of poor survival. Conversely, high expression of *DMD* in UVM (HR 0.14, 95% CI 0.06 to 0.33, p<0.001), PAAD (HR 0.46, 95% CI 0.29 to 0.73, p<0.001), LUAD (HR 0.50, 95% CI 0.34 to 0.73, p<0.001), LAML (HR 0.46, 95% CI 0.29 to 0.72, p<0.001) and BRCA (HR 0.59, 95% CI 0.43 to 0.81, p<0.001) was associated with protection (compared with high expressing cases). THYM had the highest risk of poor survival and UVM has the lowest risk of poor survival, though with wide confidence intervals in both cases.

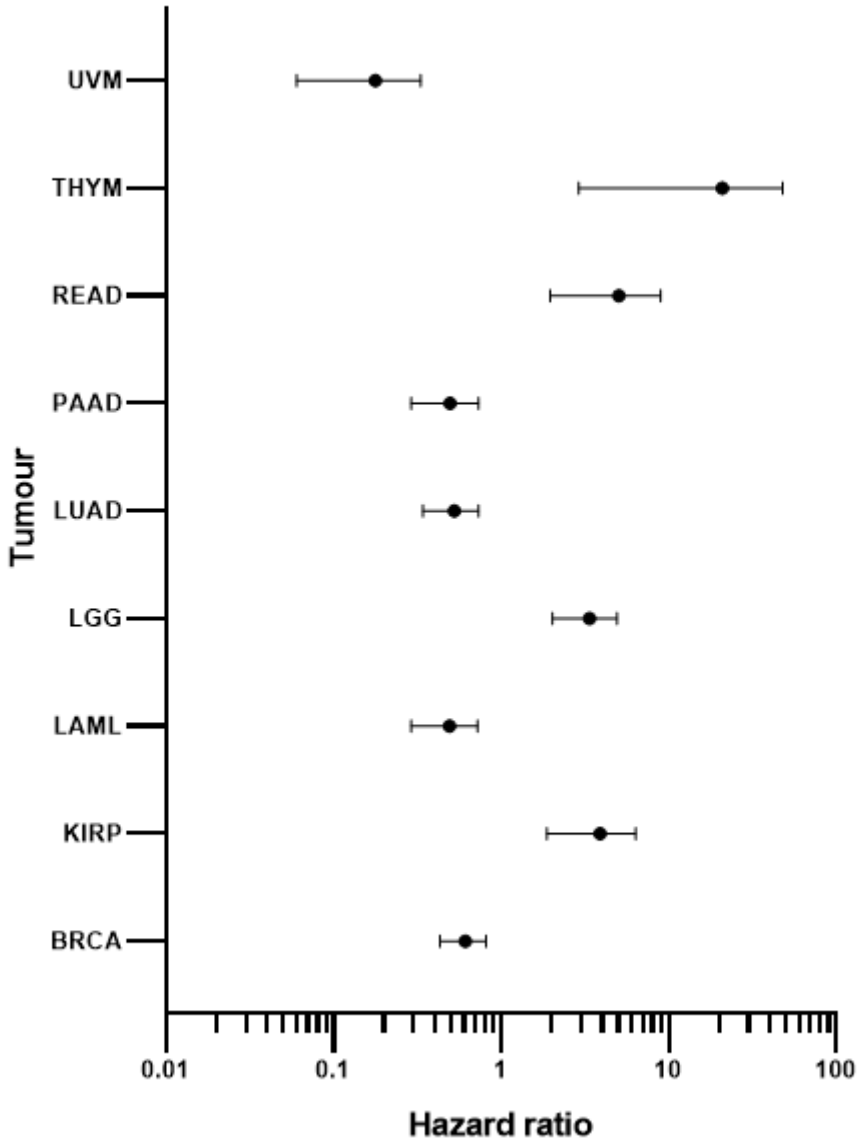


Figure 3.3 Hazard ratios in selected TCGA tumours
Forest plot revealing the log-rank hazard ratio with 95% confidence intervals. UVM n=80, THYM n=121, READ n=177, PAAD n = 182, LUAD n=594, LGG n= 529, LAML n= 151, KIRP n= 321, BRCA n=1222

3.2.3 Univariate DAPC gene hazard modelling and cluster analysis

Dystrophin (encoded by *DMD*) is a component of the Dystrophin associated protein complex (DAPC) and so it was of interest to explore whether other DAPC genes (Table 3.1) had a similar hazard ratio profile compared with *DMD* across the selected nine tumours (Figure 3.4).

Table 3.1 Genes and encoded protein products in the Dystrophin associated protein complex

Gene	Protein
DMD	Dystrophin
DAG1	Dystroglycan 1 (Alpha & Beta)
SGCA	Alpha Sarcoglycan
SGCB	Beta Sarcoglycan
SGCD	Delta Sarcoglycan
SGCE	Epsilon Sarcoglycan
SGCG	Gamma Sarcoglycan
SGCZ	Zeta Sarcoglycan
SSPN	Sarcospan
SNTA1	Syntrophin Alpha 1
SNTB1	Syntrophin Beta 1
SNTB2	Syntrophin Beta 2
DTNA	Dystrobrevin Alpha
DTNB	Dystrobrevin Beta
NOS1	Nitric oxide synthase 1 (nNOS)

Interestingly, of the nine tumours examined, only LGG had all DAPC genes providing statistically significant univariate hazard ratios (nine DAPC genes associated with increase in hazard, six DAPC genes associated with a decrease in hazard). No tumour type had all statistically significant hazard ratios trending in the same direction (i.e. all increase or decrease the hazard), however, high expression of DAPC genes in LUAD was protective for eight genes with only two increasing the hazard.

Interestingly, by doing hierarchical clustering analysis, hazard ratio data can be clustered by rows (TCGA cancer) and columns (DAPC genes) using a Ward D2 hierarchical clustering algorithm and a Euclidean distance as distance metric (Lawlor *et al.*, 2016). In clustering by

DAPC genes three clusters containing the fifteen genes are illustrated on the dendrogram (Figure 3.5A). The first cluster contained genes encoding two sarcoglycans (gamma and delta) and dystrobrevin beta, the second cluster contained sarcospan, dystrophin, nNOS, dystrobrevin alpha, sarcospan and dystroglycan. The third cluster contained sarcoglycans (alpha, beta, zeta, and epsilon) and alpha and beta syntrophins.

With clustering analysis grouped on tumours, three clusters containing the nine tumours are illustrated on the dendrogram (Figure 3.5B) which included two major clusters and a third cluster containing only THYM. The middle cluster (yellow) contained READ, KIRP and LGG and the largest cluster contained UVM, PAAD, LAML, BRCA and LUAD.

In summary, of the nine TCGA tumour types with significant survival differences between high and low *DMD* expressing tumours, three major disease clusters could be defined based on hazard ratios of *DMD* and 14 other DAPC-associated gene Hazard ratios.

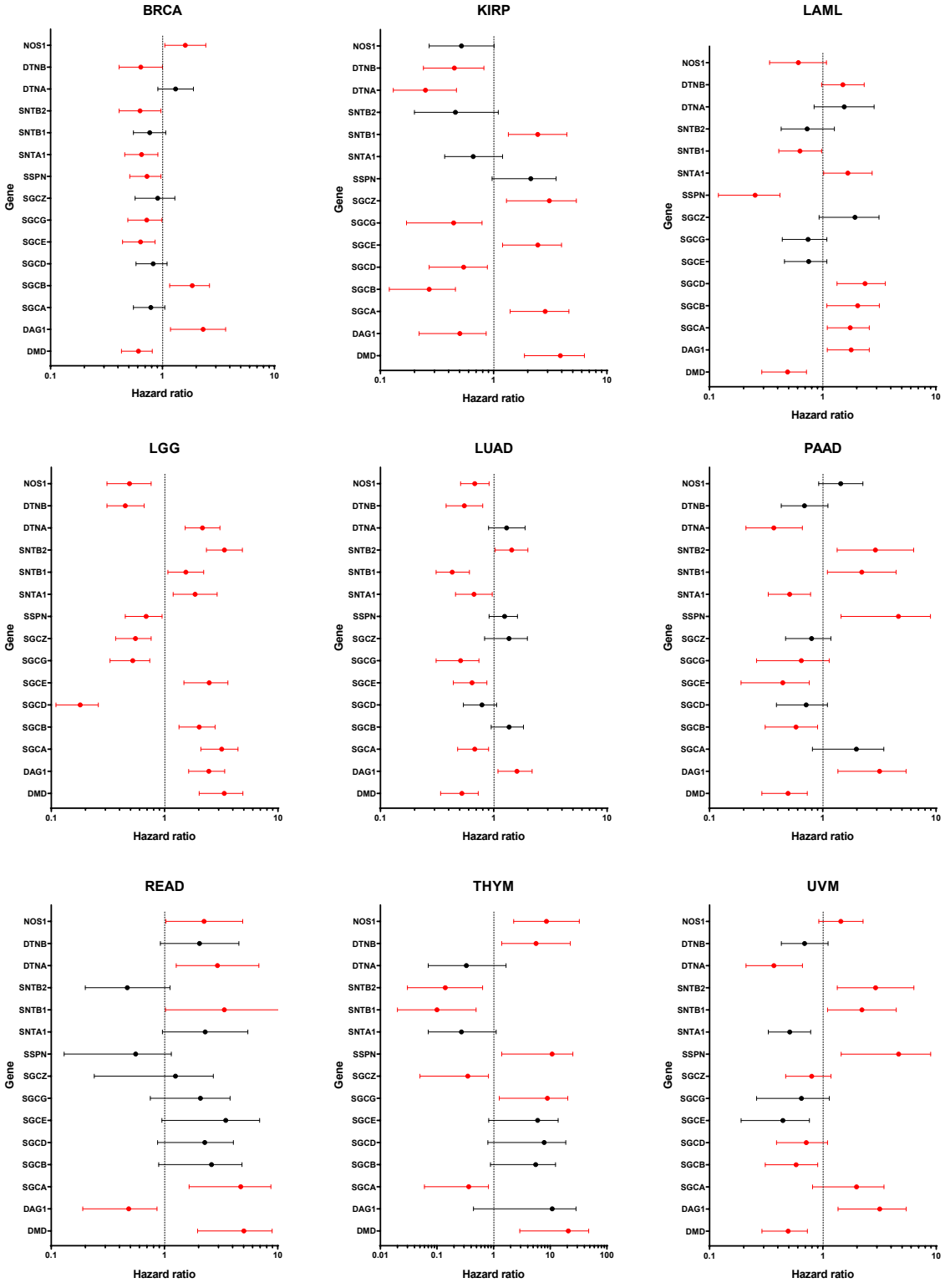


Figure 3.4 Association of DAPC gene expression with hazard ratios in selected TCGA tumours

Forest plots revealing the log-rank hazard ratio with 95% confidence intervals. Red bars are significant (alpha < 0.05) UVM n=80, THYM n=121, READ n=177, PAAD n = 182, LUAD n=594, LGG n= 529, LAML n= 151, KIRP n= 321, BRCA n=1222

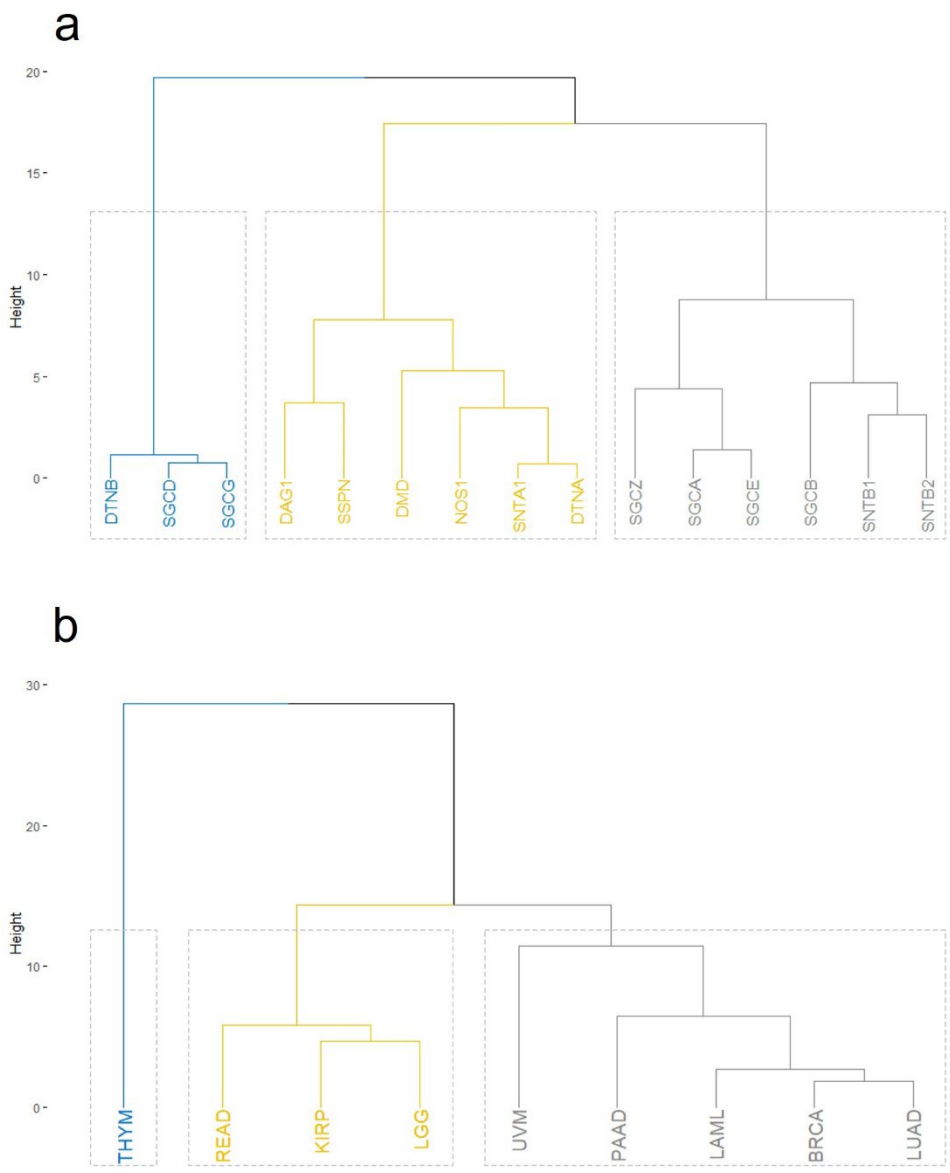


Figure 3.5 Cluster dendrograms based on hazard ratios

Identifying (**a)** DAPC-gene based dendrogram cluster analysis of tumours. (**b)** Tumour-based dendrogram cluster analysis of DAPC genes. Significant univariate Hazard ratio values were used. Three clusters were specified for both dendrograms, using Euclidean distance as a distance metric and the Ward D2 clustering algorithm.

3.2.4 Variant specific expression analysis across TCGA tumours

Most studies to date have considered dystrophin as a single protein and have not considered the complex diversity that arises from alternative splicing or promotor usage. For example, to date, 14 Dp71 isoforms have been identified (Naidoo and Anthony, 2020). The splicing of exons 71 and 78 are particularly important as their presence/absence determines subcellular localisation and function (Naidoo and Anthony, 2020). Dp71 is alternatively spliced to produce multiple isoforms, the isoforms lacking exon 71 (Dp71a, found exclusively in the nucleus) and exons 71 and 78 (Dp71ab, found exclusively in the cytoplasm) are the most predominant in soft tissue Sarcomas (STS) (Mauduit *et al.*, 2019). However, this pattern may be different in other cancers. Table 2.2 details the known correctly annotated *DMD* gene products which were analysed for expression in the nine TCGA tumours that had significant survival differences based on total DMD expression RNA levels (Figure 3.6). For LAML only Dp40 was weakly expressed. THYM had additional expression of Dp71ab. For BRCA, LUAD, PAAD and READ, Dp40, Dp71ab, Dp71b and Dp427m were expressed. KIRP and had a similar expression profile but lacked Dp427m. UVM had the additional expression of Dp260-1. The broadest expression of *DMD* gene products was observed in LGG with expression of eight different gene products (Dp40, Dp71ab, Dp71b, Dp71a, Dp71, Dp116, Dp140, Dp260-1 and Dp427m). Interestingly Dp40 was expressed in all tumours and at least one Dp71 gene product in all tumours except LAML. Hierarchical clustering into four groups based on Dp gene expression values confirms these observations (Figure 3.7)

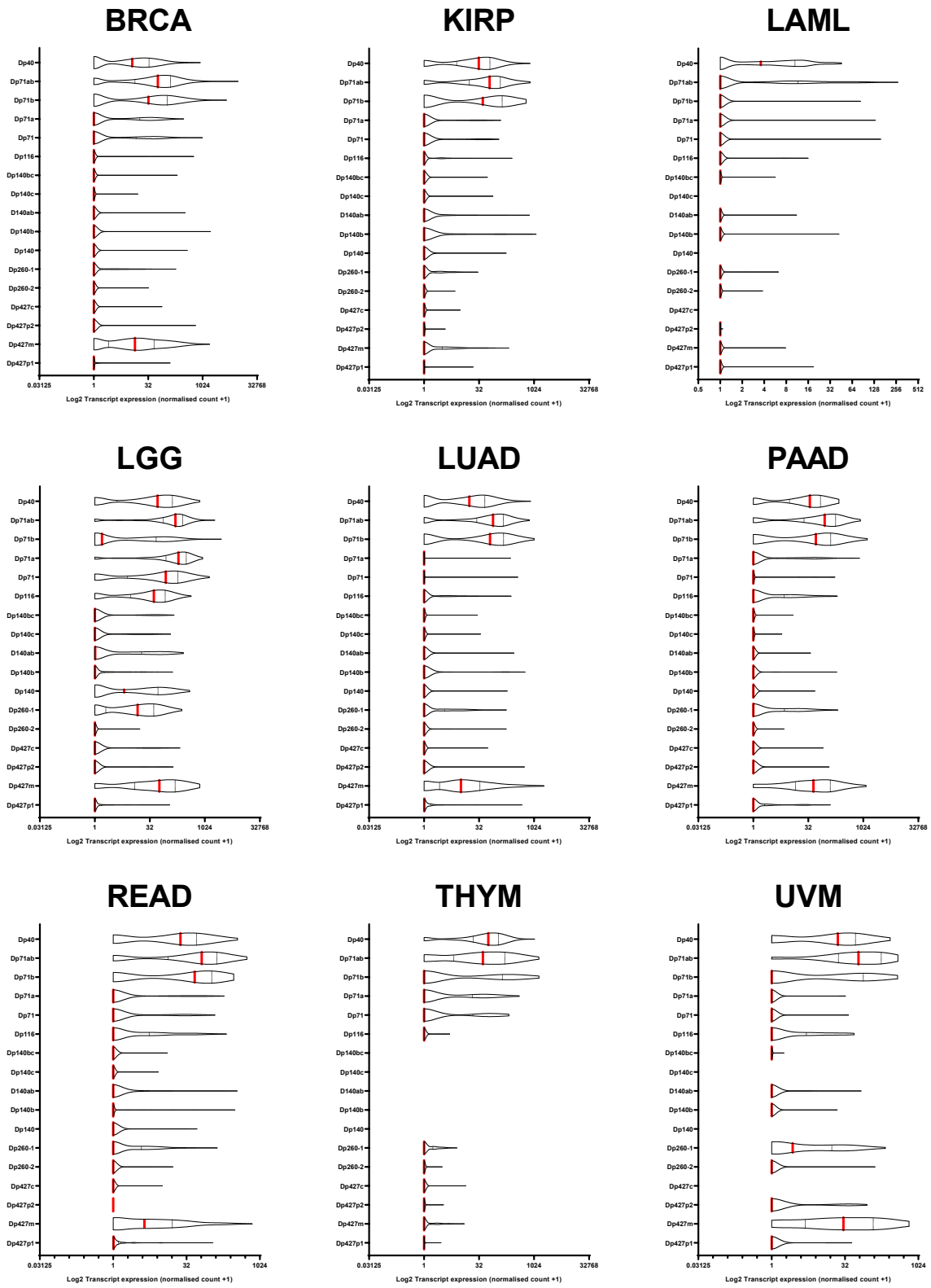


Figure 3.6 Expression of individual *DMD* gene products in selected TCGA cancers
Normalised counts were Log2 transformed +1. Red bars represent median values, dash lines represent 95% confidence intervals.

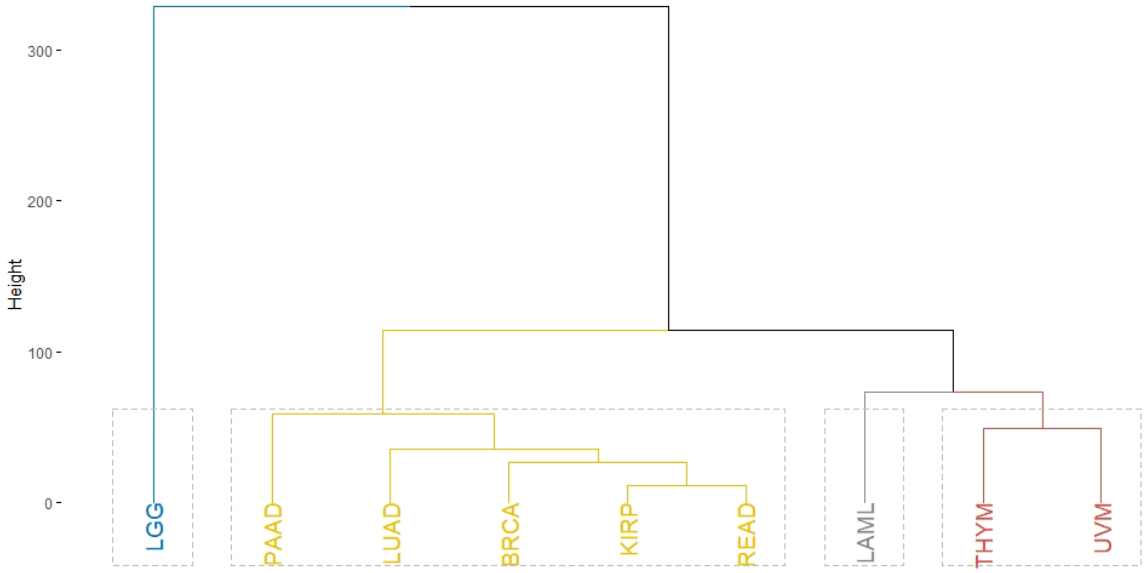


Figure 3.7 Cluster dendrograms based on Dp gene expression

A Dp gene variant based dendrogram cluster analysis of tumours. Gene expression values were used. Four clusters were specified for the dendrogram, using Euclidean distance as a distance metric and the Ward D2 clustering algorithm.

3.2.5 Variant specific survival analysis across TCGA cancers

Based on expression of specific gene products in individual cancers, BRCA was first examined to determine whether Dp40, Dp71ab, Dp71b and Dp427 were associated with overall survival (Figure 3.8). High expression of the Dp71ab (p=0.0032) and Dp71b (p=0.036) gene products were significantly associated with poor BRCA survival.

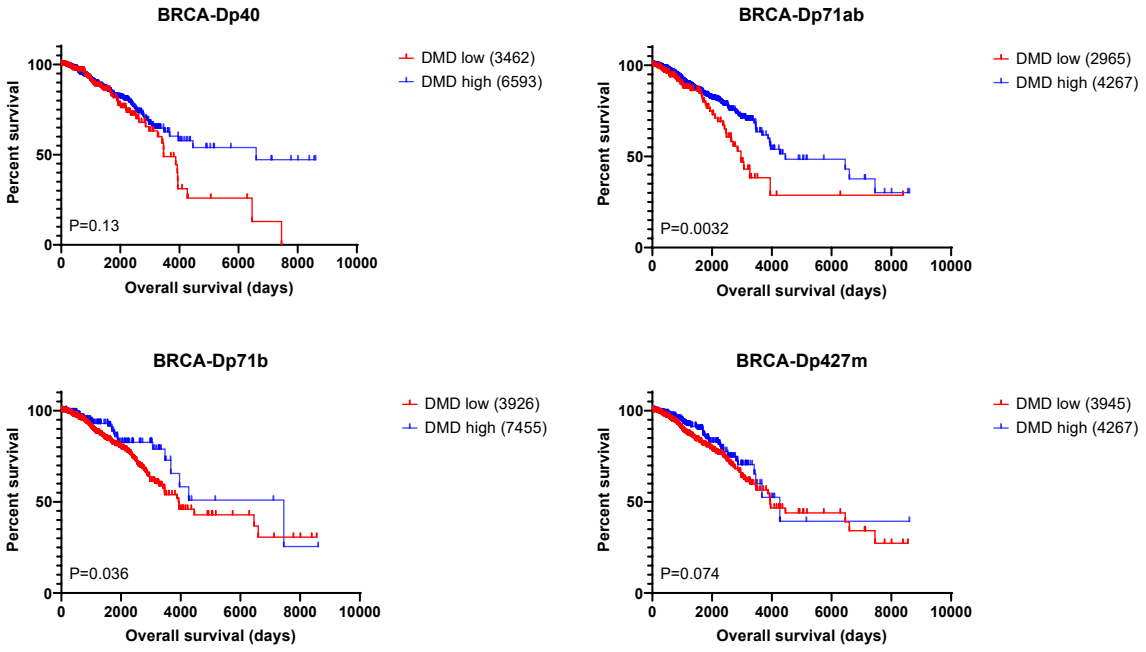


Figure 3.8 The expression of Dp71ab and Dp71b gene products are significantly associated with BRCA survival outcomes

(a) BRCA TCGA RNAseq data for each *DMD* isoform was dichotomised into high (blue) and low (red) expression groups and survival analysis performed in GraphPad using the log-rank test. Numbers in brackets are median overall survival times in months.

KIRP was examined to determine whether Dp40, Dp71ab and Dp71b were associated with overall survival (Figure 3.9). High expression of the Dp71ab (p <0001) and Dp71b (p = 0.0007) gene products were significantly associated with poor KIRP survival.

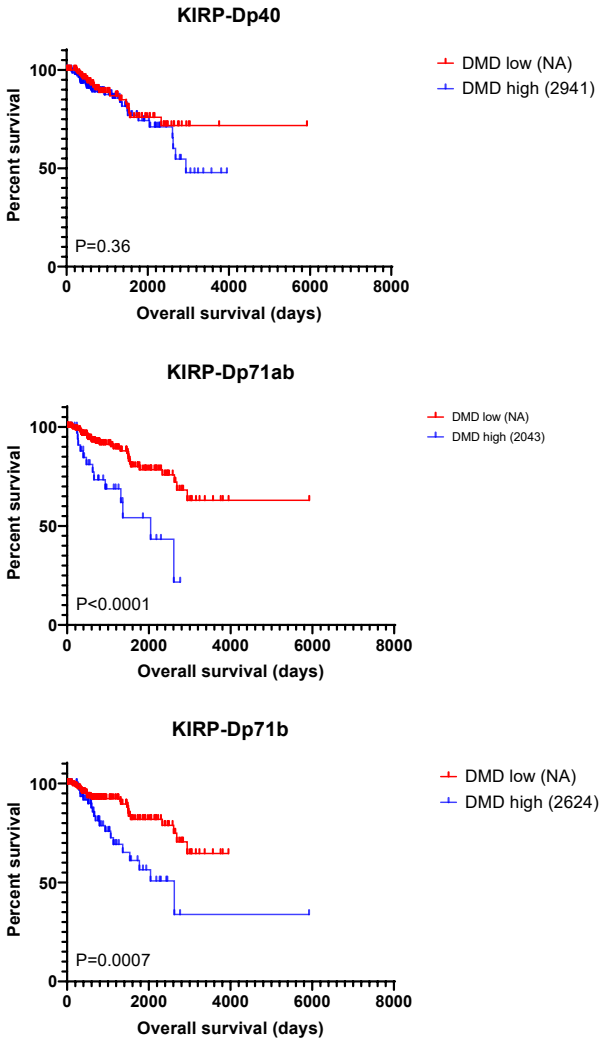


Figure 3.9 The expression of Dp71ab and Dp71b products are significantly associated with KIRP survival outcomes

KIRP TCGA RNAseq data for each *DMD* isoform was dichotomised into high (blue) and low (red) expression groups and survival analysis performed in GraphPad using the log-rank test. Numbers in brackets are median overall survival times in months.

LAML was examined to determine whether Dp40 was associated with overall survival (Figure 3.10). No significant survival differences were found.

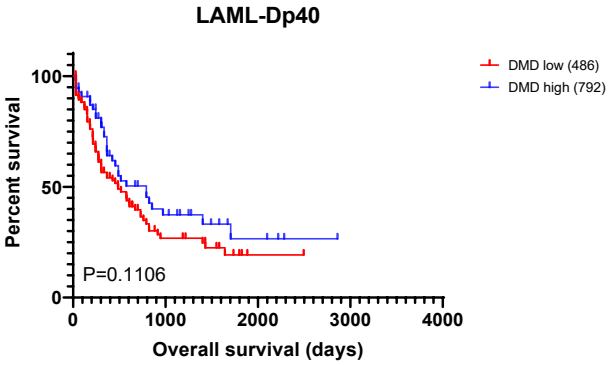


Figure 3.10 The expression of the Dp40 gene product was not significantly associated with LAML survival outcomes

LAML TCGA RNAseq data from Dp40 was dichotomised into high (blue) and low (red) expression groups and survival analysis performed in GraphPad using the log-rank test. Numbers in brackets are median overall survival times in months.

LGG was examined to determine whether Dp40, Dp71ab and Dp71b, Dp71a, Dp71, Dp116, Dp140, Dp260-1 and Dp260-2 were associated with overall survival (Figure 3.11). High expression of the Dp40 (p = 0.0486), Dp71ab (p <0.0001), Dp71 (p <0.0001), Dp116 (p <0.0001) and Dp140 (p=0.0391) gene products were significantly associated with poor LGG survival.

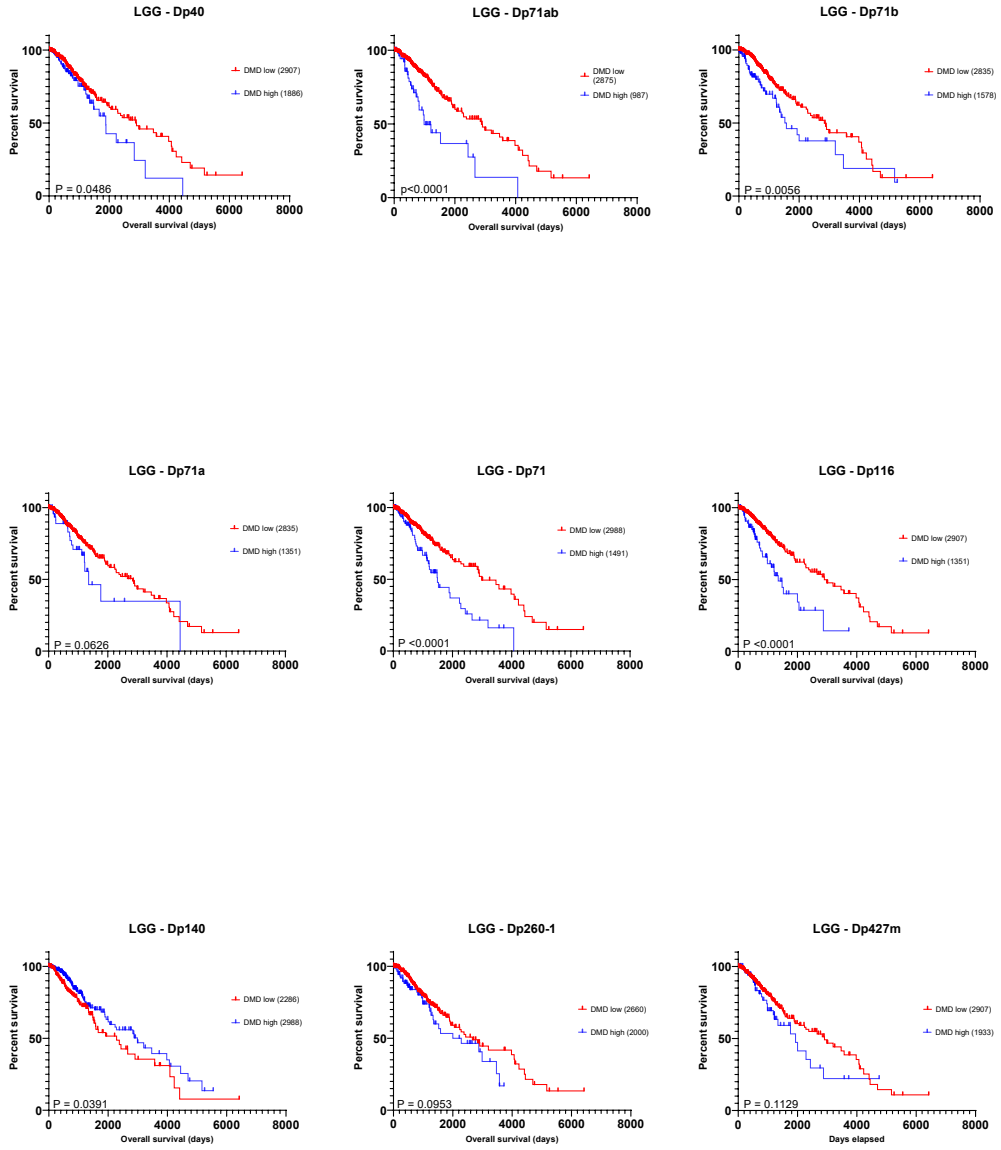


Figure 3.11 The expression of Dp40, Dp71ab, Dp71, Dp116 and Dp140 gene products are significantly associated with LGG survival outcomes

LGG TCGA RNAseq data for each *DMD* isoform was dichotomised into high (blue) and low (red) expression groups and survival analysis performed in GraphPad using the log-rank test. Numbers in brackets are median overall survival times in months.

LUAD was examined to determine whether Dp40, Dp71ab and Dp71b, and Dp427m were associated with overall survival (Figure 3.12). Low expression of the Dp71ab (p = 0.0011), and Dp427m (p=0.0041) gene products were significantly associated with poor LUAD survival.

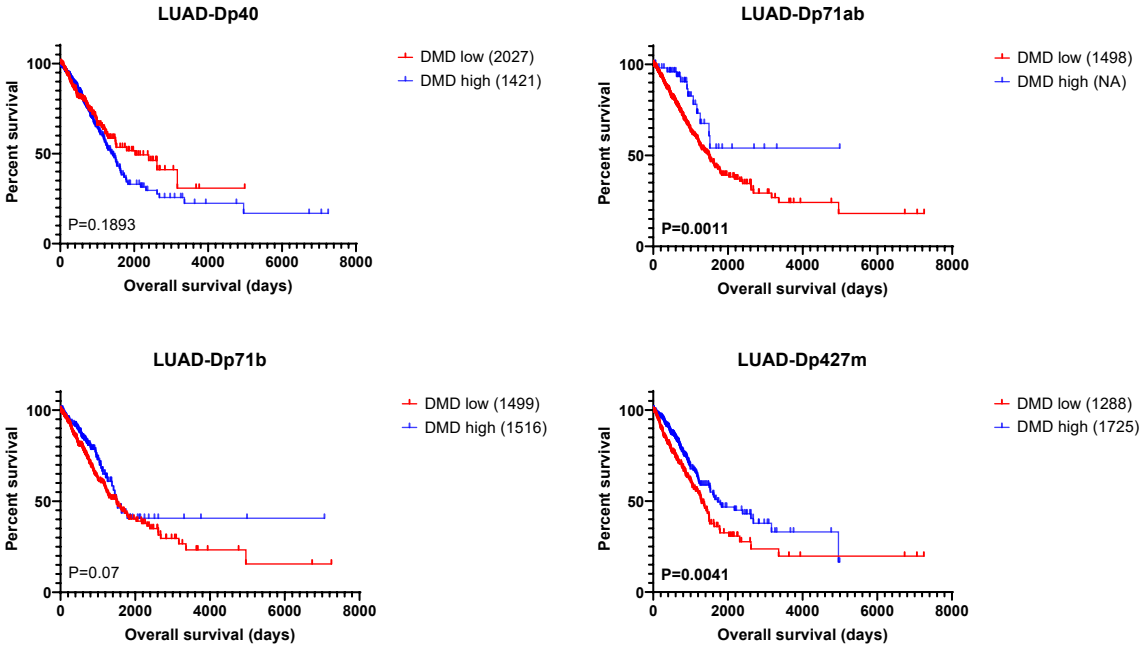


Figure 3.12 The expression of Dp71ab and Dp427m gene products are significantly associated with LUAD survival outcomes

LUAD TCGA RNAseq data for each *DMD* isoform was dichotomised into high (blue) and low (red) expression groups and survival analysis performed in GraphPad using the log-rank test. Numbers in brackets are median overall survival times in months.

PAAD was examined to determine whether Dp40, Dp71ab and Dp71b, and Dp427m were associated with overall survival (Figure 3.13). Low expression of Dp40 (p = 0.0009) was associated with poor survival and high expression of Dp71b (p = 0.026), and Dp427m (p=0.018) gene products were significantly associated with poor PAAD survival.

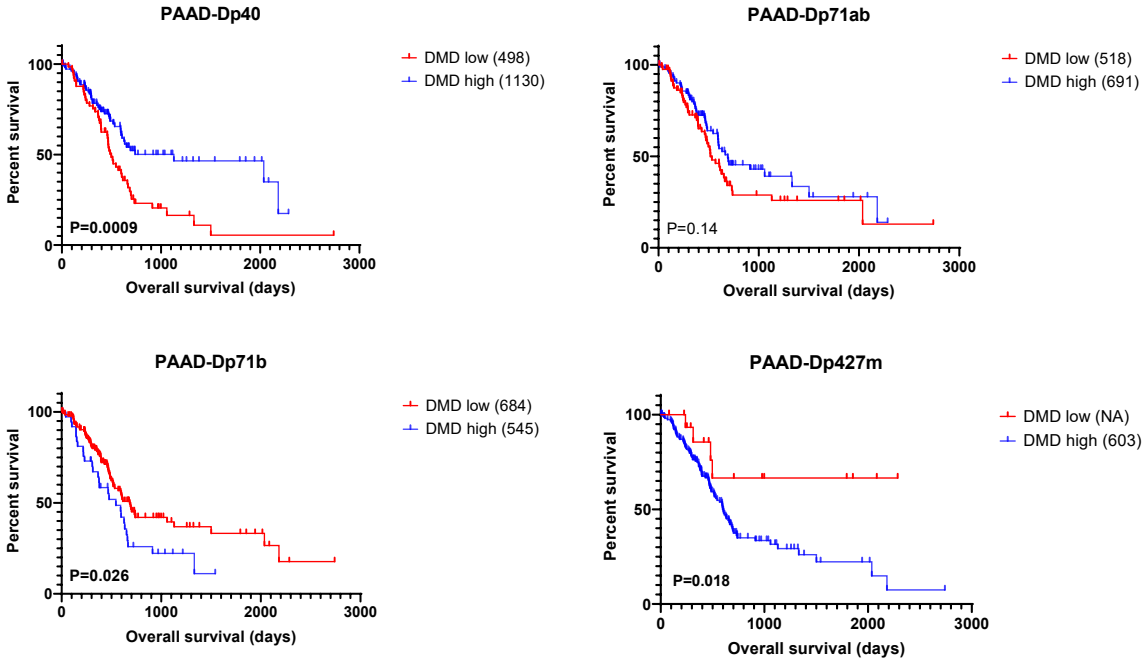


Figure 3.13 The expression of Dp40, Dp71b and Dp427m gene products are significantly associated with PAAD survival outcomes

PAAD TCGA RNAseq data for each *DMD* isoform was dichotomised into high (blue) and low (red) expression groups and survival analysis performed in GraphPad using the log-rank test. Numbers in brackets are median overall survival times in months.

READ was examined to determine whether Dp40, Dp71ab and Dp71b, and Dp427m were associated with overall survival (Figure 3.14). High expression of the Dp40 ($p < 0.0001$), Dp71ab ($p = 0.034$), and Dp427m ($p=0.00019$) gene products were significantly associated with poor READ survival.

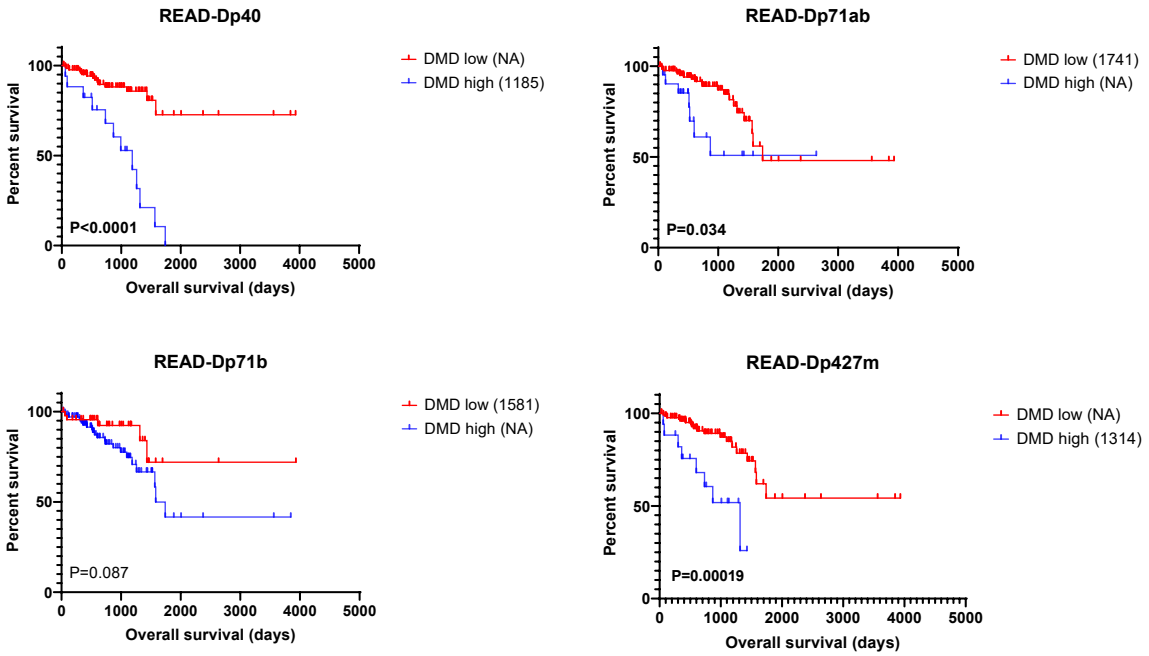


Figure 3.14 The expression of the Dp40, Dp71ab and Dp427m gene products are significantly associated with READ survival outcomes

READ TCGA RNAseq data for each *DMD* isoform was dichotomised into high (blue) and low (red) expression groups and survival analysis performed in GraphPad using the log-rank test. Numbers in brackets are median overall survival times in months.

THYM was examined to determine whether Dp40 and Dp71ab were associated with overall survival (Figure 3.15). High expression of the Dp71ab (p =0.0011), gene product was significantly associated with poor LGG survival.

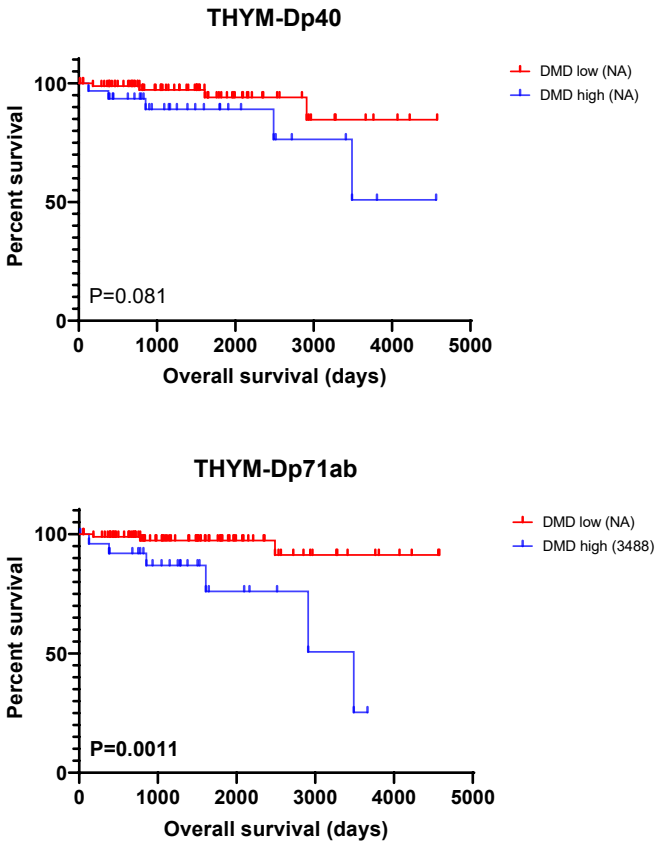


Figure 3.15 The expression of the Dp71ab gene product was significantly associated with THYM survival outcomes

THYM TCGA RNAseq data for each DMD isoform was dichotomised into high (blue) and low (red) expression groups and survival analysis performed in GraphPad using the log-rank test. Numbers in brackets are median overall survival times in months.

UVM was examined to determine whether Dp40, Dp71ab, Dp260-1 and Dp427 were associated with overall survival (Figure 3.16). High expression of the Dp40 (p = 0.031), Dp71ab (p =0.00016), Dp260-1 (p =0.0367) and Dp427m (p =0.0092), gene products were significantly associated with poor UVM survival.

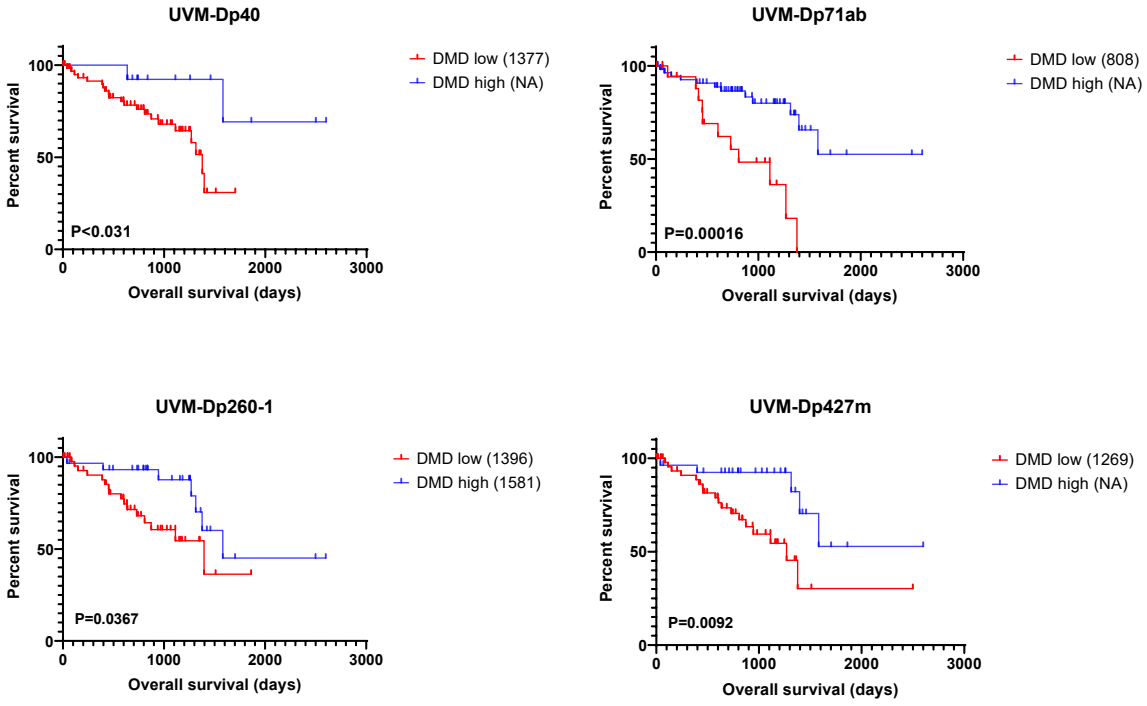


Figure 3.16 The expression of Dp40, Dp71ab, Dp260-1 and Dp427m gene products are significantly associated with UVM survival outcomes

UVM TCGA RNAseq data for each *DMD* isoform was dichotomised into high (blue) and low (red) expression groups and survival analysis performed in GraphPad using the log-rank test. Numbers in brackets are median overall survival times in months.

Based on the specific *DMD* gene products shown to be expressed in the nine selected TCGA tumours, Hazard ratios were calculated with an overview of the specific gene variant shown (Figure 3.17). In summary, UVM has four transcripts where low expression is associated with the largest increase in the hazard across all cancers. Low expression of transcripts from THYM, LGG, and KIRP are protective (HR <1) and low expression of transcripts in LUAD, BRCA (and UVM) increased the hazard. PAAD has two transcripts associated with protection and one associated with increased hazard. Finally, THYM has one protective transcript. In summary, the associated direction of the hazard across (and sometimes within) TCGA cancers expressing specific gene products is complex and variable.

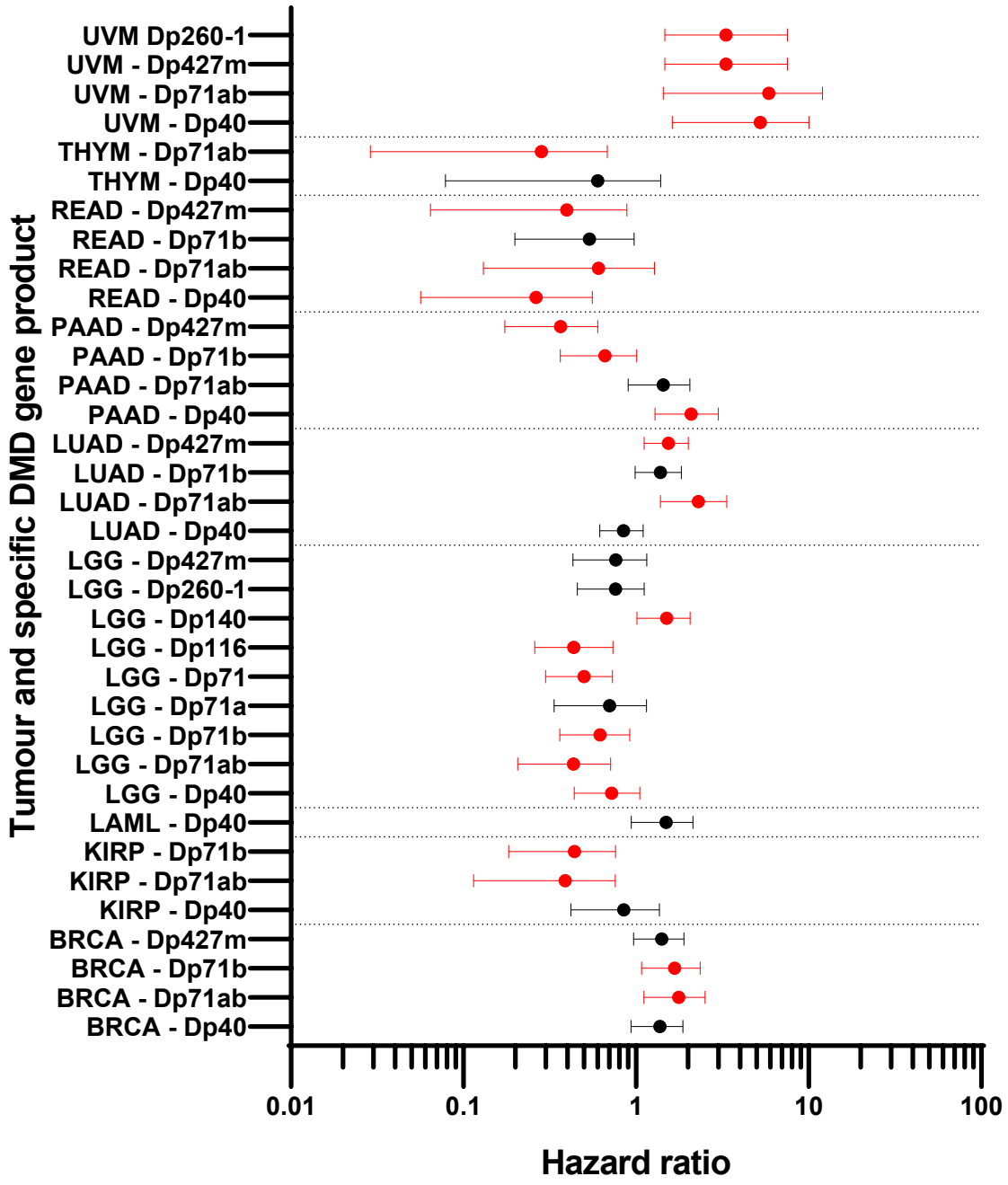


Figure 3.17 Hazard ratios of TCGA tumours expressing specific DMD gene products
Forest plots revealing the log-rank hazard ratio with 95% confidence intervals. Red bars have significant p-values (alpha < 0.05) UVM n=80, THYM n=121, READ n=177, PAAD n = 182, LUAD n=594, LGG n= 529, LAML n= 151, KIRP n= 321, BRCA n=1222. Hazard ratios below 1 indicate low gene expression is protective and values above 1 indicate low expression is a hazard.

3.2.6 iDEP pathway analysis

To aid future investigation of functional role(s) for the *DMD* gene across tumour types, a preliminary bioinformatic analysis of differentially expressed genes (DEGs) was undertaken in cases comparing high verses low *DMD* expression. iDEP was employed to identify differentially expressed genes using the DESeq2 method. To examine the functional annotations of the DEGs, an enrichment analysis (gene ontology [GO] biological processes) for the DEGs was performed.

For UVM with the DESeq2 package, 750 upregulated and 472 downregulated genes were identified (Figure 3.18a). A STRING network of protein-protein interactions (PPIs) among the top 20 upregulated genes was constructed (Figure 3.18b). The connected network includes several proto-cadherin interactions. The expected number of edges for a random set of proteins of similar size was 3 compared with an observed of 6 suggesting functional intersection of the identified DEGs. However, this did not reach significance (p=0.143). To visualise the relationship among enriched GO terms the distance among the terms was measured by the percentage of overlapped genes. Then this distance is used to construct a hierarchical clustering tree (Figure 3.18c) and a network of GO terms (Figure 3.18d). Both plots show that the enriched terms are distinct. The up-regulated genes are overwhelmingly involved in cell migration and are related to cilium mediated motility, microtubule-based movement, and adhesion pathways. The down-regulated genes are related to 2 major themes: cell division and differentiation/developmental programmes.

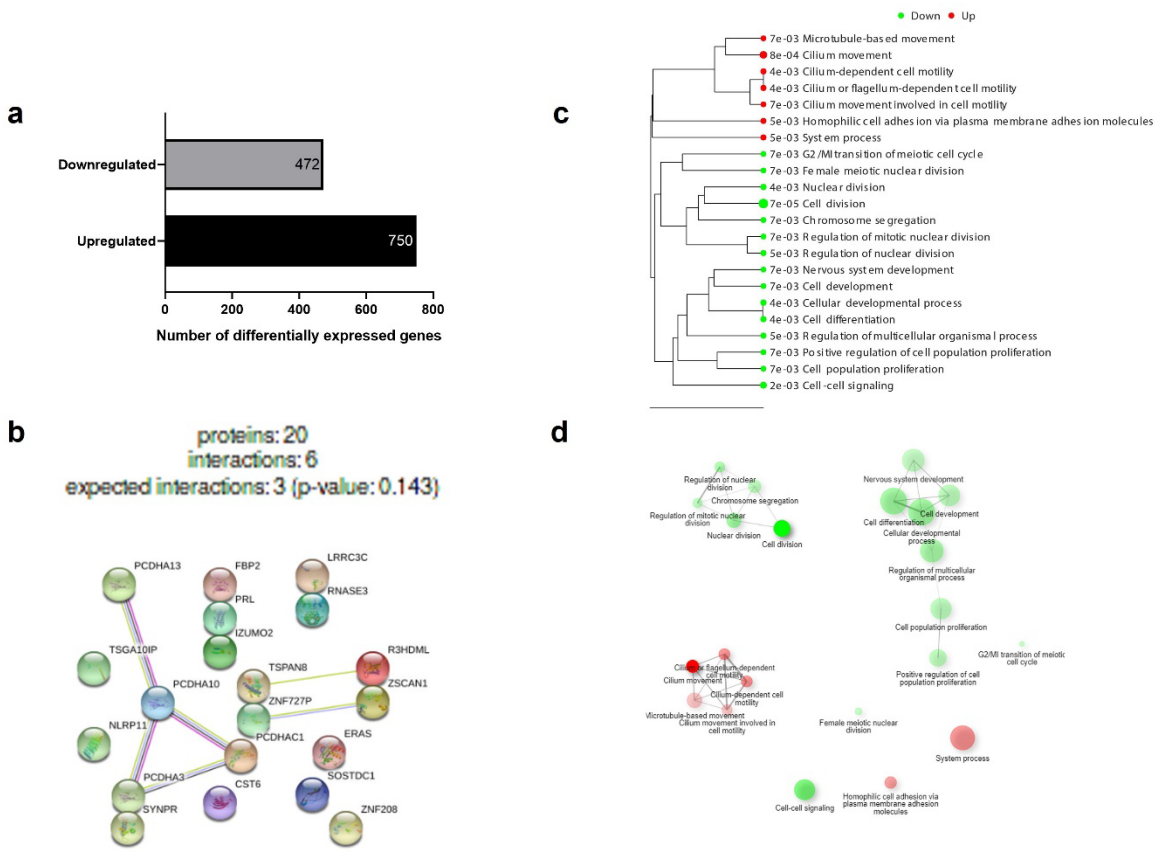


Figure 3.18 Exploratory analysis of the DEGs with high versus low *DMD* expression in UVM

(a) Column graph highlighting the number of upregulated and downregulated genes.

(b) STRING Protein-protein interactions (PPI) among top 20 up-regulated genes.

(c) Visualisation of the relationship among enriched GO categories. Connected gene sets share more genes, size of node represents adjusted P values. Upregulated and downregulated genes are indicated by red and green points respectively.

(d) Network tree Visualisation of the enriched pathways in DEGs using the GO biological processes annotation, dot size corresponds to adjusted P values.

For THYM, 1941 upregulated and 1383 downregulated genes were identified (Figure 3.19a). A STRING network of protein-protein interactions (PPIs) among the top 20 upregulated genes was constructed (Figure 3.19b). The connected network includes several Homeobox gene interactions. The expected number of edges for a random set of proteins of similar size was 3 compared with an observed of 12 strongly suggesting functional intersection of the identified DEGs (p=0.000198). To visualise the relationship among enriched GO terms the distance among the terms was measured by the percentage of overlapped genes. Then this distance is used to construct a hierarchical clustering tree (Figure 3.19c) and a network of GO terms (Figure 3.19d). Both plots show that the enriched terms are distinct. The downregulated genes are overwhelmingly involved in nuclear organisation and mitosis pathways. The upregulated genes are related to 2 major themes: cell adhesion and cell migration morphogenesis.

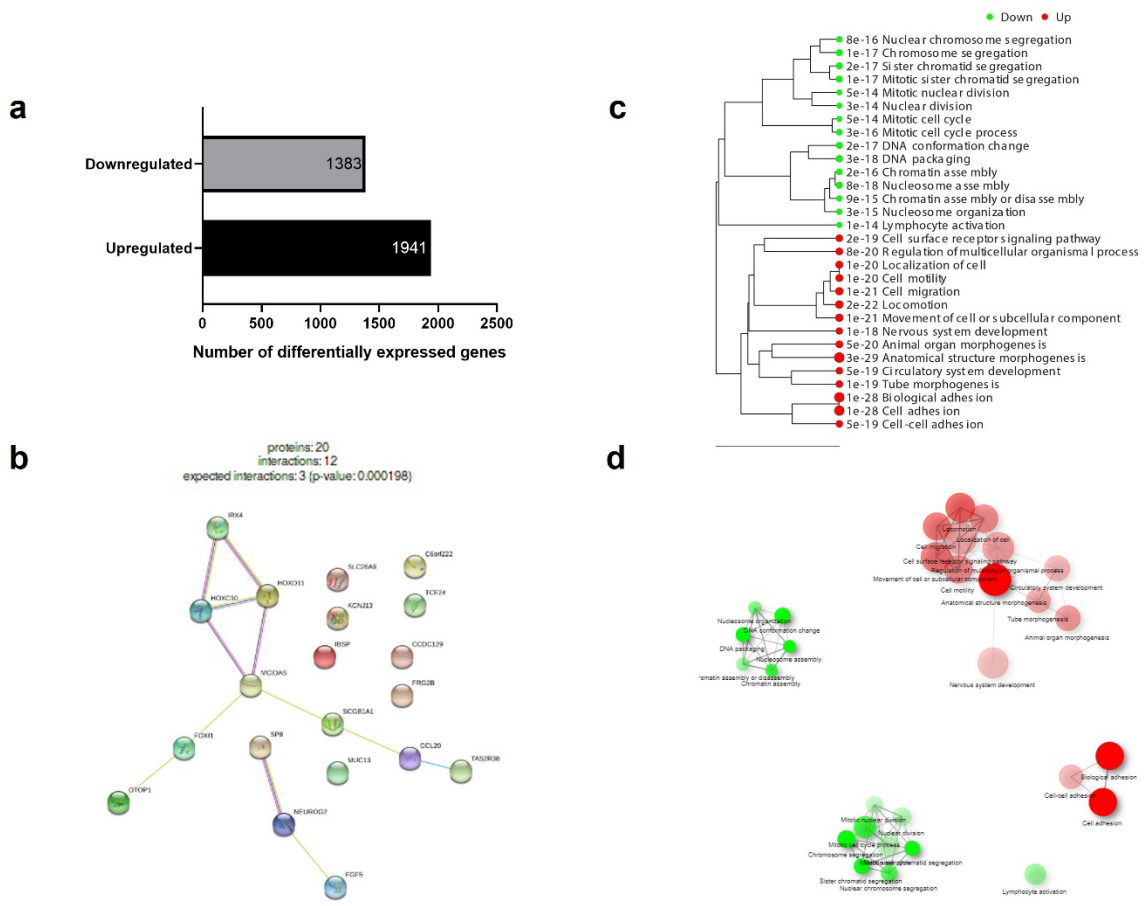


Figure 3.19 Exploratory analysis of the DEGs with high versus low *DMD* expression in THYM

(a) Column graph highlighting the number of upregulated and downregulated genes.
(b) STRING Protein-protein interactions (PPI) among top 20 up-regulated genes.
(c) Visualisation of the relationship among enriched GO categories. Connected gene sets share more genes, size of node represents adjusted P values. Upregulated and downregulated genes are indicated by red and green points respectively.
(d) Network tree Visualisation of the enriched pathways in DEGs using the GO biological processes annotation, dot size corresponds to adjusted P values.

For READ, 290 upregulated and 44 downregulated genes were identified (Figure 3.20a). A STRING network of protein-protein interactions (PPIs) among the top 20 upregulated genes (Figure 3.20b) identified several Melanoma Antigen Gene (MAGE) family members. The expected number of edges for a random set of proteins of similar size was 3 compared with an observed of 35 strongly suggesting functional intersection of the identified DEGs ($p<0.0001$). To visualise the relationship among enriched GO terms hierarchical clustering tree (Figure 3.20c) and a network analysis (Figure 3.20d) showed that the enriched terms were distinct with downregulated genes overwhelmingly involved in humoral immune responses and upregulated genes relating to calcium regulation and muscle processes.

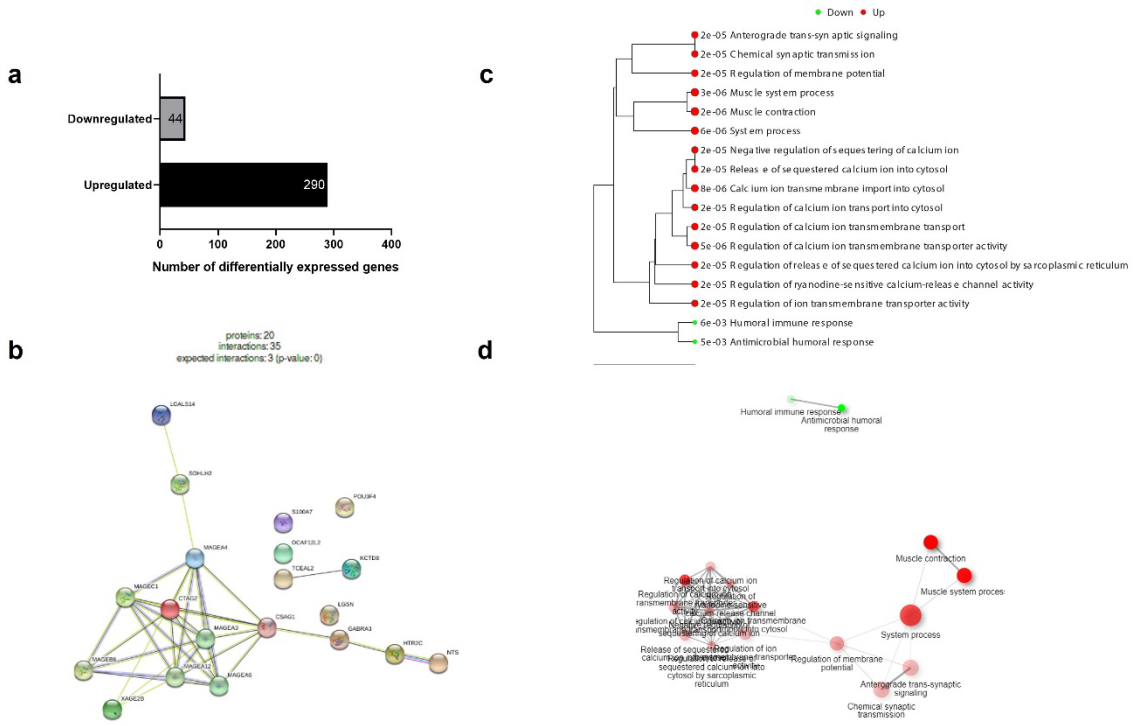


Figure 3.20 Exploratory analysis of the DEGs with high versus low *DMD* expression in READ

(a) Column graph highlighting the number of upregulated and downregulated genes.
(b) STRING Protein-protein interactions (PPI) among top 20 up-regulated genes.
(c) Visualisation of the relationship among enriched GO categories. Connected gene sets share more genes, size of node represents adjusted P values. Upregulated and downregulated genes are indicated by red and green points respectively.
(d) Network tree Visualisation of the enriched pathways in DEGs using the GO biological processes annotation, dot size corresponds to adjusted P values.

For PAAD, 1239 upregulated and 187 downregulated genes were identified (Figure 3.21a) with a STRING network of protein-protein interactions (PPIs) among the top 10 upregulated genes (Figure 3.21b) that identified several pancreatic enzyme family members (e.g. Carboxypeptidases). The expected number of edges for a random set of proteins of similar size was 1 compared with an observed of 16 strongly suggesting functional intersection of the identified DEGs (p = 7.53 x10-12). To visualise the relationship among enriched GO terms hierarchical clustering tree (Figure 3.21c) and a network analysis (Figure 3.21d) showed that the enriched terms were not distinct for downregulated genes and upregulated genes related to cation transport and regulation.

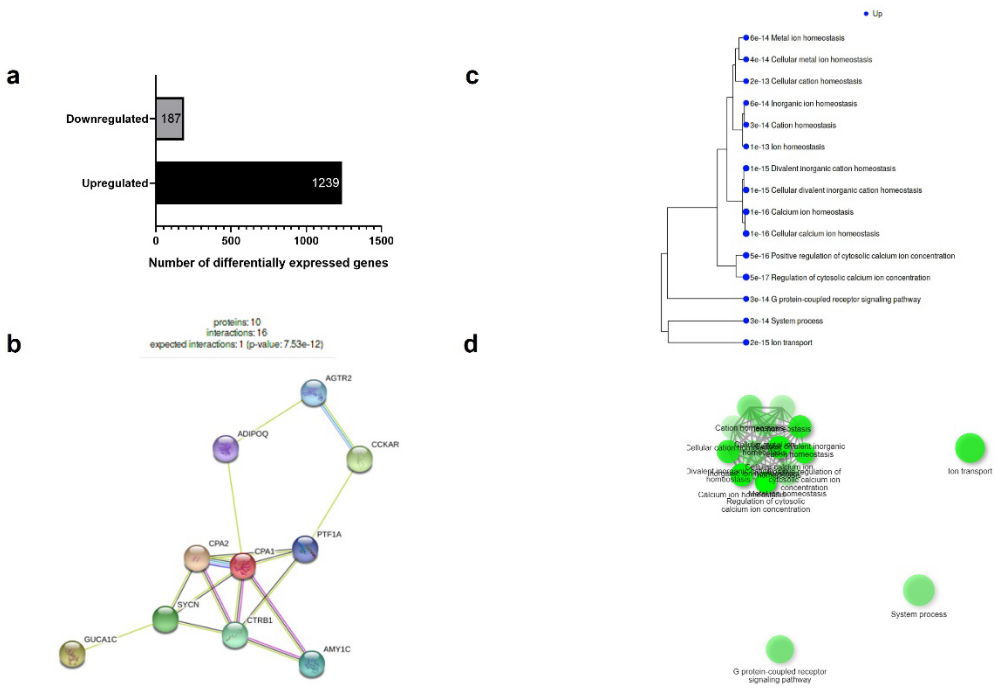


Figure 3.21 Exploratory analysis of the DEGs with high versus low *DMD* expression in PAAD

(a) Column graph highlighting the number of upregulated and downregulated genes.
(b) STRING Protein-protein interactions (PPI) among top 10 up-regulated genes.
(c) Visualisation of the relationship among enriched GO categories. Connected gene sets share more genes, size of node represents adjusted P values. Upregulated genes are indicated by blue points.
(d) Network tree Visualisation of the enriched pathways in DEGs using the GO biological processes annotation, dot size corresponds to adjusted P values.

In LUAD, 755 upregulated and 108 downregulated genes were identified (Figure 3.22a). A STRING network of protein-protein interactions (PPIs) among the top 20 upregulated genes (Figure 3.22b) showed some interaction centering on DDX53 (p = 0.0272). Hierarchical clustering (Figure 3.22c) and a network analysis (Figure 3.22d) showed that enriched terms represented immune responses for downregulated genes and stimulus detection for upregulated genes.

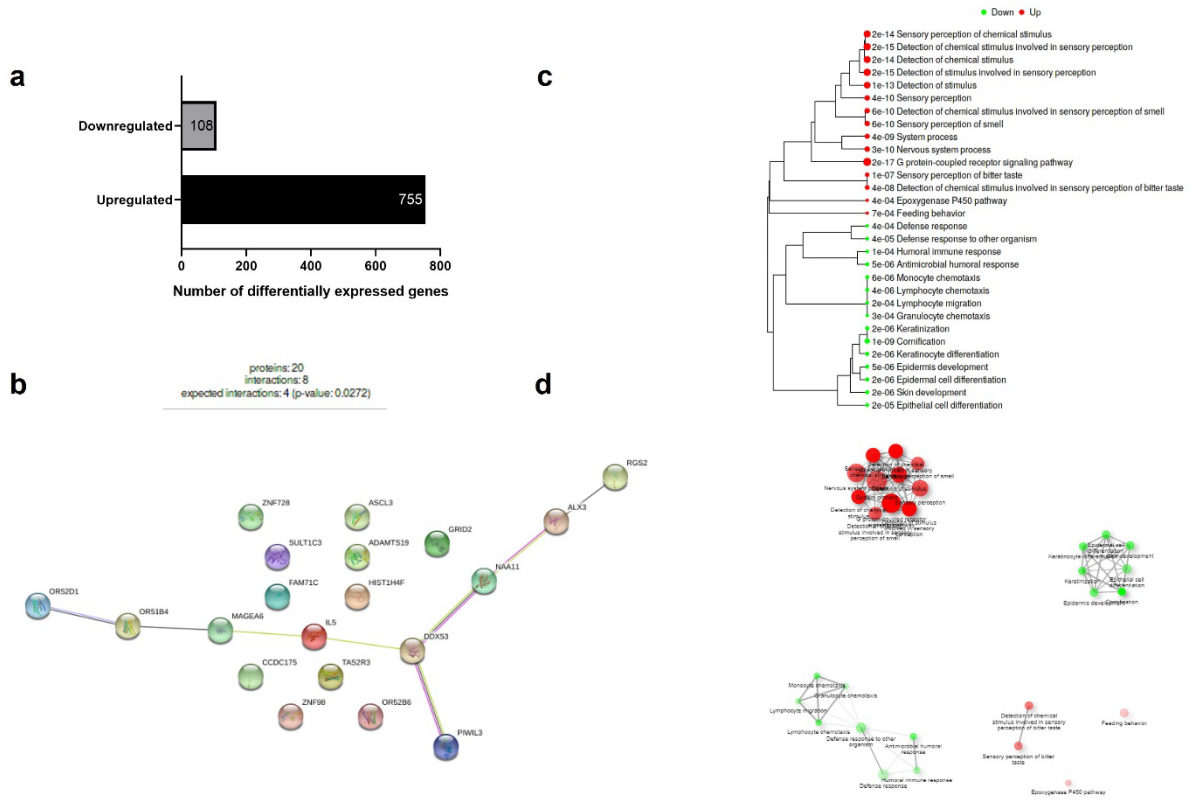


Figure 3.22 Exploratory analysis of the DEGs with high versus low *DMD* expression in LUAD

(a) Column graph highlighting the number of upregulated and downregulated genes.
(b) STRING Protein-protein interactions (PPI) among top 20 up-regulated genes.
(c) Visualisation of the relationship among enriched GO categories. Connected gene sets share more genes, size of node represents adjusted P values. Upregulated and downregulated genes are indicated by red and green points respectively.
(d) Network tree Visualisation of the enriched pathways in DEGs using the GO biological processes annotation, dot size corresponds to adjusted P values.

In LGG, 537 upregulated and 191 downregulated genes were identified (Figure 3.23a). A STRING network of protein-protein interactions (PPIs) among the top 20 upregulated genes (Figure 3.23b) showed interactions focusing on HOX family members (p = 0.000242). Hierarchical clustering (Figure 3.23c) and a network analysis (Figure 3.23d) showed that enriched terms represented hormone regulation and cell signalling for downregulated genes and development/morphogenesis and cell motility for upregulated genes.

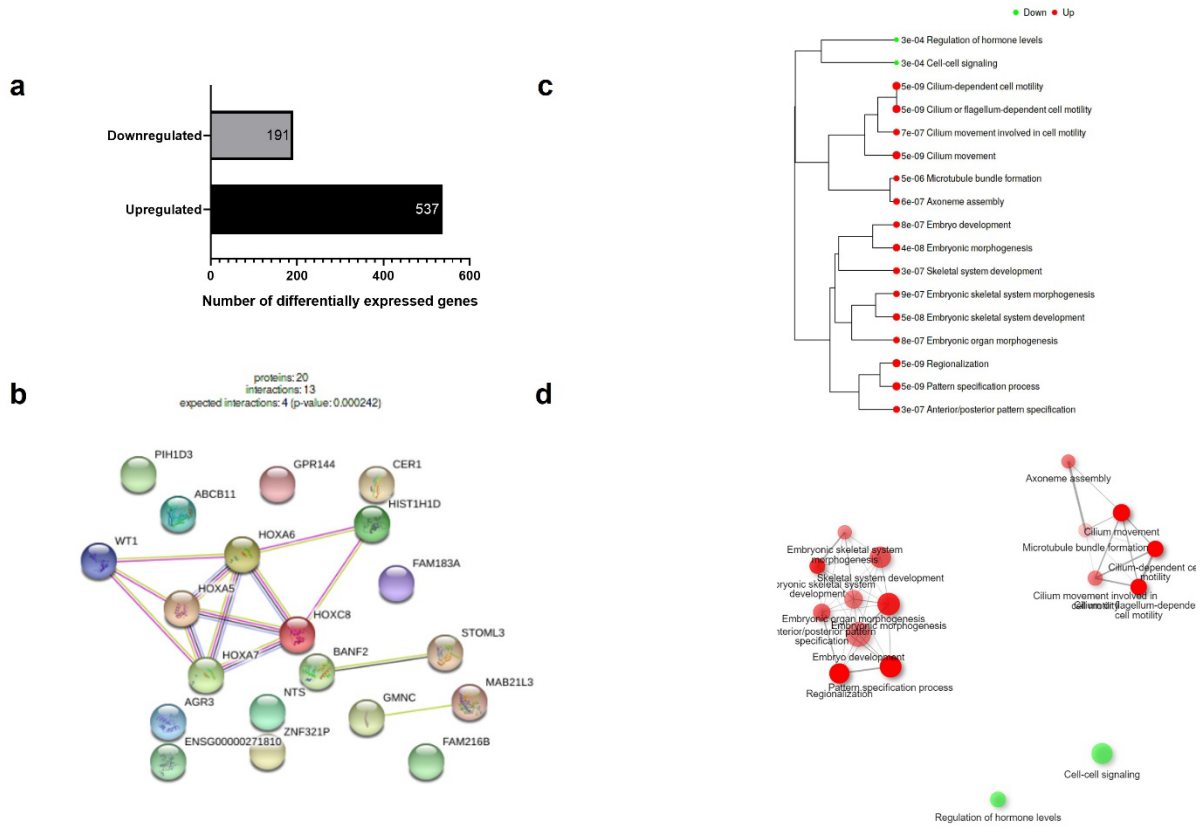


Figure 3.23 Exploratory analysis of the DEGs with high versus low *DMD* expression in LGG

(a) Column graph highlighting the number of upregulated and downregulated genes.
(b) STRING Protein-protein interactions (PPI) among top 20 up-regulated genes.
(c) Visualisation of the relationship among enriched GO categories. Connected gene sets share more genes, size of node represents adjusted P values. Upregulated and downregulated genes are indicated by red and green points respectively.
(d) Network tree Visualisation of the enriched pathways in DEGs using the GO biological processes annotation, dot size corresponds to adjusted P values.

In LAML, 965 upregulated and 88 downregulated genes were identified (Figure 3.24a). A STRING network of protein-protein interactions (PPIs) among the top 20 upregulated genes (Figure 3.24b) showed interactions of migration/adhesion molecules (p = 4.34 x10-11). Confirming this hierarchical clustering (Figure 3.24c) and a network analysis (Figure 3.24d) showed that enriched terms represented cell motility, morphogenesis a notably cell adhesion for upregulated genes.

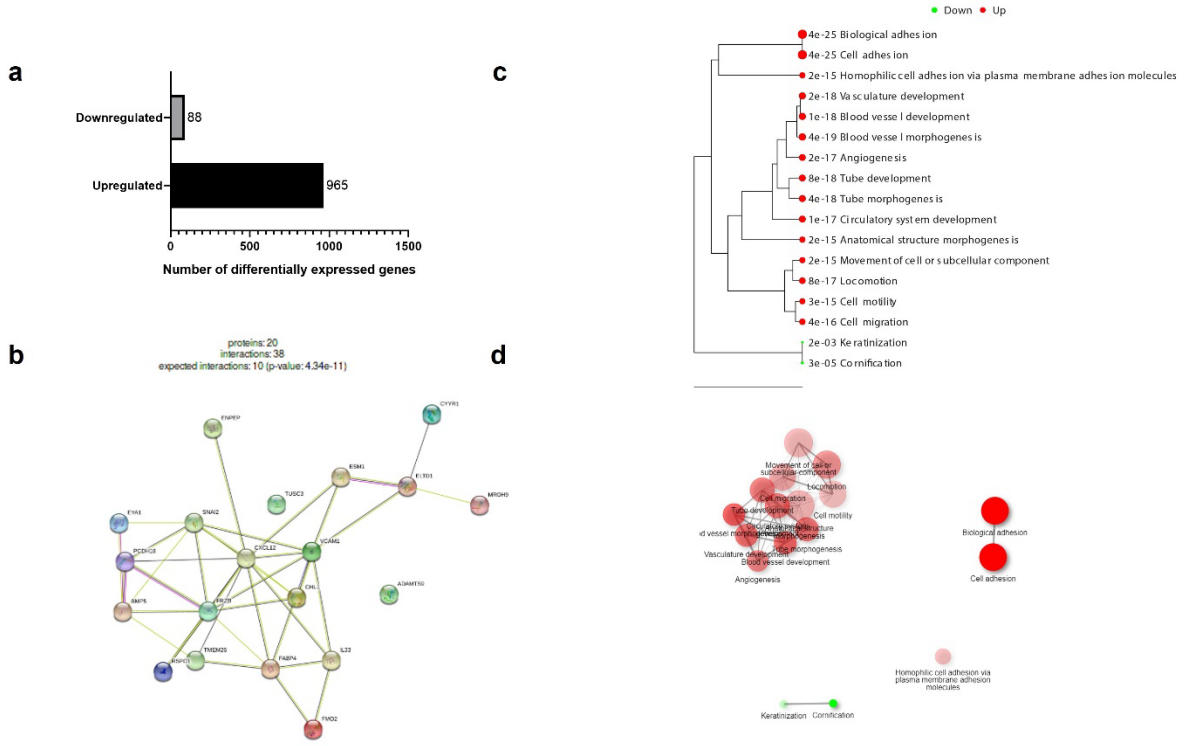


Figure 3.24 Exploratory analysis of the DEGs with high versus low *DMD* expression in LAML

(a) Column graph highlighting the number of upregulated and downregulated genes.
(b) STRING Protein-protein interactions (PPI) among top 20 up-regulated genes.
(c) Visualisation of the relationship among enriched GO categories. Connected gene sets share more genes, size of node represents adjusted P values. Upregulated and downregulated genes are indicated by red and green points respectively.
(d) Network tree Visualisation of the enriched pathways in DEGs using the GO biological processes annotation, dot size corresponds to adjusted P values.

In KIRP, 1349 upregulated and 132 downregulated genes were identified (Figure 3.25a). A STRING network of protein-protein interactions (PPIs) among the top 20 upregulated genes (Figure 3.25b) showed interactions focusing fibrinogen family members (p = 2.79 x10-5). Hierarchical clustering (Figure 3.25c) and a network analysis (Figure 3.25d) showed that enriched terms represented development/morphogenesis for upregulated genes.

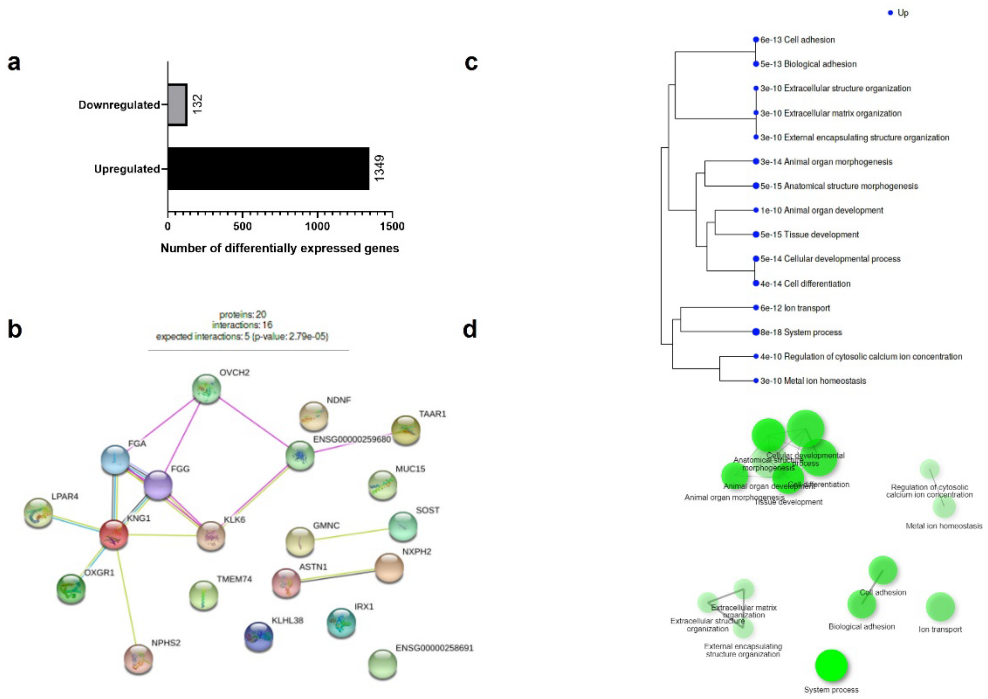


Figure 3.25 Exploratory analysis of the DEGs with high versus low *DMD* expression in KIRP

(a) Column graph highlighting the number of upregulated and downregulated genes.
(b) STRING Protein-protein interactions (PPI) among top 20 up-regulated genes.
(c) Visualisation of the relationship among enriched GO categories. Connected gene sets share more genes, size of node represents adjusted P values. Upregulated genes are indicated by blue points.
(d) Network tree Visualisation of the enriched pathways in DEGs using the GO biological processes annotation, dot size corresponds to adjusted P values.

In BRCA, 1886 upregulated and 235 downregulated genes were identified (Figure 3.26a). A STRING network of protein-protein interactions (PPIs) among the top 20 upregulated genes (Figure 3.26b) did not contain interactions focusing on obvious family members although there were more interactions than expected (p = 0.0384). Hierarchical clustering (Figure 3.26c) and a network analysis (Figure 3.26d) showed that enriched terms represented cornification and adhesion for upregulated genes.

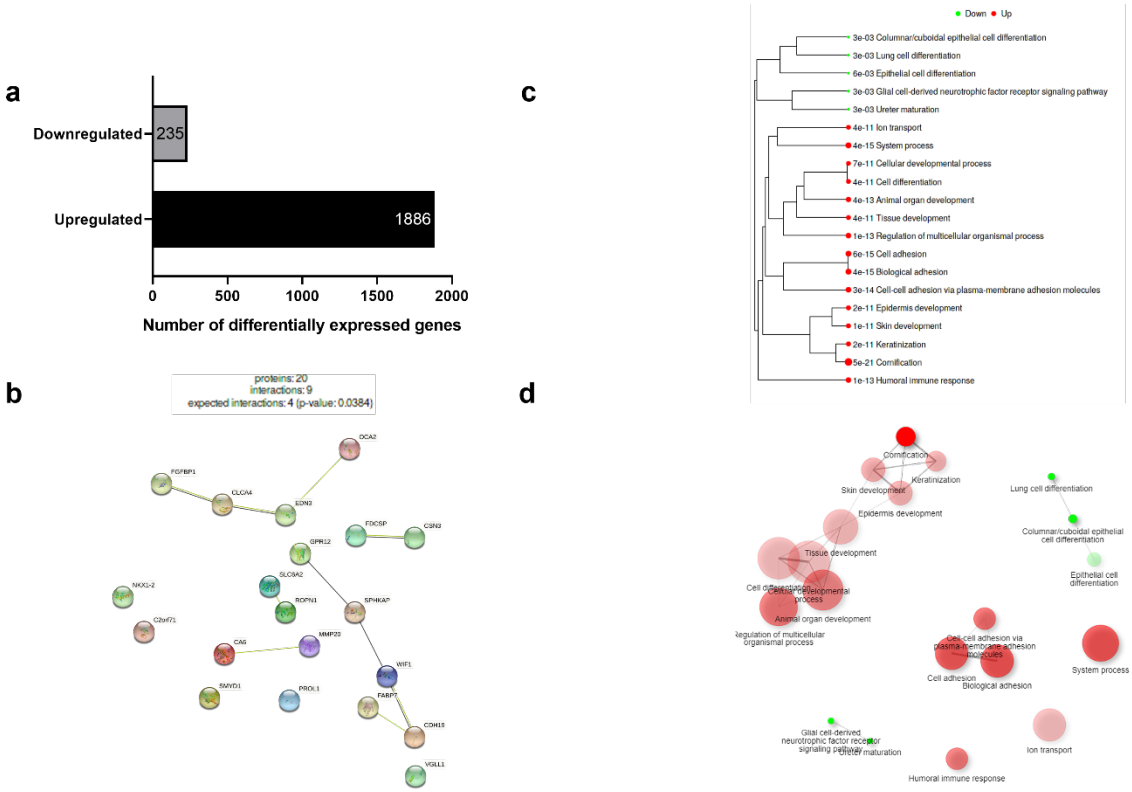


Figure 3.26 Exploratory analysis of the DEGs with high versus low *DMD* expression in BRCA

(a) Column graph highlighting the number of upregulated and downregulated genes.
(b) STRING Protein-protein interactions (PPI) among top 20 up-regulated genes.
(c) Visualisation of the relationship among enriched GO categories. Connected gene sets share more genes, size of node represents adjusted P values. Upregulated and downregulated genes are indicated by red and green points respectively.
(d) Network tree Visualisation of the enriched pathways in DEGs using the GO biological processes annotation, dot size corresponds to adjusted P values.

Chapter 4

4. Discussion

4.1 *DMD* is associated with survival outcomes in cancer patients

The aim of this project was to determine whether *DMD* and derived gene product expression was associated with survival outcomes in patient tumours across 33 TCGA cancer types. This pan-cancer analysis identified nine/33 tumours (after Bonferroni correction) where high vs. low *DMD* expression was significantly associated with overall survival outcome differences which is novel and worth further exploration. This has been done more comprehensively for LGG where our research team has used an independent bioinformatic approach using Cbioportal (Gao *et al.*, 2013; Cerami *et al.*, 2012) for data extraction, and X-tile (Camp *et al.*, 2004) to do cutpoint selection. This allowed subsequent survival and pathway analysis (Naidoo *et al.*, 2022). The key findings between the two approaches are largely concordant, validating this novel bioinformatic approach, although for specific dystrophin gene products (i.e. Dp427m) significance in survival was not quite reached using the pipeline described herein. This may reflect the fact that gene expression values obtained from Cbioportal (RSEM) and the GDC (FPKM-UQ) were derived using different bioinformatic pipelines. Indeed, during this project, GDC have continued to update their pipeline again providing FPKM, FPKM-UQ and now additionally TPM normalized gene expression data (see code availability section for additional updated code). For LGG, this project showed preliminary data suggesting differentially transcribed genes belonging to pathways relevant to development/morphogenesis and cell motility. Furthermore, detailed pathway analysis in the published study (Naidoo *et al.*, 2022) identified biological processes relating to ribosome biogenesis, synaptic signalling, neurodevelopment and immune pathways as well. The genes spanning chromosome 1 were globally upregulated in high vs. low expressing *DMD* cohorts.

Importantly immunohistochemistry was used to demonstrate dystrophin protein expression in these tumours validating the RNAseq analysis.

Of the other tumour types significantly associated with survival outcomes, high expression of *DMD* in THYM, READ and KIRP were also associated with increased risks of poor survival. Conversely, high expression of *DMD* in UVM, PAAD, LUAD, LAML and BRCA were associated with protection (compared with high expressing cases). The reasons for this are currently unclear but may reflect the specific biological pathways affected and/or the composition of the DAPC in different tumour tissues. For example, in healthy tissue, the composition of the DAPC at the sarcolemma, neuromuscular junction (NMJ), CNS and retina is known to differ (Figure 4.1). At the NMJ the DAPC is comprised of the dystrophin-related protein utrophin and α-dystrobrevin1 replaces α-dystrobrevin2 (Ohlendieck *et al.*, 1991). In cancer, the muscle wasting condition cachexia has been linked to dysfunction of the DAPC in an animal model (Acharyya *et al.*, 2005). Muscles from mice bearing subcutaneous colon-26 (C-26) tumours were severely atrophic, with altered histology showing abnormal sarcolemma and associated basal lamina from cachectic tibialis anterior muscles. This was associated with a switch from dystrophin to utrophin expression and a higher migrating band for both β-DG and β-SG suggested by the authors to be a hyperglycosylated form. This data suggests that cross talk exists between tumour cells and the local tissue microenvironment modifying DAPC formation. Interestingly, *DMD* gene products may not just play a role as a scaffold for structural and signalling proteins at the plasma membrane. Dp71d is known to undergo nuclear import employing an atypical nuclear localization signal by a ZZ-domain of the α2/β1 importin system. After import Dp71d aids in the maintenance of nuclear architecture, through interaction with the nuclear envelope proteins emerin and lamins A, C and B1. (Suárez-Sánchez *et al.*, 2014).

Given the likely different composition of the DAPC across tumour types, we sought to determine whether other gene members of the DAPC were associated with patient survival outcomes. Across the nine tumours, several DAPC genes were not associated with survival outcomes of those specific cancers. The exception was for LGG where intriguingly, every DAPC gene was associated with different survival outcomes between high and low expressing cases. As there were no easily immediately discernable patterns in the pattern of hazard ratios across the genes in each tumour, cluster analysis was done to see if any patterns emerged. Whether clustering was based on genes or tumour types, discernable clusters could be identified showing the relationship between tumours based on DAPC gene expression hazard ratios or genes based on tumour clustering. As discussed above the lack of a clear pattern across tumours may reflect the different structures of the complex in different tissues and the distinct roles DAPC gene products may contribute within the cell, based on cellular location. *In vitro* and *in vivo* models as well as tissue immunohistochemical studies have been deployed to interrogate the role of DAPC members in cancer with a number of studies implicating dystroglycan (in particular) in cancer biology (Cross *et al.*, 2008; Brennan *et al.*, 2004; Mitchell *et al.*, 2013; Sgambato and Brancaccio, 2005; Fernandez *et al.*, 2010; Mathew *et al.*, 2013; Calogero *et al.*, 2006). Mice lacking sarcoglycan also spontaneously develop eRMS tumours (Fernandez *et al.*, 2010). Therefore, although this study has focused on the *DMD* gene in particular, evidence implicates the DAPC in cancer biology although further studies are urgently needed.

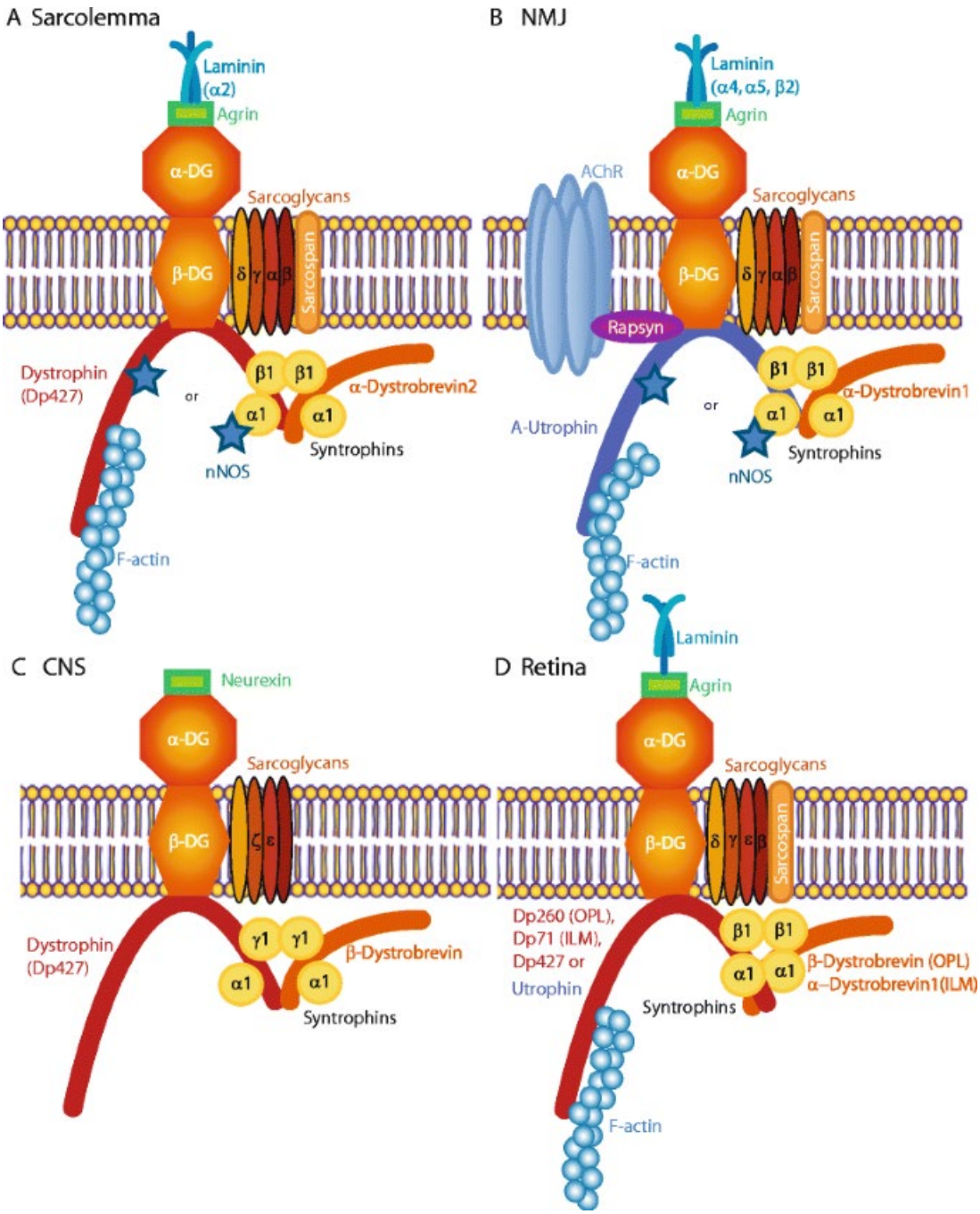


Figure 4.1 Composition of the DAPC at different tissue sites

DAPC components differ in the (a) Sarcolemma, (b) neuromuscular junction, (c) central nervous system and (d) Retina with respect to specific *DMD* gene products and other DAPC members. taken from (Pilgram *et al.*, 2010)

Currently most studies of *DMD* associated cancers have ignored the complex pattern of gene product usage. Therefore, this project explored the expression pattern of *DMD* gene variants in tumours with significant survival associations with total *DMD* and highlighted novel patterns of gene expression. Most tumours expressed Dp40, two of the Dp71 variants and Dp427. However, there were differences with some tumours not expressing Dp427 and LGG having a particularly broad pattern of gene product expression which is perhaps unsurprising given the known significant role of *DMD* in the brain (Naidoo and Anthony, 2020). Dp40 was expressed in all nine tumours which is interesting given its relatively poor characterisation compared with other dystrophins. Dp40 is the shortest dystrophin reported, transcribed from intron 62 to exon 70 and shares a promoter with Dp71. It lacks syntrophin and dystrobrevin binding domains but contains a β-dystroglycan (β-DG) WW binding domain. Mass spectrometry analysis showed Dp40 is expressed in synaptic vesicles and is associated with syntaxin1A and SNAP25 (presynaptic proteins) (Tozawa *et al.*, 2012). The effect of Dp40 and Dp40L170P (leucine to proline in residue 170 of Dp40 which promotes exclusive nuclear localisation of Dp40) stable overexpression during neuronal differentiation of PC12 Tet-On cells was evaluated (García-Cruz *et al.*, 2022). Overexpression was shown to modify neurite outgrowth and the protein expression profile of PC12 cells. Specifically, Dp40 overexpression increased the proportion of PC12 cells with neurites and neurite length. Conversely, Dp40L170P overexpression decreased neurites and neurite length.

Except for LAML, at least one Dp71 variant was expressed in all tumours examined. This may reflect the ubiquitous expression of Dp71 family members in the body. Others have shown that decreased Dp71 expression (in GBM cell lines compared with astrocytic control cells) is associated with increased cancer cell proliferation (Ki-67 as a marker) and poor prognosis (from cytoplasmic to nuclear relocalisation) in glioblastoma (Ruggieri *et al.*, 2019). In gastric

cancer cell lines Dp71 overexpression (both Dp71d and Dp71f) inhibited proliferation of SGC7901 gastric cells. Dp71 protein was bound with lamin B1 in GES-1 cells demonstrated with immunoprecipitation experiments (Sipin Tan *et al.*, 2016).

Survival analysis showed that several specific transcripts were associated with different survival outcomes across the analysed tumours. Typically for tumours with significant hazard ratios, the specific gene products in each tumour trend in the same direction but not always. Comparing across all cancers, UVM had the highest hazard ratios, for all four gene products expressed (Dp40, Dp71ab, Dp260-1 and Dp427m). Interestingly, UVM cases have virtually no mutation across the *DMD* gene suggesting non-mutational mechanism predominates in this setting (Figure 1.7). This data generally supports our model where Dp427 and Dp71 expression may play a key role in the pathogenesis of tumours and/or may cooperate with existing cell mutations and genomic instability to influence the progression to full neoplastic disease. In a proposed model of *DMD* driven cancer development the relative balance of the Dp427 and Dp71 gene products could influence the progression to full neoplastic disease (Figure 4.2). The rationale for this model arises from soft tissue sarcomas studies where recurrent mutation largely restricted to the 5' region of the gene abrogates Dp427 expression but retains Dp71 expression (Mauduit *et al.*, 2019). Knockdown of Dp71 results in reduced proliferation and cell cycle progression. This work built upon earlier studies where intragenic deletions of *DMD* were frequently found (63%) in high grade myogenic tumors (Wang *et al.*, 2014). Restoration of dystrophin expression with a miniDMD construct (240-kDa dystrophin product) in DMD-inactivated GIST, eRMS and LMS cells inhibits invasiveness, migration, invadeapodia formation and anchorage independent growth. To confirm these findings in additional cancer backgrounds including those identified herein, functional cell experiments will be required that overexpress or knockdown/delete dystrophin variants in these different

tumour settings. This is ongoing work in our laboratory where, for example, we are currently overexpressing specific Dp71 variants in glioma cell lines.

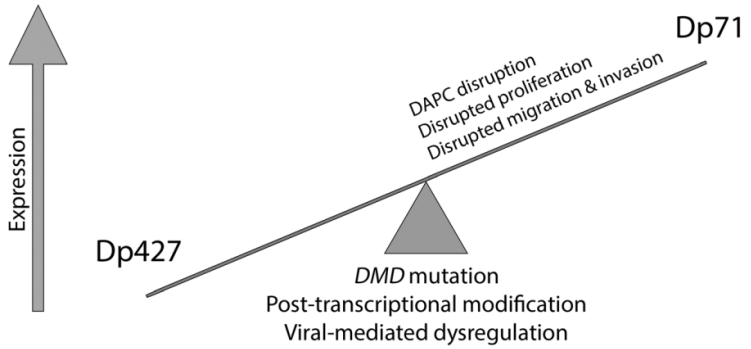


Figure 4.2 Proposed model of *DMD* driven cancer development
The balance of Dp71 and Dp427 gene products contributes to neoplasia. Altered *DMD* gene product levels have tissue specific effects on cancer hallmarks, such as proliferation and invasion, and disrupt the dystrophin associated protein complex (DAPC) (Jones *et al.*, 2021).

In a recent pre-publication, *DMD* gene expression was characterised across 25 TCGA cancers and their corresponding normal tissues (where possible). This work showed that the largest transcript Dp427 was downregulated in most tumours compared with healthy tissue (Alnassar *et al.*, 2022). Dp71 expression had variable transcript expression and a 10-gene signature could identify discrete disease clusters. Pathway analysis was used from cell line transcriptomic data was used to identify putative functional pathways (i.e. ECM-receptor interactions) that are implicated in those tumour types (Alnassar *et al.*, 2022). A limitation of this work was the lack of comprehensive survival analysis. However, in pooling survival data across 14 carcinomas and sarcoma, the overall survival of these patients with decreased *DMD* expression in tumours was 27 months lower than that of patients with high *DMD* expression

(Alnassar *et al.*, 2022). The survival analysis herein indicates pooling across so many cancer types with distinct survival characteristics may not be instructive. Furthermore, the authors compared patients at the bottom 25% of *DMD* expression values and those at the top 25% of *DMD* expression. This approach of dichotomising a continuous covariate based on percentiles, medians, means or a proposed clinical threshold value are arbitrary and may miss the true prognostic value of a putative biomarker. Our cutpoint approach incorporates an outcome based method where the optimal cutpoint is defined by the threshold of the distribution which optimally separates low and high risk patients with respect to an outcome (i.e. overall survival) (Williams *et al.*, 2006). In this project, the outcome-based method is based on log rank statistics. Typically, in time-to-event analysis, outcome-oriented methods perform better than data orientated methods (Mandrekar, n.d.). In addition, Alnassar *et al.*, 2022 focuses on the top 10 expressed *DMD* transcripts across all tissues, and therefore may miss interesting findings with novel highly expressed transcripts in a specific tumour type. Notably, analysis of Dp40 expression was not reported which was expressed in all nine tumours in this study.

Having dichotomised patients into high or low gene expression groups based on total *DMD* expression, preliminary differential transcriptome expression analysis allowed identification of putative biological pathways impacted across the nine tumours examined. In many cases GO Biological terms related to motility and adhesion were identified which is unsurprising given the role of *DMD* as a structural/scaffold protein that facilitates cellular interaction of the actin cytoskeleton with the extracellular matrix. However, in some cancers novel terms relating to Ion homeostasis (PAAD and READ) and chemical/sensory perception (LUAD) were identified and the biological significance of this is currently unclear.

4.2 Is *DMD* a driver gene in cancer?

DMD is not considered a classic cancer driver but in a recent publication, deep learning was used to reveal the exclusive functional contributions of individual cancer mutations (Gupta *et al.*, 2022). Their approach enabled identification and future exploration of putative driver genes including *DMD*, *RSK4*, *OFD1*, *WDR44*, and *AFF2*. This approach described a newly developed technique, Continuous Representation of Codon Switches (CRCS), that enabled the generation of numerical vector representations of mutations, applicable in several machine learning-based tasks. One task involved the authors constructing a novel deep learning architecture constituting bidirectional long short-term memory with attention & CRCS embeddings (BLAC) and demonstrated that a substantial chunk of cancer mutations are distinguishable from noncancer mutations. The model differentiated between driver gene-specific noncancerous and cancerous mutations and by merging multiple driver gene databases they identified 33 potential driver genes on the X chromosome including *DMD*. There is considerable scope for the use of novel machine learning approaches applied to gene expression analysis for cancer prediction and this has been summarised in a recent review (Khalsan *et al.*, 2022).

As discussed in our prior review (Jones *et al.*, 2021) ongoing studies should determine whether *DMD* acts as a driver or passenger in neoplasia. In some tumours *DMD* is frequently altered (i.e., single base mutations and copy number alterations) and varies across cancers (Figure 1.5). In some cases, there is clear evidence that recurrent mutations abolish expression of *DMD* gene products (i.e., Dp427 in soft tissue sarcomas) and/or specific focal functional mutations in domains such as the actin-binding domain as seen in meningioma (Juratli *et al.*, 2018). The results presented in this thesis suggest that dysregulated gene expression of dystrophin and/or the DAPC, through non-mutational mechanisms, may also be

relevant for disease development (i.e., in virus-induced alterations in gene expression programmes or epigenetic modifications).

4.3 Do Duchenne and Becker patients have an increased risk of cancer?

Given the increasingly recognised association and putative functional role(s) of *DMD* in cancer, an obvious question arises as to whether patients with dystrophinopathies (e.g. Becker and Duchenne) are at greater risk of cancer. The challenge(s) in addressing this issue are considerable as DMD and BMD are rare diseases and historically patients (typically boys) do not survive long enough for cancer to be an issue. However, with improved treatment/management patients are living longer and some specific case studies have now been identified from the literature (Table 4.1). For example, an interesting case report identified Rhabdomyosarcoma (RMS) in a Patient with Duchenne muscular dystrophy (Chandler *et al*., 2021). These cases provide interesting insights and highlight the need to explore further the connection between DMD and cancer. The authors reported a case of alveolar rhabdomyosarcoma (ARMS) in a five-year-old male with DMD who showed stable disease after radiotherapy and maintenance chemotherapy (Chandler *et al*., 2021). The co-occurrence of RMS with DMD in this individual is likely not a coincidence. Indeed, and although missed in their assessment of the literature in this area (where the authors state their subject was the third such case of a DMD patient developing RMS) there were at least another four cases reported, two of which are ARMS (Vita *et al*., 2021; Saldanha *et al*., 2005; Büget *et al*., 2014). Moreover, there are an additional seven case reports of DMD co-occurring with other cancer types, four of which are tumours of the central nervous system (Johnston *et al.*, 1986; Svarch *et al.*, 1988; D *et al.*, 2001; Doddihal and Jalali, 2007; Van Den Akker *et al.*, 2012; Vita *et al.*, 2021).

Table 4.1 Published case reports of cancer in individuals with *DMD*.

Publication year [reference]	*DMD* gene mutation	Age at cancer diagnosis (years)	Tumour type	Outcome
1986 (Johnston *et al.*, 1986)	Unknown	0.75	Stage III neuroblastoma	Alive > 25 months after surgery and chemotherapy
1988 (Svarch *et al.*, 1988)	Unknown	Unknown	Acute lymphoblastic leukaemia	Unknown
1999 (Rossbach *et al.*, 1999)	Exon 47-50 deletion	4	Alveolar RMS	Alive after chemotherapy and radiation
2001 (Korones *et al.*, 2001)	Unknown	3	Stage III Wilms tumour	Alive 6 months after surgery, chemotherapy and radiation
2002 (Jakab *et al.*, 2002)	Unknown	7	Stage II embryonal RMS	Dead 2.5 years after combined therapy
2005 (Saldanha *et al.*, 2005)	Unknown	5	RMS	Unknown after surgery
2007 (Doddihal and Jalali, 2007)	Exon 44 deletion	7	Medulloblastoma	Treated with surgery and radiation; tumour progressed; alive at 8 months post-treatment
2012 (Van Den Akker *et al.*, 2012)	Point mutation in exon 32 (c.4483C > T)	9	Anaplastic medulloblastoma	Alive 30 months after surgery, chemotherapy and radiation
2014 (Büget *et al.*, 2014)	Unknown	17	Massive RMS	Unknown after discharge following left arm amputation
2021 ((Vita *et al.*, 2021)	Exon 46-47 deletion	35	Brain tumour	Dead 50 days after onset
	Exon 48-50 deletion	14	Alveolar RMS	Dead 9 months after onset
	No deletions/duplications	17	Alveolar RMS	Dead after 1 year from lung metastases
	Exon 45-54 deletion	11	Enchondroma	Dead after 13 years from *DMD*-related respiratory failure
2021(Chandler *et al.*, 2021)	Exon 45-62 deletion	5	RMS	Stable disease after radiotherapy and on maintenance chemo
2022 (Okuno *et al.*, 2022)	Unknown	9	Alveolar RMS	Died due to ARMS exacerbation 5 months after treatment (chemo and radio) interruption

Chandler *et al*. state there is no literature examining the prevalence of specific *DMD* mutations amongst DMD individuals with RMS. This was confirmed in our review article (Jones *et al.*, 2021) and in a subsequent study (Vita *et al*., 2021). It should be noted that for most case reports the location and/or type of *DMD* mutation was unknown or unreported precluding definitive conclusions from such a small cohort. Our comprehensive review of the literature (Jones *et al.*, 2021) surrounding the *DMD* gene and cancer prompted Vita *et al*. to undertake the first dedicated study to test this through examination of patient records from all Italian Neuromuscular Centres for incidence of cancer (Vita *et al*., 2021). Their data

suggests that, considering the lower risk of cancer in children, DMD individuals may indeed have an increased risk and annual incidence (6/100,000) of RMS in Italian DMD population developing RMS (Vita *et al*., 2021). This compares with annual childhood RMS incidence in the USA is 0.44/100,000 (alveolar RMS is 0.15/100,000) suggesting up to a 40-fold increase over what one would have expected for alveolar RMS in childhood.

In summary, the incidence of cancer in DMD people has been historically under-reported and novel studies posit that RMS may be more common than is suggested from the case report literature by Chandler *et al*. Speculatively, the degenerative muscle environment of individuals with Duchenne may promote the development of RMS through increased tissue turnover. In summary, multicentre reviews and fundamental investigations into the role of the *DMD* gene in tumorigenesis are urgently required.

4.4 Future work and limitations

As mentioned, multicentre reviews and fundamental investigations into the role of the *DMD* gene in tumorigenesis are urgently required as:

1. There is support for an increased risk of RMS in people with DMD
2. Therapeutic advances may not mitigate cancer risk
3. Increasingly, people with DMD are living longer

Therefore, a comprehensive population-based survey to ascertain the risk and incidence of cancer in DMD is warranted which we are in the process of setting up to access data from the paediatric North Star UK national neuromuscular database which was established in 2003 to help drive improvements in services and set national standards of care for children living with Duchenne muscular dystrophy (DMD).

During this project, univariate analyses were done reporting hazard ratios where *DMD* was the sole covariate. Associations with other clinical co-variates have not been determined or whether *DMD* expression as a potential biomarker is confounded by other clinic-pathological variables. In part, this was to simplify analysis as each cancer type is likely to have specific additional covariates that are important for prognosis but specific for that disease. For example, in LGG, a multivariate model containing *DMD*, IDH status, tumour subtype and age revealed both *DMD* and IDH status remained independent prognostic markers (Naidoo *et al.*, 2022). IDH status is a valuable biomarker currently used in clinical practice for LGG and is included in recently updated WHO criteria (Bale and Rosenblum, 2022).

It is currently unclear whether survival associations with *DMD* (and variant gene products) replicate in independent disease cohorts. We employed the Chinese Glioma Genome Atlas (CGGA) dataset for our LGG study (Naidoo *et al.*, 2022) but this has not yet been done for the other eight tumour types.

Survival analysis was limited to overall survival although there are other outcome measures that could have been determined from the raw data (disease specific survival, progression free survival, disease-free interval). There are pros and cons to using these different outcome measures which has been previously discussed and evaluated (Liu *et al.*, 2018). For example, short-term clinical follow-up favours outcome analyses for more aggressive cancers, as multiple events are observed within a short timeframe. Studies with less aggressive cancers, where patients relapse after years or decades, may observe too few events during their follow-up intervals to support reliable outcome determinations. Of the thirty-three tumours analysed, OS was not recommended for 4 tumour types (DLBC, PCPG, TGCT and THYM). As *DMD* was associated with survival outcomes in THYM, this result should be considered more

cautiously as this report suggested the number of events was too small for OS and DSS; with a longer follow-up needed. Only PFI was recommended for THYM.

Our analysis of the altered transcriptomic response between high and low *DMD* expressing cases is exploratory and more detailed analysis is possible and required using other R packages available in iDEP. The functional follow-on experiments from this will test whether biological pathways (i.e. motility, morphogenesis and developmental programmes in LGG) are relevant for the identified cancer.

Subset analysis (i.e. by different histologic sites, anatomical location or existing clinical biomarkers) was beyond the scope of this project, however, it is clear from our LGG study (Naidoo *et al.*, 2022) that *DMD* expression can further stratify patients based on these other clinical markers. For example, *DMD* expression further stratified IDH mutant LGG to identify those at risk of poor survival. This knowledge may improve risk stratification and management of LGG.

Ultimately, RNA expression does not necessarily translate to equivalent protein levels and so for each tumour, accurate characterisation of protein expression (both levels and tissue distribution) is needed. Similar survival analysis can then be employed based on the association of Dystrophin protein expression and survival outcomes. A current caveat is the lack of suitable antibody reagents that distinguish the different Dp protein gene products expressed in cancer tissue. However, we are currently screening cell lines, and tissues by western blotting to better address dystrophin expression. Finally, the composition of the DAPC in different tumours is a fertile area for future research using either mass spectrometry or immunoprecipitation approaches to clarify the relevant molecular players.

4.5 Conclusions

DMD (both total and specific gene product) RNA expression was associated with the overall survival outcomes of patients across nine different cancers. These included BRCA, KIRP, LAML, LGG, LUAD, PAAD, READ, THYM and UVM. *DMD* is associated with the DAPC and other genes encoding proteins in the complex are also associated with different survival hazards in these tumours. Genes differentially expressed between high and low total *DMD* expressing cases can be used to define putative biological pathways dysregulated during disease. Functional studies will help unlock the importance of these pathways in *DMD*-associated tumourigenesis.

5. References

Abbadi, N., Philippe, C., Chery, M., Gilgenkrantz, H., Tome, F., Collin, H., Theau, D., Recan, D., Broux, O., Fardeau, M., Kaplan, J.C., Gilgenkrantz, S. (1994) Additional case of female monozygotic twins discordant for the clinical manifestations of Duchenne muscular dystrophy due to opposite X-chromosome inactivation. *American journal of medical genetics*. **52**(2), 198–206.

Acharyya, S., Butchbach, M.E.R., Sahenk, Z., Wang, H., Saji, M., Carathers, M., Ringel, M.D., Skipworth, R.J.E., Fearon, K.C.H., Hollingsworth, M.A., Muscarella, P., Burghes, A.H.M., Rafael-Fortney, J.A., Guttridge, D.C. (2005) Dystrophin glycoprotein complex dysfunction: A regulatory link between muscular dystrophy and cancer cachexia. *Cancer Cell*. **8**(5), 421–432.

Van Den Akker, M., Northcott, P., Taylor, M.D., Halliday, W., Bartels, U., Bouffet, E. (2012) Anaplastic medulloblastoma in a child with Duchenne muscular dystrophy: Case report. *Journal of Neurosurgery: Pediatrics*. **10**(1), 21–24.

Alnassar, N., Borczyk, M., Tsagkogeorga, G., Korostynski, M., Han, N., Górecki, D.C. (2022) Loss of DMD gene expression results in Duchenne-like molecular abnormalities across diverse tissues. *bioRxiv*, 2022.04.04.486990.

Bale, T.A., Rosenblum, M.K. (2022) The 2021 WHO Classification of Tumors of the Central Nervous System: An update on pediatric low-grade gliomas and glioneuronal tumors. *Brain Pathology*. **32**(4), e13060.

Baumforth, K.R.N., Birgersdotter, A., Reynolds, G.M., Wei, W., Kapatai, G., Flavell, J.R., Kalk, E., Piper, K., Lee, S., Machado, L., others, Hadley, K., Sundblad, A., Sjoberg, J., Bjorkholm, M., Porwit, A. a, Yap, L.-F., Teo, S., Grundy, R.G., Young, L.S., Ernberg, I., Woodman, C.B.J., Murray, P.G. (2008) Expression of the Epstein-Barr virus-encoded Epstein-Barr virus nuclear antigen 1 in Hodgkin's lymphoma cells mediates up-regulation of CCL20 and the migration of regulatory T cells. *The American journal of pathology*. **173**(1), 195–204.

Bradburn, M.J., Clark, T.G., Love, S.B., Altman, D.G. (2003a) Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods. *British Journal of Cancer 2003 89:3*. **89**(3), 431–436.

Bradburn, M.J., Clark, T.G., Love, S.B., Altman, D.G. (2003b) Survival Analysis Part III: Multivariate data analysis – choosing a model and assessing its adequacy and fit. *British Journal of Cancer 2003 89:4*. **89**(4), 605–611.

Brennan, P.A., Jing, J., Ethunandan, M., Górecki, D. (2004) Dystroglycan complex in cancer. *European Journal of Surgical Oncology*. **30**(6), 589–592.

Büget, M.I., Eren, I., Küçükay, S. (2014) Regional anaesthesia in a Duchenne muscular dystrophy patient for upper extremity amputation. *Agri*. **26**(4), 191–195.

Calogero, A., Pavoni, E., Gramaglia, T., D'Amati, G., Ragona, G., Brancaccio, A., Petrucci, T.C. (2006) Altered expression of α-dystroglycan subunit in human gliomas. *Cancer Biology and Therapy*. **5**(4), 441–448.

Camp, R.L., Dolled-Filhart, M., Rimm, D.L. (2004) X-tile: A new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clinical Cancer Research*. **10**(21), 7252–7259.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A.P., Sander, C., Schultz, N. (2012) The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery*. **2**(5), 401–404.

Cereda, S., Cefalo, G., Terenziani, M., Catania, S., Fossati-Bellani, F. (2004) Becker Muscular Dystrophy in a Patient with Hodgkin's Disease. *Journal of Pediatric Hematology/Oncology*. **26**(1), 72–73.

Chamberlain, J.S., Metzger, J., Reyes, M., Townsend, D., Faulkner, J.A. (2007) Dystrophin-deficient mdx mice display a reduced life span and are susceptible to spontaneous rhabdomyosarcoma . *The FASEB Journal*. **21**(9), 2195–2204.

Chandler, E., Rawson, L., Debski, R., McGowan, K., Lakhotia, A. (2021) Rhabdomyosarcoma in a Patient With Duchenne Muscular Dystrophy: A Possible Association. *Child Neurology Open*. **8**, 2329048X2110414.

Cirak, S., Feng, L., Anthony, K., Arechavala-Gomeza, V., Torelli, S., Sewry, C., Morgan, J.E., Muntoni, F. (2012) Restoration of the dystrophin-associated glycoprotein complex after

exon skipping therapy in Duchenne muscular dystrophy. *Molecular therapy : the journal of the American Society of Gene Therapy*. **20**(2), 462–7.

Clark, T.G., Bradburn, M.J., Love, S.B., Altman, D.G. (2003) Survival Analysis Part I: Basic concepts and first analyses. *British Journal of Cancer*. **89**(2), 232.

Colaprico, A., Silva, T.C., Olsen, C., Garofano, L., Cava, C., Garolini, D., Sabedot, T.S., Malta, T.M., Pagnotta, S.M., Castiglioni, I., Ceccarelli, M., Bontempi, G., Noushmehr, H. (2016) TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*. **44**(8), e71.

Cox, G.A., Phelps, S.F., Chapman, V.M., Chamberlain, J.S. (1993) New mdx mutation disrupts expression of muscle and nonmuscle isoforms of dystrophin. *Nature Genetics 1993 4:1*. **4**(1), 87–93.

Cross, S.S., Lippitt, J., Mitchell, A., Hollingsbury, F., Balasubramanian, S.P., Reed, M.W.R., Eaton, C., Catto, J.W., Hamdy, F., Winder, S.J. (2008) Expression of β-dystroglycan is reduced or absent in many human carcinomas. *Histopathology*. **53**(5), 561–566.

D, K., MR, B., J, P. (2001) 'Liver Function Tests' Are Not Always Tests of Liver Function. *American journal of hematology*. **66**(1), 46–48.

Dobin, A., Gingeras, T.R. (2015) Mapping RNA-seq Reads with STAR. *Current Protocols in Bioinformatics*. **51**(1), 11.14.1-11.14.19.

Doddihal, H., Jalali, R. (2007) Medulloblastoma in a child with Duchenne muscular dystrophy. *Child's Nervous System*. **23**(5), 595–597.

Fernandez, K., Serinagaoglu, Y., Hammond, S., Martin, L.T., Martin, P.T. (2010) Mice lacking dystrophin or α sarcoglycan spontaneously develop embryonal rhabdomyosarcoma with cancer-associated p53 mutations and alternatively spliced or mutant Mdm2 transcripts. *American Journal of Pathology*. **176**(1), 416–434.

Gallia, G.L., Zhang, M., Ning, Y., Haffner, M.C., Batista, D., Binder, Z.A., Bishop, J.A., Hann, C.L., Hruban, R.H., Ishii, M., Klein, A.P., Reh, D.D., Rooper, L.M., Salmasi, V., Tamargo, R.J., Wang, Q., Williamson, T., Zhao, T., Zou, Y., Meeker, A.K., Agrawal, N., Vogelstein, B., Kinzler, K.W., Papadopoulos, N., Bettegowda, C. (2018) Genomic analysis identifies frequent

deletions of Dystrophin in olfactory neuroblastoma. *Nature communications*. **9**(1), 5410.

Gao, G.F., Parker, J.S., Reynolds, S.M., Silva, T.C., Wang, L.B., Zhou, W., Akbani, R., Bailey, M., Balu, S., Berman, B.P., Brooks, D., Chen, H., Cherniack, A.D., Demchok, J.A., Ding, L., Felau, I., Gaheen, S., Gerhard, D.S., Heiman, D.I., Hernandez, K.M., Hoadley, K.A., Jayasinghe, R., Kemal, A., Knijnenburg, T.A., Laird, P.W., Mensah, M.K.A., Mungall, A.J., Robertson, A.G., Shen, H., Tarnuzzer, R., Wang, Z., Wyczalkowski, M., Yang, L., Zenklusen, J.C., Zhang, Z., Liang, H., Noble, M.S. (2019) Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Systems*. **9**(1), 24-34.e10.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., Schultz, N. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*. **6**(269), pl1.

Gao, Q.Q., McNally, E.M. (2015) The Dystrophin Complex: Structure, Function, and Implications for Therapy. *Comprehensive Physiology*. **5**(3), 1223–1239.

García-Cruz, C., Merino-Jiménez, C., Aragón, J., Ceja, V., González-Assad, B., Reyes-Grajeda, J.P., Montanez, C. (2022) Overexpression of the dystrophins Dp40 and Dp40L170P modifies neurite outgrowth and the protein expression profile of PC12 cells. *Scientific Reports 2022 12:1*. **12**(1), 1–11.

GDC (2022) Genomic Data Commons. *Genomic Data Commons*. [online]. Available from: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/images/gene-expression-quantification-pipeline-v3.png [Accessed August 25, 2022].

Ge, S.X., Son, E.W., Yao, R. (2018) iDEP: An integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics*. **19**(1), 1–24.

Gupta, P., Jindal, A., Ahuja, G., Jayadeva, Sengupta, D. (2022) A new deep learning technique reveals the exclusive functional contributions of individual cancer mutations. *Journal of Biological Chemistry*. **298**(8), 102177.

Hanahan, D. (2022) Hallmarks of Cancer: New Dimensions. *Cancer discovery*. **12**(1), 31–46.

Jakab, Z., Szegedi, I., Balogh, E., Kiss, C., Oláh, É. (2002) Duchenne muscular dystrophy-

rhabdomyosarcoma, ichthyosis vulgaris/acute monoblastic leukemia: Association of rare genetic disorders and childhood malignant diseases. *Medical and Pediatric Oncology*. **39**(1), 66–68.

Johnston, K.M., Zoger, S., Golabi, M., Mulvihill, J.J. (1986) Neuroblastoma in Duchenne muscular dystrophy. *Pediatrics*. **78**(6), 1170–1171.

Jones, L., Naidoo, M., Machado, L.R., Anthony, K. (2021) The Duchenne muscular dystrophy gene and cancer. *Cellular Oncology*. **44**(1), 19–32.

Juratli, T.A., McCabe, D., Nayyar, N., Williams, E.A., Silverman, I.M., Tummala, S.S., Fink, A.L., Baig, A., Martinez-Lage, M., Selig, M.K., Bihun, I. V., Shankar, G.M., Penson, T., Lastrapes, M., Daubner, D., Meinhardt, M., Hennig, S., Kaplan, A.B., Fujio, S., Kuter, B.M., Bertalan, M.S., Miller, J.J., Batten, J.M., Ely, H.A., Christiansen, J., Baretton, G.B., Stemmer-Rachamimov, A.O., Santagata, S., Rivera, M.N., Barker, F.G., Schackert, G., Wakimoto, H., Iafrate, A.J., Carter, S.L., Cahill, D.P., Brastianos, P.K. (2018) DMD genomic deletions characterize a subset of progressive/higher-grade meningiomas with poor outcome. *Acta Neuropathologica*. **136**(5), 779–792.

Kahana, E., Gratzer, W.B. (1995) Minimum folding unit of dystrophin rod domain. *Biochemistry*. **34**(25), 8110–8114.

Kaplan, E.L., Meier, P. (1958) Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*. **53**(282), 457–481.

Khalsan, M., MacHado, L.R., Al-Shamery, E.S., Ajit, S., Anthony, K., Mu, M., Agyeman, M.O. (2022) A Survey of Machine Learning Approaches Applied to Gene Expression Analysis for Cancer Prediction. *IEEE Access*. **10**, 27522–27534.

Körner, H., Epanchintsev, A., Berking, C., Schuler-Thurner, B., Speicher, M.R., Menssen, A., Hermeking, H. (2007) Digital karyotyping reveals frequent inactivation of the Dystrophin/DMD gene in malignant melanoma. *Cell Cycle*. **6**(2), 189–198.

Korones, D.N., Brown, M.R., Palis, J. (2001) 'Liver Function Tests' Are Not Always Tests of Liver Function. *J. Hematol*. **66**, 46–48.

Lausen, B., Schumacher, M. (1992) Maximally Selected Rank Statistics. *Biometrics*. **48**(1), 73.

Lawlor, N., Fabbri, A., Guan, P., George, J., Karuturi, R.K.M. (2016) MultiClust: An R-package for identifying biologically relevant clusters in cancer transcriptome profiles. *Cancer Informatics*. **15**, 103–114.

Liu, Jianfang, Lichtenberg, T., Hoadley, K.A., Poisson, L.M., Lazar, A.J., Cherniack, A.D., Kovatich, A.J., Benz, C.C., Levine, D.A., Lee, A. V., Omberg, L., Wolf, D.M., Shriver, C.D., Thorsson, V., Caesar-Johnson, S.J., Demchok, J.A., Felau, I., Kasapi, M., Ferguson, M.L., Hutter, C.M., Sofia, H.J., Tarnuzzer, R., Wang, Z., Yang, L., Zenklusen, J.C., Zhang, J. (Julia), Chudamani, S., Liu, Jia, Lolla, L., Naresh, R., Pihl, T., Sun, Q., Wan, Y., Wu, Y., Cho, J., DeFreitas, T., Frazer, S., Gehlenborg, N., Getz, G., Heiman, D.I., Kim, J., Lawrence, M.S., Lin, P., Meier, S., Noble, M.S., Saksena, G., Voet, D., Zhang, Hailei, Bernard, B., Chambwe, N., Dhankani, V., Knijnenburg, T., Kramer, R., Leinonen, K., Liu, Y., Miller, M., Reynolds, S., Shmulevich, I., Thorsson, V., Zhang, W., Akbani, R., Broom, B.M., Hegde, A.M., Ju, Z., Kanchi, R.S., Korkut, A., Li, J., Liang, H., Ling, S., Liu, W., Lu, Y., Mills, G.B., Ng, K.S., Rao, A., Ryan, M., Wang, Jing, Weinstein, J.N., Zhang, J., Abeshouse, A., Armenia, J., Chakravarty, D., Chatila, W.K., de Bruijn, I., Gao, J., Gross, B.E., Heins, Z.J., Kundra, R., La, K., Ladanyi, M., Luna, A., Nissan, M.G., Ochoa, A., Phillips, S.M., Reznik, E., Sanchez-Vega, F., Sander, C., Schultz, N., Sheridan, R., Sumer, S.O., Sun, Y., Taylor, B.S., Wang, Jioajiao, Zhang, Hongxin, Anur, P., Peto, M., Spellman, P., Benz, C., Stuart, J.M., Wong, C.K., Yau, C., Hayes, D.N., Parker, J.S., Wilkerson, M.D., Ally, A., Balasundaram, M., Bowlby, R., Brooks, D., Carlsen, R., Chuah, E., Dhalla, N., Holt, R., Jones, S.J.M., Kasaian, K., Lee, D., Ma, Y., Marra, M.A., Mayo, M., Moore, R.A., Mungall, A.J., Mungall, K., Robertson, A.G., Sadeghi, S., Schein, J.E., Sipahimalani, P., Tam, A., Thiessen, N., Tse, K., Wong, T., Berger, A.C., Beroukhim, R., Cherniack, A.D., Cibulskis, C., Gabriel, S.B., Gao, G.F., Ha, G., Meyerson, M., Schumacher, S.E., Shih, J., Kucherlapati, M.H., Kucherlapati, R.S., Baylin, S., Cope, L., Danilova, L., Bootwalla, M.S., Lai, P.H., Maglinte, D.T., Van Den Berg, D.J., Weisenberger, D.J., Auman, J.T., Balu, S., Bodenheimer, T., Fan, C., Hoadley, K.A., Hoyle, A.P., Jefferys, S.R., Jones, C.D., Meng, S., Mieczkowski, P.A., Mose, L.E., Perou, A.H., Perou, C.M., Roach, J., Shi, Y., Simons, J. V., Skelly, T., Soloway, M.G., Tan, D., Veluvolu, U., Fan, H., Hinoue, T., Laird, P.W., Shen, H., Zhou, W., Bellair, M., Chang, K., Covington, K., Creighton, C.J., Dinh, H., Doddapaneni, H.V., Donehower, L.A., Drummond, J., Gibbs, R.A., Glenn, R., Hale, W., Han, Y., Hu, J., Korchina, V., Lee, S., Lewis, L., Li, W., Liu, X., Morgan, M., Morton, D., Muzny, D., Santibanez, J., Sheth, M.,

Shinbro, E., Wang, L., Wang, M., Wheeler, D.A., Xi, L., Zhao, F., Hess, J., Appelbaum, E.L., Bailey, M., Cordes, M.G., Ding, L., Fronick, C.C., Fulton, L.A., Fulton, R.S., Kandoth, C., Mardis, E.R., McLellan, M.D., Miller, C.A., Schmidt, H.K., Wilson, R.K., Crain, D., Curley, E., Gardner, J., Lau, K., Mallery, D., Morris, S., Paulauskis, J., Penny, R., Shelton, C., Shelton, T., Sherman, M., Thompson, E., Yena, P., Bowen, J., Gastier-Foster, J.M., Gerken, M., Leraas, K.M., Lichtenberg, T.M., Ramirez, N.C., Wise, L., Zmuda, E., Corcoran, N., Costello, T., Hovens, C., Carvalho, A.L., de Carvalho, A.C., Fregnani, J.H., Longatto-Filho, A., Reis, R.M., Scapulatempo-Neto, C., Silveira, H.C.S., Vidal, D.O., Burnette, A., Eschbacher, J., Hermes, B., Noss, A., Singh, R., Anderson, M.L., Castro, P.D., Ittmann, M., Huntsman, D., Kohl, B., Le, X., Thorp, R., Andry, C., Duffy, E.R., Lyadov, V., Paklina, O., Setdikova, G., Shabunin, A., Tavobilov, M., McPherson, C., Warnick, R., Berkowitz, R., Cramer, D., Feltmate, C., Horowitz, N., Kibel, A., Muto, M., Raut, C.P., Malykh, A., Barnholtz-Sloan, J.S., Barrett, W., Devine, K., Fulop, J., Ostrom, Q.T., Shimmel, K., Wolinsky, Y., Sloan, A.E., De Rose, A., Giuliante, F., Goodman, M., Karlan, B.Y., Hagedorn, C.H., Eckman, J., Harr, J., Myers, J., Tucker, K., Zach, L.A., Deyarmin, B., Hu, H., Kvecher, L., Larson, C., Mural, R.J., Somiari, S., Vicha, A., Zelinka, T., Bennett, J., Iacocca, M., Rabeno, B., Swanson, P., Latour, M., Lacombe, L., Têtu, B., Bergeron, A., McGraw, M., Staugaitis, S.M., Chabot, J., Hibshoosh, H., Sepulveda, A., Su, T., Wang, T., Potapova, O., Voronina, O., Desjardins, L., Mariani, O., Roman-Roman, S., Sastre, X., Stern, M.H., Cheng, F., Signoretti, S., Berchuck, A., Bigner, D., Lipp, E., Marks, J., McCall, S., McLendon, R., Secord, A., Sharp, A., Behera, M., Brat, D.J., Chen, A., Delman, K., Force, S., Khuri, F., Magliocca, K., Maithel, S., Olson, J.J., Owonikoko, T., Pickens, A., Ramalingam, S., Shin, D.M., Sica, G., Van Meir, E.G., Zhang, Hongzheng, Eijckenboom, W., Gillis, A., Korpershoek, E., Looijenga, L., Oosterhuis, W., Stoop, H., van Kessel, K.E., Zwarthoff, E.C., Calatozzolo, C., Cuppini, L., Cuzzubbo, S., DiMeco, F., Finocchiaro, G., Mattei, L., Perin, A., Pollo, B., Chen, C., Houck, J., Lohavanichbutr, P., Hartmann, A., Stoehr, C., Stoehr, R., Taubert, H., Wach, S., Wullich, B., Kycler, W., Murawa, D., Wiznerowicz, M., Chung, K., Edenfield, W.J., Martin, J., Baudin, E., Bubley, G., Bueno, R., De Rienzo, A., Richards, W.G., Kalkanis, S., Mikkelsen, T., Noushmehr, H., Scarpace, L., Girard, N., Aymerich, M., Campo, E., Giné, E., Guillermo, A.L., Van Bang, N., Hanh, P.T., Phu, B.D., Tang, Y., Colman, H., Evason, K., Dottino, P.R., Martignetti, J.A., Gabra, H., Juhl, H., Akeredolu, T., Stepa, S., Hoon, D., Ahn, K., Kang, K.J., Beuschlein, F., Breggia, A., Birrer, M., Bell, D., Borad, M., Bryce, A.H., Castle, E., Chandan, V., Cheville, J., Copland, J.A., Farnell, M., Flotte, T., Giama, N., Ho, T., Kendrick, M.,

Kocher, J.P., Kopp, K., Moser, C., Nagorney, D., O'Brien, D., O'Neill, B.P., Patel, T., Petersen, G., Que, F., Rivera, M., Roberts, L., Smallridge, R., Smyrk, T., Stanton, M., Thompson, R.H., Torbenson, M., Yang, J.D., Zhang, L., Brimo, F., Ajani, J.A., Angulo Gonzalez, A.M., Behrens, C., Bondaruk, J., Broaddus, R., Czerniak, B., Esmaeli, B., Fujimoto, J., Gershenwald, J., Guo, C., Logothetis, C., Meric-Bernstam, F., Moran, C., Ramondetta, L., Rice, D., Sood, A., Tamboli, P., Thompson, T., Troncoso, P., Tsao, A., Wistuba, I., Carter, C., Haydu, L., Hersey, P., Jakrot, V., Kakavand, H., Kefford, R., Lee, K., Long, G., Mann, G., Quinn, M., Saw, R., Scolyer, R., Shannon, K., Spillane, A., Stretch, J., Synott, M., Thompson, J., Wilmott, J., Al-Ahmadie, H., Chan, T.A., Ghossein, R., Gopalan, A., Levine, D.A., Reuter, V., Singer, S., Singh, B., Tien, N.V., Broudy, T., Mirsaidi, C., Nair, P., Drwiega, P., Miller, J., Smith, J., Zaren, H., Park, J.W., Hung, N.P., Kebebew, E., Linehan, W.M., Metwalli, A.R., Pacak, K., Pinto, P.A., Schiffman, M., Schmidt, L.S., Vocke, C.D., Wentzensen, N., Worrell, R., Yang, H., Moncrieff, M., Goparaju, C., Melamed, J., Pass, H., Botnariuc, N., Caraman, I., Cernat, M., Chemencedji, I., Clipca, A., Doruc, S., Gorincioi, G., Mura, S., Pirtac, M., Stancul, I., Tcaciuc, D., Albert, M., Alexopoulou, I., Arnaout, A., Bartlett, J., Engel, J., Gilbert, S., Parfitt, J., Sekhon, H., Thomas, G., Rassl, D.M., Rintoul, R.C., Bifulco, C., Tamakawa, R., Urba, W., Hayward, N., Timmers, H., Antenucci, A., Facciolo, F., Grazi, G., Marino, M., Merola, R., de Krijger, R., Gimenez-Roqueplo, A.P., Piché, A., Chevalier, S., McKercher, G., Birsoy, K., Barnett, G., Brewer, C., Farver, C., Naska, T., Pennell, N.A., Raymond, D., Schilero, C., Smolenski, K., Williams, F., Morrison, C., Borgia, J.A., Liptay, M.J., Pool, M., Seder, C.W., Junker, K., Omberg, L., Dinkin, M., Manikhas, G., Alvaro, D., Bragazzi, M.C., Cardinale, V., Carpino, G., Gaudio, E., Chesla, D., Cottingham, S., Dubina, M., Moiseenko, F., Dhanasekaran, R., Becker, K.F., Janssen, K.P., Slotta-Huspenina, J., Abdel-Rahman, M.H., Aziz, D., Bell, S., Cebulla, C.M., Davis, A., Duell, R., Elder, J.B., Hilty, J., Kumar, B., Lang, J., Lehman, N.L., Mandt, R., Nguyen, P., Pilarski, R., Rai, K., Schoenfield, L., Senecal, K., Wakely, P., Hansen, P., Lechan, R., Powers, J., Tischler, A., Grizzle, W.E., Sexton, K.C., Kastl, A., Henderson, J., Porten, S., Waldmann, J., Fassnacht, M., Asa, S.L., Schadendorf, D., Couce, M., Graefen, M., Huland, H., Sauter, G., Schlomm, T., Simon, R., Tennstedt, P., Olabode, O., Nelson, M., Bathe, O., Carroll, P.R., Chan, J.M., Disaia, P., Glenn, P., Kelley, R.K., Landen, C.N., Phillips, J., Prados, M., Simko, J., Smith-McCune, K., VandenBerg, S., Roggin, K., Fehrenbach, A., Kendler, A., Sifri, S., Steele, R., Jimeno, A., Carey, F., Forgie, I., Mannelli, M., Carney, M., Hernandez, B., Campos, B., Herold-Mende, C., Jungk, C., Unterberg, A., von Deimling, A., Bossler, A., Galbraith, J., Jacobus, L., Knudson, M., Knutson, T., Ma, D., Milhem,

M., Sigmund, R., Godwin, A.K., Madan, R., Rosenthal, H.G., Adebamowo, C., Adebamowo, S.N., Boussioutas, A., Beer, D., Giordano, T., Mes-Masson, A.M., Saad, F., Bocklage, T., Landrum, L., Mannel, R., Moore, K., Moxley, K., Postier, R., Walker, J., Zuna, R., Feldman, M., Valdivieso, F., Dhir, R., Luketich, J., Mora Pinero, E.M., Quintero-Aguilo, M., Carlotti, C.G., Dos Santos, J.S., Kemp, R., Sankarankuty, A., Tirapelli, D., Catto, J., Agnew, K., Swisher, E., Creaney, J., Robinson, B., Shelley, C.S., Godwin, E.M., Kendall, S., Shipman, C., Bradford, C., Carey, T., Haddad, A., Moyer, J., Peterson, L., Prince, M., Rozek, L., Wolf, G., Bowman, R., Fong, K.M., Yang, I., Korst, R., Rathmell, W.K., Fantacone-Campbell, J.L., Hooke, J.A., Kovatich, A.J., Shriver, C.D., DiPersio, J., Drake, B., Govindan, R., Heath, S., Ley, T., Van Tine, B., Westervelt, P., Rubin, M.A., Lee, J. Il, Aredes, N.D., Mariamidze, A., Hu, H. (2018) An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*. **173**(2), 400.

Lu, Z., Zou, J., Li, S., Topper, M.J., Tao, Y., Zhang, H., Jiao, X., Xie, W., Kong, X., Vaz, M., Li, H., Cai, Y., Xia, L., Huang, P., Rodgers, K., Lee, B., Riemer, J.B., Day, C.P., Yen, R.W.C., Cui, Y., Wang, Yujiao, Wang, Yanni, Zhang, W., Easwaran, H., Hulbert, A., Kim, K.B., Juergens, R.A., Yang, S.C., Battafarano, R.J., Bush, E.L., Broderick, S.R., Cattaneo, S.M., Brahmer, J.R., Rudin, C.M., Wrangle, J., Mei, Y., Kim, Y.J., Zhang, B., Wang, K.K.H., Forde, P.M., Margolick, J.B., Nelkin, B.D., Zahnow, C.A., Pardoll, D.M., Housseau, F., Baylin, S.B., Shen, L., Brock, M. V. (2020) Epigenetic therapy inhibits metastases by disrupting premetastatic niches. *Nature*. **579**(7798), 284–290.

Luce, L.N., Abbate, M., Cotignola, J., Giliberto, F. (2017) Non-myogenic tumors display altered expression of dystrophin (DMD) and a high frequency of genetic alterations. *Oncotarget*. **8**(1), 145–155.

Mandrekar, J.N. Cutpoint Determination Methods in Survival Analysis using SAS ®. *SUGI 28*. [online]. Available from: https://support.sas.com/resources/papers/proceedings/proceedings/sugi28/261-28.pdf [Accessed August 16, 2022].

Mathew, G., Mitchell, A., Down, J.M., Jacobs, L.A., Hamdy, F.C., Eaton, C., Rosario, D.J., Cross, S.S., Winder, S.J. (2013) Nuclear targeting of dystroglycan promotes the expression of androgen regulated transcription factors in prostate cancer. *Scientific Reports*. **3**(Sep 30),

2792.

Mauduit, O., Delcroix, V., Lesluyes, T., Pérot, G., Lagarde, P., Lartigue, L., Blay, J.-Y., Chibon, F. (2019) Recurrent DMD Deletions Highlight Specific Role of Dp71 Isoform in Soft-Tissue Sarcomas. *Cancers*. **11**(7), 922.

McAvoy, S., Ganapathiraju, S., Perez, D.S., James, C.D., Smith, D.I. (2007) DMD and IL1RAPL1: Two large adjacent genes localized within a common fragile site (FRAXC) have reduced expression in cultured brain tumors. *Cytogenetic and Genome Research*. **119**(3–4), 196–203.

Mitchell, A., Mathew, G., Jiang, T., Hamdy, F.C., Cross, S.S., Eaton, C., Winder, S.J. (2013) Dystroglycan function is a novel determinant of tumor growth and behavior in prostate cancer. *Prostate*. **73**(4), 398–408.

Muntoni, F., Torelli, S., Ferlini, A. (2003) Dystrophin and mutations: One gene, several proteins, multiple phenotypes. *Lancet Neurology*. **2**(12), 731–740.

Naidoo, M., Anthony, K. (2020) Dystrophin Dp71 and the Neuropathophysiology of Duchenne Muscular Dystrophy. *Molecular Neurobiology*. **57**(3), 1748–1767.

Naidoo, M., Jones, L., Conboy, B., Hamarneh, W., D'Souza, D., Anthony, K., Machado, L.R. (2022) Duchenne muscular dystrophy gene expression is an independent prognostic marker for IDH mutant low-grade glioma. *Scientific reports*. **12**(1), 19–32.

National Cancer Institute (2022) TCGA Research Network Publications. [online]. Available from: cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/publications [Accessed August 17, 2022].

NCI The Cancer Genome Atlas - Ethics and Policies. [online]. Available from: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/policies [Accessed August 8, 2022].

Nikitin, E.A., Malakho, S.G., Biderman, B. V., Baranova, A. V., Lorie, Y.Y., Shevelev, A.Y., Peklo, M.M., Vlasik, T.N., Moskalev, E.A., Zingerman, B. V., Vorob'ev, I.A., Poltaraus, A.B., Sudarikov, A.B., Vorobjev, A.I. (2007) Expression level of lipoprotein lipase and dystrophin genes predict survival in B-cell chronic lymphocytic leukemia. *Leukemia and Lymphoma*.

48(5), 912–922.

Ohlendieck, K., Ervasti, J.M., Matsumura, K., Kahl, S.D., Leveille, C.J., Campbell, K.P. (1991) Dystrophin-related protein is localized to neuromuscular junctions of adult skeletal muscle. *Neuron*. **7**(3), 499–508.

Okuno, K., Kawaba, D., Maejima, A., Kakee, S., Namba, N. (2022) A high-risk alveolar rhabdomyosarcoma case with Duchenne muscular dystrophy. *Pediatrics International*. **64**(1), e14754.

Olson, E.N. (2021) Toward the correction of muscular dystrophy by gene editing. *Proceedings of the National Academy of Sciences of the United States of America*. **118**(22), e2004840117.

Paramasivam, N., Hübschmann, D., Toprak, U.H., Ishaque, N., Neidert, M., Schrimpf, D., Stichel, D., Reuss, D., Sievers, P., Reinhardt, A., Wefers, A.K., Jones, D.T.W., Gu, Z., Werner, J., Uhrig, S., Wirsching, H.G., Schick, M., Bewerunge-Hudler, M., Beck, K., Brehmer, S., Urbschat, S., Seiz-Rosenhagen, M., Hänggi, D., Herold-Mende, C., Ketter, R., Eils, R., Ram, Z., Pfister, S.M., Wick, W., Weller, M., Grossmann, R., von Deimling, A., Schlesner, M., Sahm, F. (2019) Mutational patterns and regulatory networks in epigenetic subgroups of meningioma. *Acta Neuropathologica*. **138**(2), 295–308.

Percival, J.M. (2018) Perspective: Spectrin-like repeats in dystrophin have unique binding preferences for syntrophin adaptors that explain the mystery of how nNOSμ localizes to the sarcolemma. *Frontiers in Physiology*. **9**(Oct 8), 1369.

Peto, R., Pike, M.C., Armitage, P., Breslow, N.E., Cox, D.R., Howard, S. V., Mantel, N., McPherson, K., Peto, J., Smith, P.G. (1977) Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *British Journal of Cancer 1977 35:1*. **35**(1), 1–39.

Pilgram, G.S.K., Potikanond, S., Baines, R.A., Fradkin, L.G., Noordermeer, J.N. (2010) The roles of the dystrophin-associated glycoprotein complex at the synapse. *Molecular Neurobiology*. **41**(1), 1–21.

Rani, A.Q.M., Farea, M., Maeta, K., Kawaguchi, T., Awano, H., Nagai, M., Nishio, H., Matsuo, M. (2019) Identification of the shortest splice variant of Dp71, together with five known

variants, in glioblastoma cells. *Biochemical and Biophysical Research Communications*. **508**(2), 640–645.

van den Reek, J.M.P.A., Kievit, W., Gniadecki, R., Goeman, J.J., Zweegers, J., van de Kerkhof, P.C.M., Seyger, M.M.B., de Jong, E.M.G.J. (2015) Drug Survival Studies in Dermatology:Principles, Purposes, and Pitfalls. *Journal of Investigative Dermatology*. **135**(7), 1–5.

Romitti, P.A., Zhu, Y., Puzhankara, S., James, K.A., Nabukera, S.K., Zamba, G.K.D., Ciafaloni, E., Cunniff, C., Druschel, C.M., Mathews, K.D., Matthews, D.J., Meaney, F.J., Andrews, J.G., Caspers Conway, K.M., Fox, D.J., Street, N., Adams, M.M., Bolen, J. (2015) Prevalence of Duchenne and Becker muscular dystrophies in the United States. *Pediatrics*. **135**(3), 513–521.

Rossbach, H.C., Lacson, A., Grana, N.H., Barbosa, J.L. (1999) Duchenne muscular dystrophy and concomitant metastatic alveolar rhabdomyosarcoma. *Journal of Pediatric Hematology/Oncology*. **21**(6), 528–530.

Ruggieri, S., De Giorgis, M., Annese, T., Tamma, R., Notarangelo, A., Marzullo, A., Senetta, R., Cassoni, P., Notarangelo, M., Ribatti, D., Nico, B. (2019) Dp71 expression in human glioblastoma. *International Journal of Molecular Sciences*. **20**(21), 5429.

Saldanha, R.M., Gasparini, J.R., Silva, L.S., de Carli, R.R., Castilhos, V.U.D. de, Neves, M.M.P. das, Araújo, F.P., Sales, P.C. de A., Neves, J.F.N.P. das (2005) Anestesia em paciente portador de distrofia muscular de Duchenne: relato de casos. *Revista Brasileira de Anestesiologia*. **55**(4), 445–449.

Schmidt, W.M., Uddin, M.H., Dysek, S., Moser-Thier, K., Pirker, C., Höger, H., Ambros, I.M., Ambros, P.F., Berger, W., Bittner, R.E. (2011) DNA Damage, Somatic Aneuploidy, and Malignant Sarcoma Susceptibility in Muscular Dystrophies H. T. Orr, ed. *PLoS Genetics*. **7**(4), e1002042.

Sgambato, A., Brancaccio, A. (2005) The dystroglycan complex: From biology to cancer. *Journal of Cellular Physiology*. **205**(2), 163–169.

Skrypek, N., Goossens, S., De Smedt, E., Vandamme, N., Berx, G. (2017) Epithelial-to-Mesenchymal Transition: Epigenetic Reprogramming Driving Cellular Plasticity. *Trends in*

genetics : TIG. **33**(12), 943–959.

Suárez-Sánchez, R., Aguilar, A., Wagstaff, K.M., Velez, G., Azuara-Medina, P.M., Gomez, P., Vásquez-Limeta, A., Hernández-Hernández, O., Lieu, K.G., Jans, D.A., Cisneros, B. (2014) Nucleocytoplasmic shuttling of the Duchenne muscular dystrophy gene product dystrophin Dp71d is dependent on the importin α/β and CRM1 nuclear transporters and microtubule motor dynein. *Biochimica et biophysica acta*. **1843**(5), 985–1001.

Svarch, E., Menéndez, A., González, A. (1988) Duchenne muscular dystrophy and acute lymphoblastic leukaemia. *Haematologia*. **21**(2), 123–124.

Tan, Sichuang, Tan, Sipin, Chen, Zhikang, Cheng, K., Chen, Zhicao, Wang, W., Wen, Q., Zhang, W. (2016) Knocking down Dp71 expression in A549 cells reduces its malignancy in vivo and in vitro. *Cancer Investigation*. **34**(1), 16–25.

Tan, Sipin, Tan, J., Tan, Sichuang, Zhao, S., Cao, X., Chen, Z., Weng, Q., Zhang, H., Wang, K.K., Zhou, J., Xiao, X. (2016) Decreased Dp71 expression is associated with gastric adenocarcinoma prognosis. *Oncotarget*. **7**(33), 53702–53711.

Therneau, T.M. (2021) Survival Analysis [R package survival version 3.2-11].

Thienpont, B., Van Dyck, L., Lambrechts, D. (2016) Tumors smother their epigenome. *Molecular & cellular oncology*. **3**(6), e1240549.

Tozawa, T., Itoh, K., Yaoi, T., Tando, S., Umekage, M., Dai, H., Hosoi, H., Fushiki, S. (2012) The shortest isoform of dystrophin (Dp40) interacts with a group of presynaptic proteins to form a presumptive novel complex in the mouse brain. *Molecular neurobiology*. **45**(2), 287–297.

Uotani, H., Hirokawa, S., Saito, F., Tauchi, K., Shimoda, M., Ishizawa, S., Kawaguchi, M., Nomura, K., Kanegane, H., Tsukada, K. (2001) Non-Hodgkin's lymphoma of the ascending colon in a patient with Becker muscular dystrophy: Report of a case. *Surgery Today*. **31**(11), 1016–1019.

Vita, G.L., Politano, L., Berardinelli, A., Vita, G. (2021) Have Duchenne Muscular Dystrophy Patients an Increased Cancer Risk? *Journal of Neuromuscular Diseases*. **Preprint**(Preprint), 1–5.

Wang, Y., Marino-Enriquez, A., Bennett, R.R., Zhu, M., Shen, Y., Eilers, G., Lee, J.-C., Henze, J.,

Fletcher, B.S., Gu, Z., Fox, E.A., Antonescu, C.R., Fletcher, C.D.M., Guo, X., Raut, C.P., Demetri, G.D., van de Rijn, M., Ordog, T., Kunkel, L.M., Fletcher, J.A. (2014) Dystrophin is a tumor suppressor in human cancers with myogenic programs. *Nature genetics*. **46**(6), 601–6.

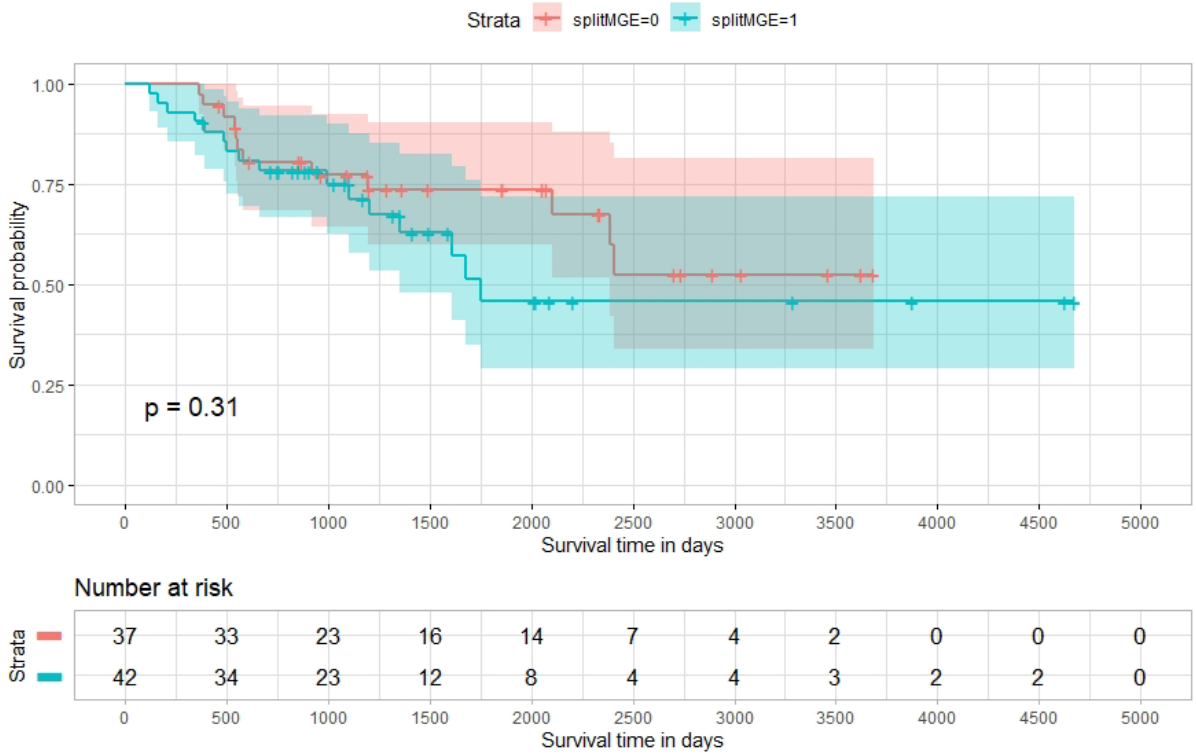
Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Sander, C., Stuart, J.M., Chang, K., Creighton, C.J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, Y.S.N., Chu, A., Chuah, E., Chun, H.J.E., Dhalla, N., Guin, R., Hirst, M., Hirst, C., Holt, R.A., Jones, S.J.M., Lee, D., Li, H.I., Marra, M.A., Mayo, M., Moore, R.A., Mungall, A.J., Robertson, A.G., Schein, J.E., Sipahimalani, P., Tam, A., Thiessen, N., Varhol, R.J., Beroukhim, R., Bhatt, A.S., Brooks, A.N., Cherniack, A.D., Freeman, S.S., Gabriel, S.B., Helman, E., Jung, J., Meyerson, M., Ojesina, A.I., Pedamallu, C.S., Saksena, G., Schumacher, S.E., Tabak, B., Zack, T., Lander, E.S., Bristow, C.A., Hadjipanayis, A., Haseley, P., Kucherlapati, R., Lee, S., Lee, E., Luquette, L.J., Mahadeshwar, H.S., Pantazi, A., Parfenov, M., Park, P.J., Protopopov, A., Ren, X., Santoso, N., Seidman, J., Seth, S., Song, X., Tang, J., Xi, R., Xu, A.W., Yang, Lixing, Zeng, D., Auman, J.T., Balu, S., Buda, E., Fan, C., Hoadley, K.A., Jones, C.D., Meng, S., Mieczkowski, P.A., Parker, J.S., Perou, C.M., Roach, J., Shi, Y., Silva, G.O., Tan, D., Veluvolu, U., Waring, S., Wilkerson, M.D., Wu, J., Zhao, W., Bodenheimer, T., Hayes, D.N., Hoyle, A.P., Jeffreys, S.R., Mose, L.E., Simons, J. V., Soloway, M.G., Baylin, S.B., Berman, B.P., Bootwalla, M.S., Danilova, L., Herman, J.G., Hinoue, T., Laird, P.W., Rhie, S.K., Shen, H., Triche, T., Weisenberger, D.J., Carter, S.L., Cibulskis, K., Chin, L., Zhang, Jianhua, Sougnez, C., Wang, M., Getz, G., Dinh, H., Doddapaneni, H.V., Gibbs, R., Gunaratne, P., Han, Y., Kalra, D., Kovar, C., Lewis, L., Morgan, M., Morton, D., Muzny, D., Reid, J., Xi, L., Cho, J., Dicara, D., Frazer, S., Gehlenborg, N., Heiman, D.I., Kim, J., Lawrence, M.S., Lin, P., Liu, Yingchun, Noble, M.S., Stojanov, P., Voet, D., Zhang, H., Zou, L., Stewart, C., Bernard, B., Bressler, R., Eakin, A., Iype, L., Knijnenburg, T., Kramer, R., Kreisberg, R., Leinonen, K., Lin, J., Liu, Yuexin, Miller, M., Reynolds, S.M., Rovira, H., Shmulevich, I., Thorsson, V., Yang, D., Zhang, W., Amin, S., Wu, C.J., Wu, C.C., Akbani, R., Aldape, K., Baggerly, K.A., Broom, B., Casasent, T.D., Cleland, J., Dodda, D., Edgerton, M., Han, L., Herbrich, S.M., Ju, Z., Kim, H., Lerner, S., Li, J., Liang, H., Liu, W., Lorenzi, P.L., Lu, Y., Melott, J., Nguyen, L., Su, X., Verhaak, R., Wang, W., Wong, A., Yang, Y., Yao, J., Yao, R., Yoshihara, K., Yuan, Y., Yung, A.K., Zhang, N., Zheng, S., Ryan, M., Kane, D.W., Aksoy, B.A., Ciriello, G.,

Dresdner, G., Gao, J., Gross, B., Jacobsen, A., Kahles, A., Ladanyi, M., Lee, W., Lehmann, K. Van, Miller, M.L., Ramirez, R., Rätsch, G., Reva, B., Schultz, N., Senbabaoglu, Y., Shen, R., Sinha, R., Sumer, S.O., Sun, Y., Taylor, B.S., Weinhold, N., Fei, S., Spellman, P., Benz, C., Carlin, D., Cline, M., Craft, B., Goldman, M., Haussler, D., Ma, S., Ng, S., Paull, E., Radenbaugh, A., Salama, S., Sokolov, A., Swatloski, T., Uzunangelov, V., Waltman, P., Yau, C., Zhu, J., Hamilton, S.R., Abbott, S., Abbott, R., Dees, N.D., Delehaunty, K., Ding, L., Dooling, D.J., Eldred, J.M., Fronick, C.C., Fulton, R., Fulton, L.L., Kalicki-Veizer, J., Kanchi, K.L., Kandoth, C., Koboldt, D.C., Larson, D.E., Ley, T.J., Lin, L., Lu, C., Magrini, V.J., Mardis, E.R., McLellan, M.D., McMichael, J.F., Miller, C.A., O'Laughlin, M., Pohl, C., Schmidt, H., Smith, S.M., Walker, J., Wallis, J.W., Wendl, M.C., Wilson, R.K., Wylie, T., Zhang, Q., Burton, R., Jensen, M.A., Kahn, A., Pihl, T., Pot, D., Wan, Y., Levine, D.A., Black, A.D., Bowen, J., Frick, J., Gastier-Foster, J.M., Harper, H.A., Helsel, C., Leraas, K.M., Lichtenberg, T.M., McAllister, C., Ramirez, N.C., Sharpe, S., Wise, L., Zmuda, E., Chanock, S.J., Davidsen, T., Demchok, J.A., Eley, G., Felau, I., Sheth, M., Sofia, H., Staudt, L., Tarnuzzer, R., Wang, Z., Yang, Liming, Zhang, Jiashan, Omberg, L., Margolin, A., Raphael, B.J., Vandin, F., Wu, H.T., Leiserson, M.D.M., Benz, S.C., Vaske, C.J., Noushmehr, H., Wolf, D., Veer, L.V.T., Anastassiou, D., Yang, T.H.O., Lopez-Bigas, N., Gonzalez-Perez, A., Tamborero, D., Xia, Z., Li, W., Cho, D.Y., Przytycka, T., Hamilton, M., McGuire, S., Nelander, S., Johansson, P., Jörnsten, R., Kling, T. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics 2013 45:10*. **45**(10), 1113–1120.

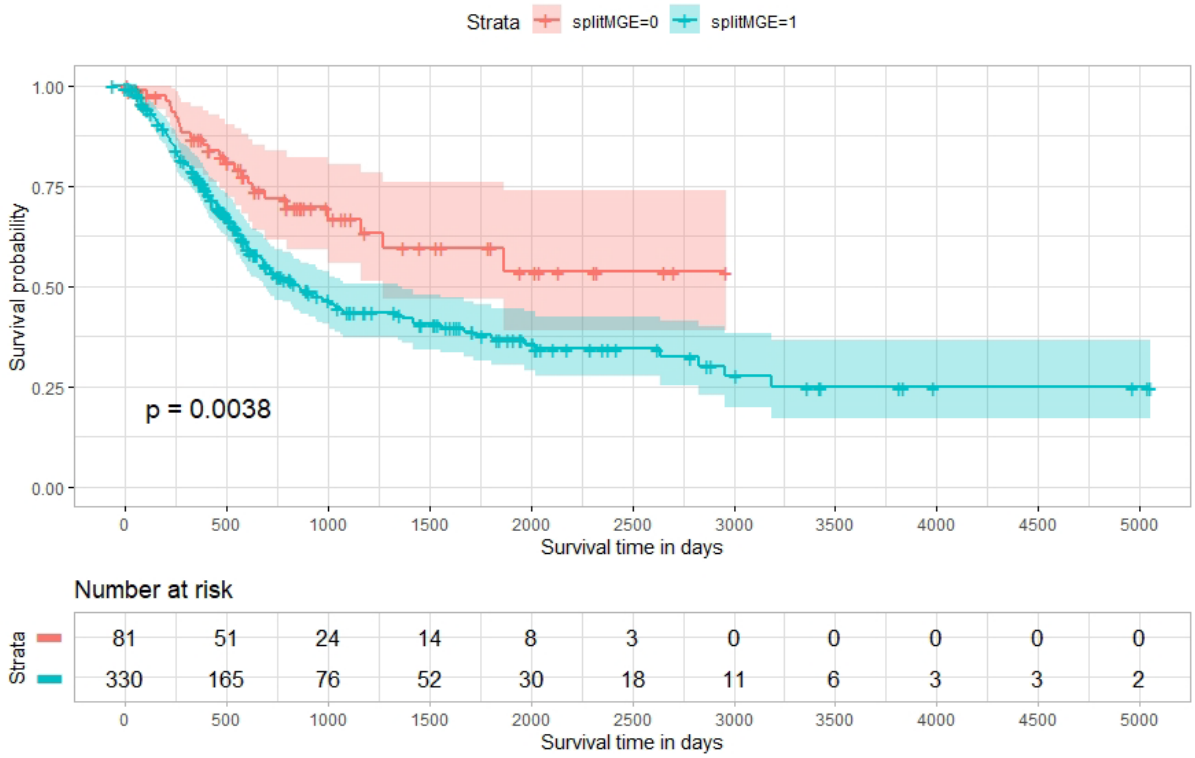
Williams, B.A., Mandrekar, J.N., Mandrekar, S.J., Cha, S.S., Furth, A.F., Williams, B., Jayawant Mandrekar, M.N., Alfred Furth, M.F. (2006) Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes. *Mayo Foundation*. [online]. Available from: https://www.mayo.edu/research/documents/biostat-79pdf/doc-10027230 [Accessed August 16, 2022].

Zuo, X.-Y., Feng, Q.-S., Sun, J., Wei, P.-P., Chin, Y.-M., Guo, Y.-M., Xia, Y.-F., Li, B., Xia, X.-J., Jia, W.-H., Liu, J.-J., Khoo, A.S.-B., Mushiroda, T., Ng, C.-C., Su, W.-H., Zeng, Y.-X., Bei, J.-X. (2019) X-chromosome association study reveals genetic susceptibility loci of nasopharyngeal carcinoma. *Biology of sex differences*. **10**(1), 13.

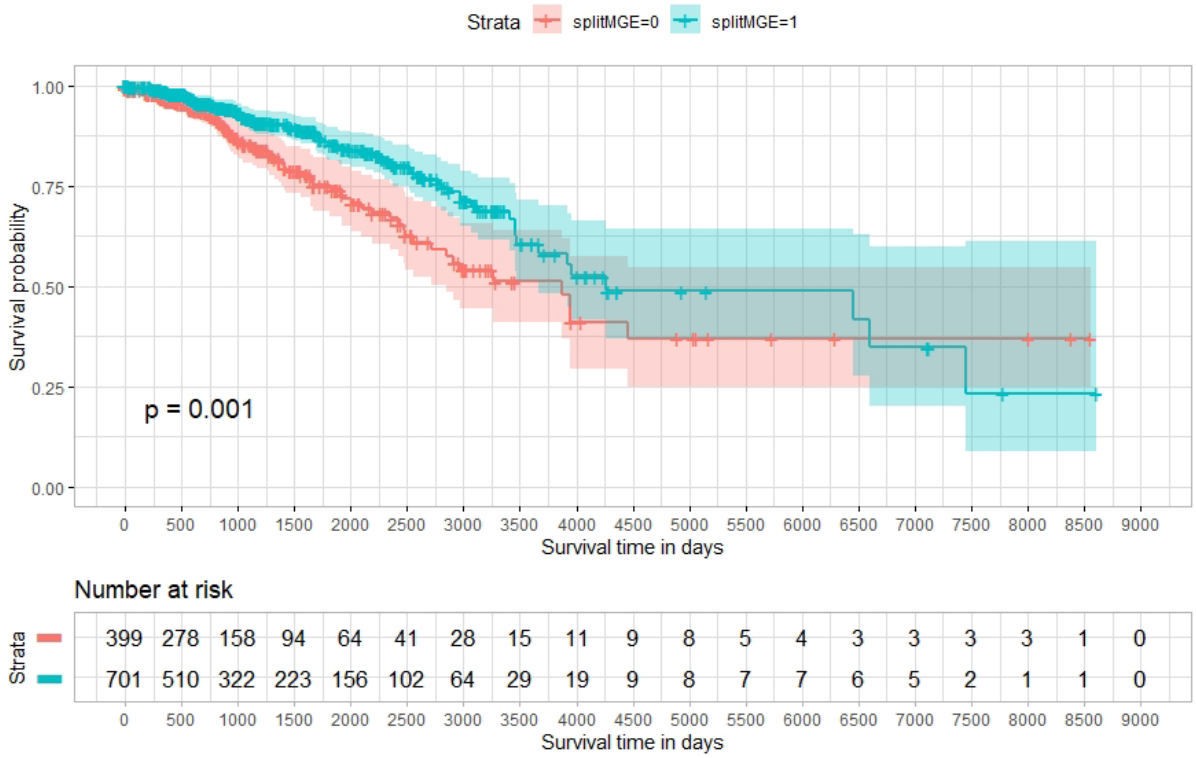
6. Appendices



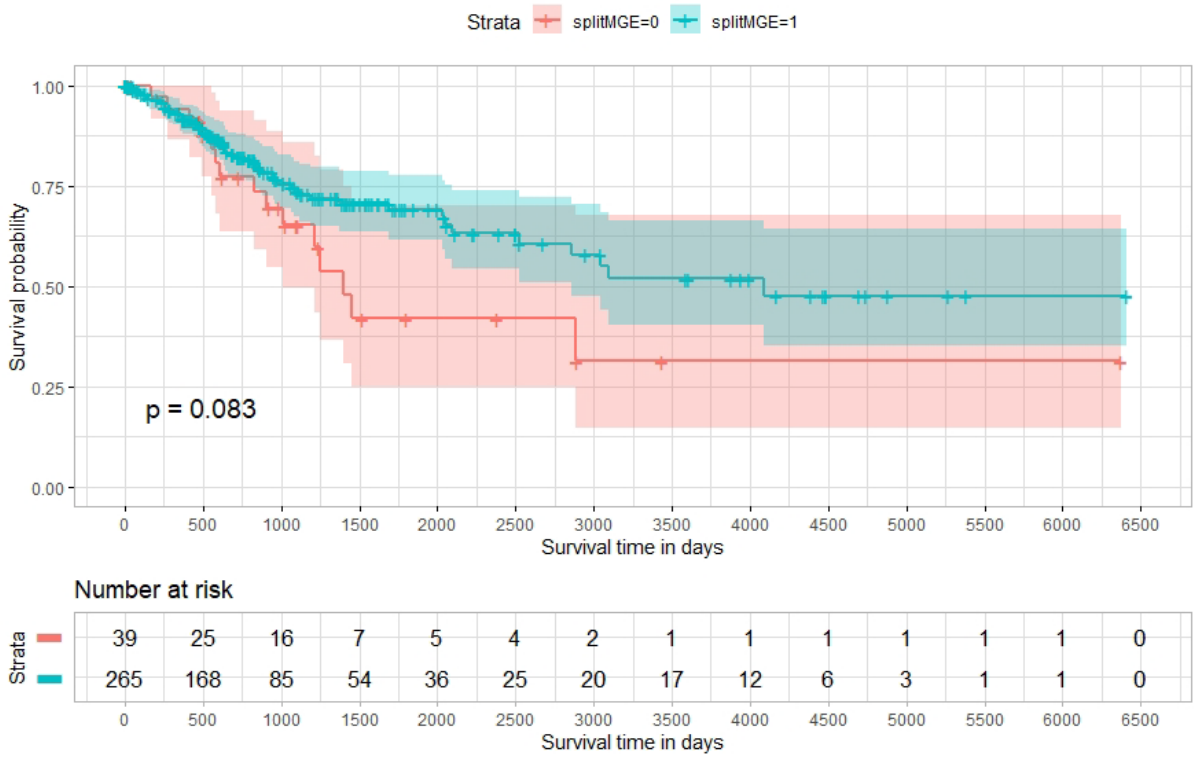
Supplementary figure 1. Kaplan-Meier survival curves for high vs. low *DMD* expression in ACC. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



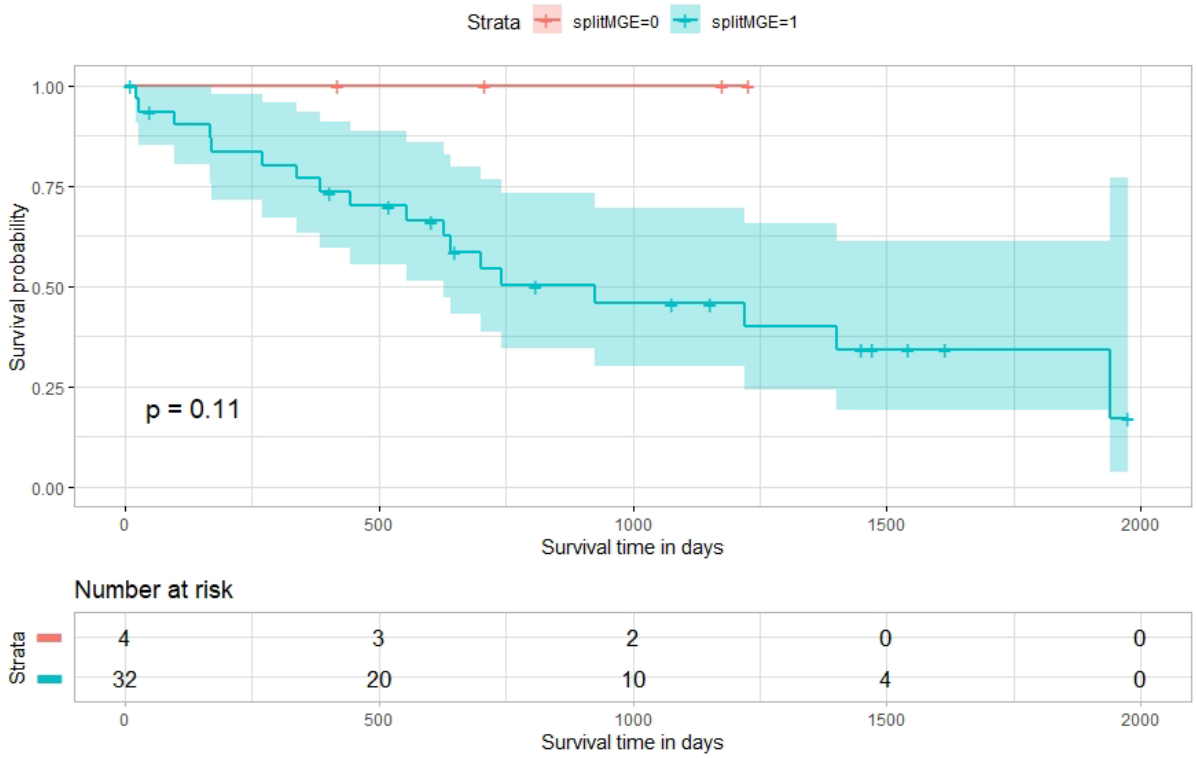
Supplementary figure 2. Kaplan-Meier survival curves for high vs. low *DMD* expression in BLCA. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



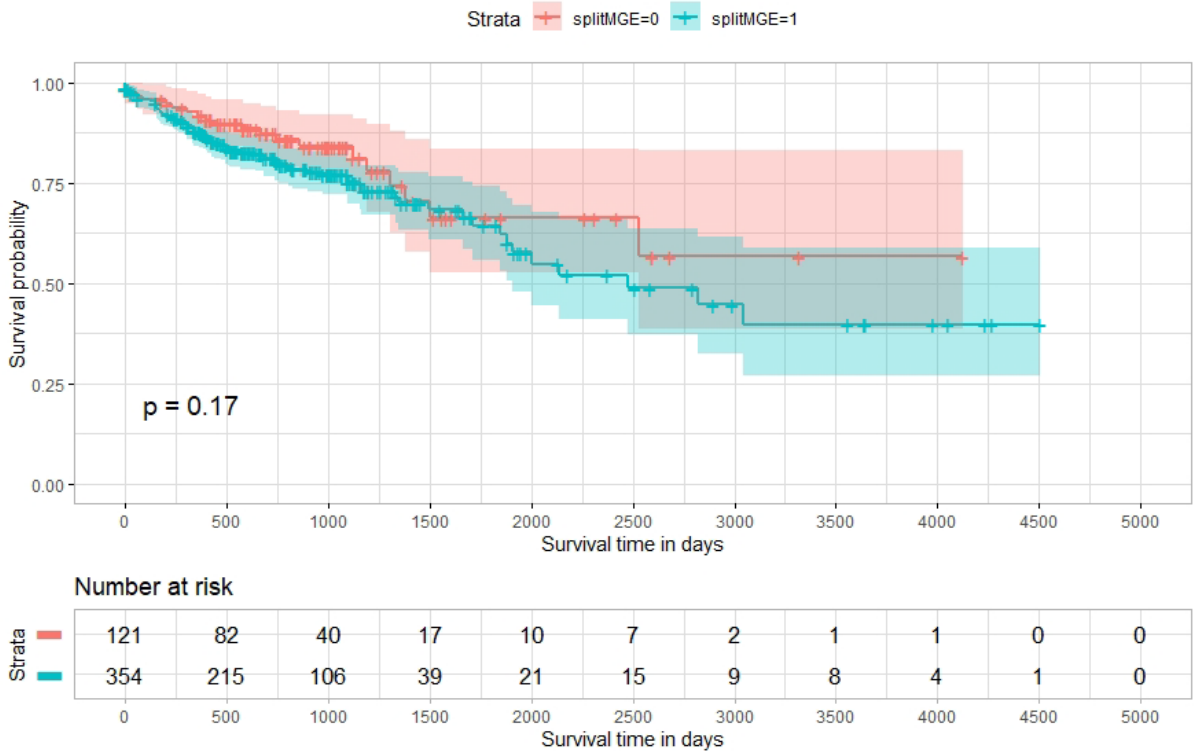
Supplementary figure 3. Kaplan-Meier survival curves for high vs. low *DMD* expression in BRCA. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



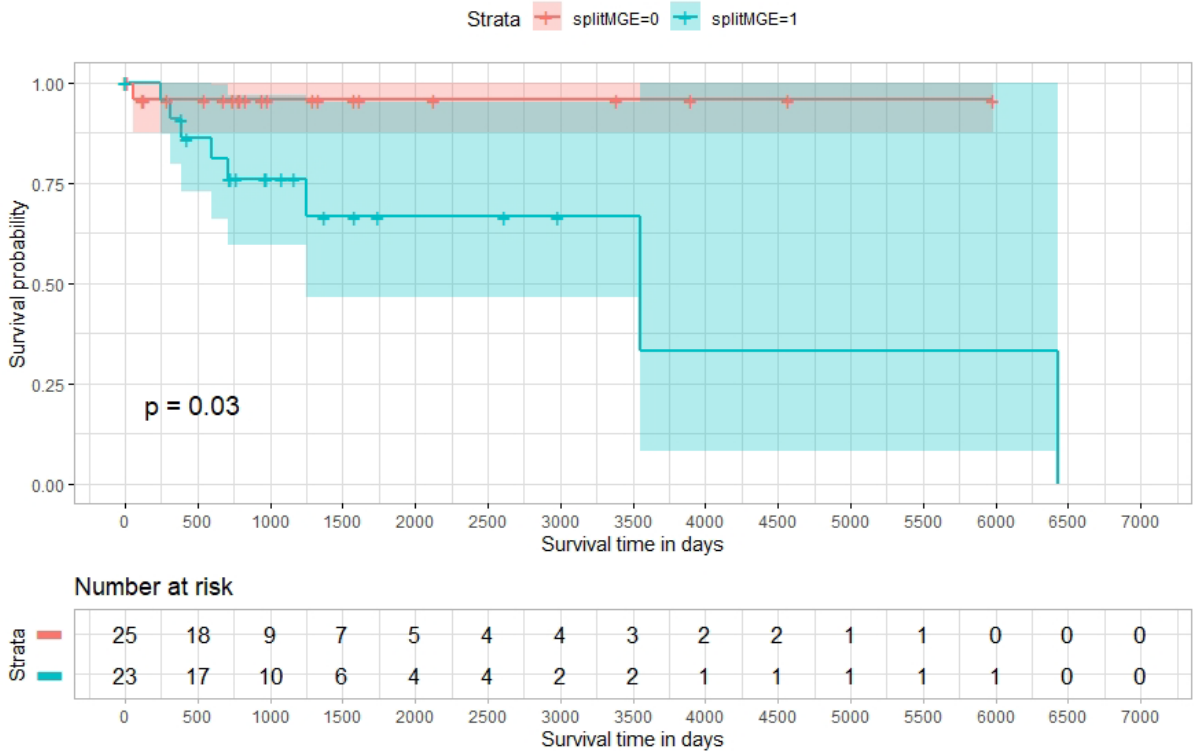
Supplementary figure 4. Kaplan-Meier survival curves for high vs. low *DMD* expression in CESC. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



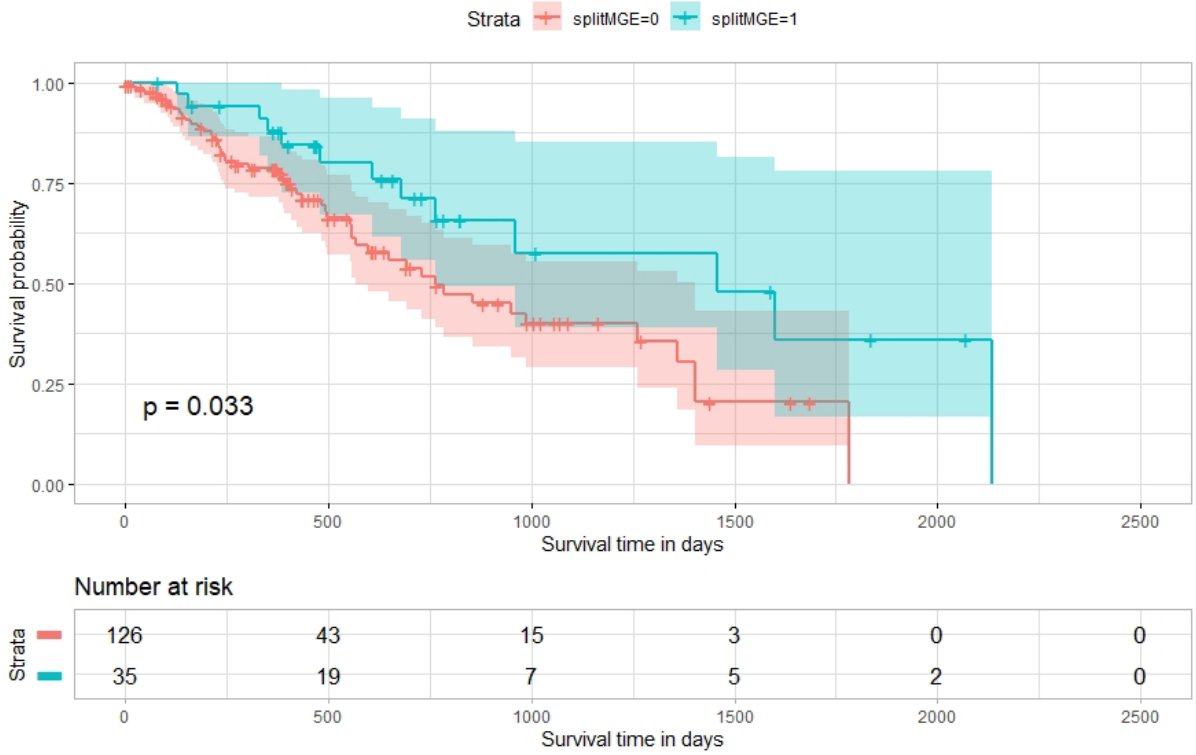
Supplementary figure 5. Kaplan-Meier survival curves for high vs. low *DMD* expression in CHOL. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



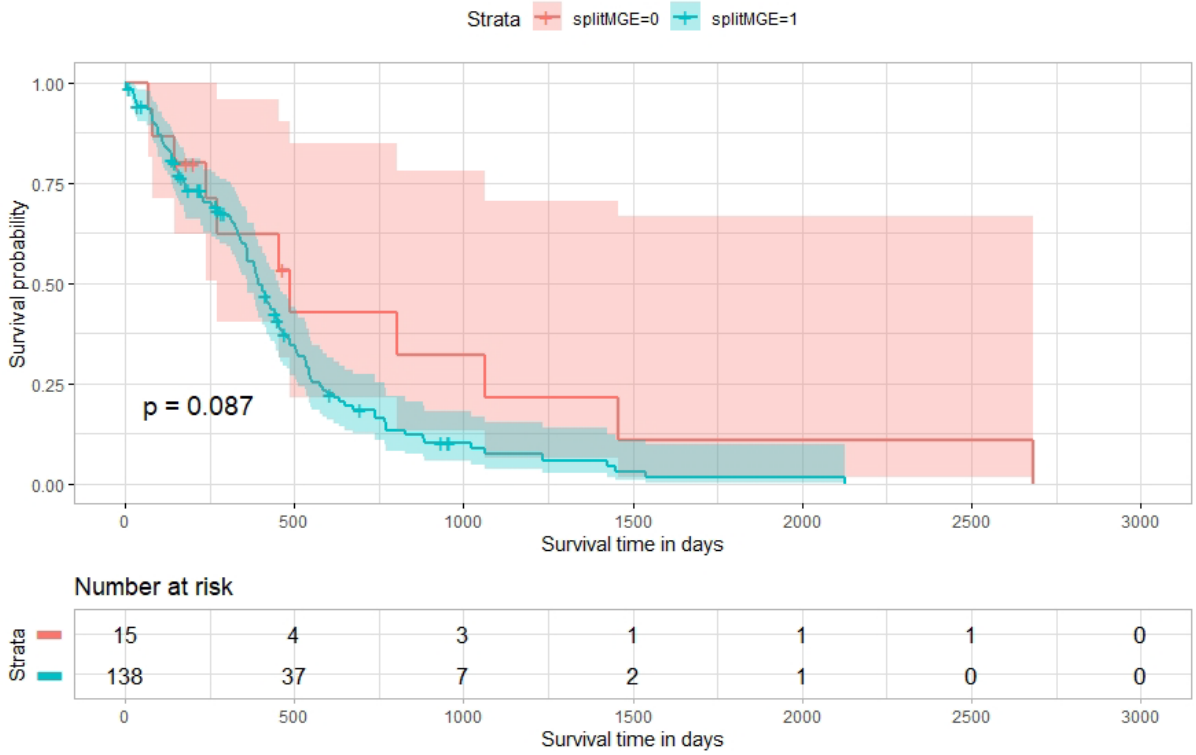
Supplementary figure 6. Kaplan-Meier survival curves for high vs. low *DMD* expression in COAD. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



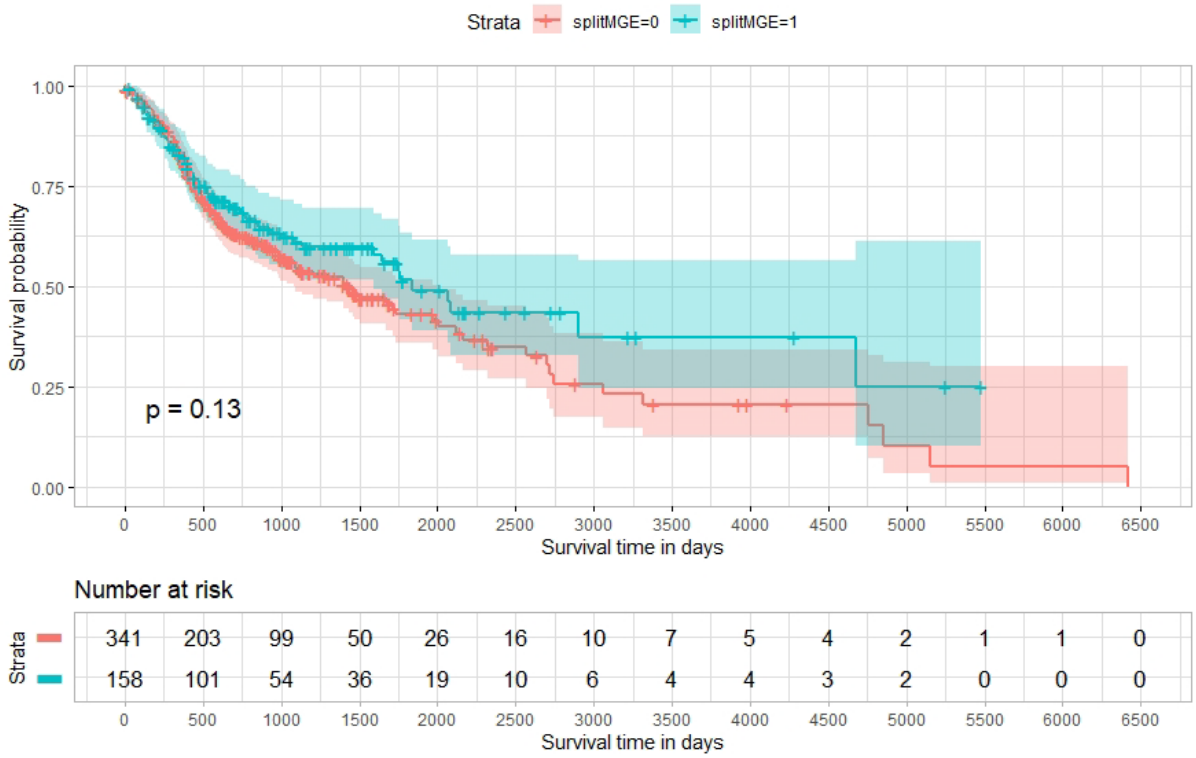
Supplementary figure 7. Kaplan-Meier survival curves for high vs. low *DMD* expression in DLBC. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



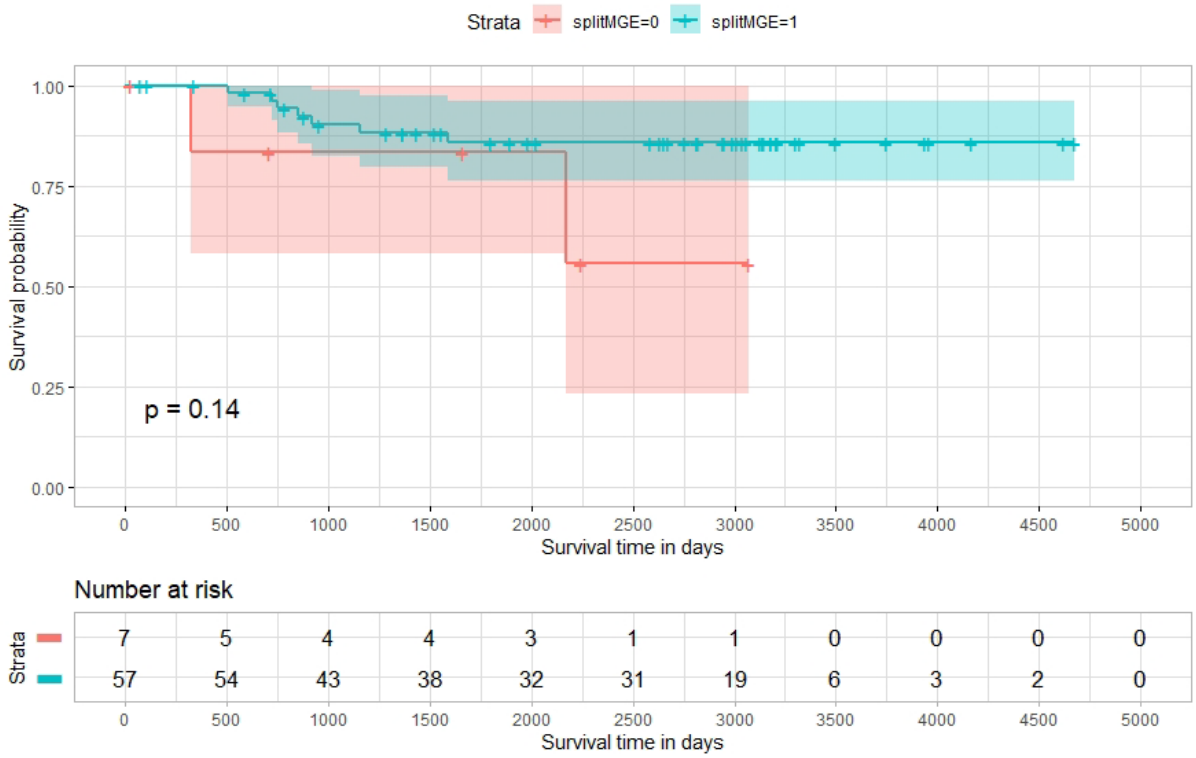
Supplementary figure 8. Kaplan-Meier survival curves for high vs. low *DMD* expression in ESCA. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



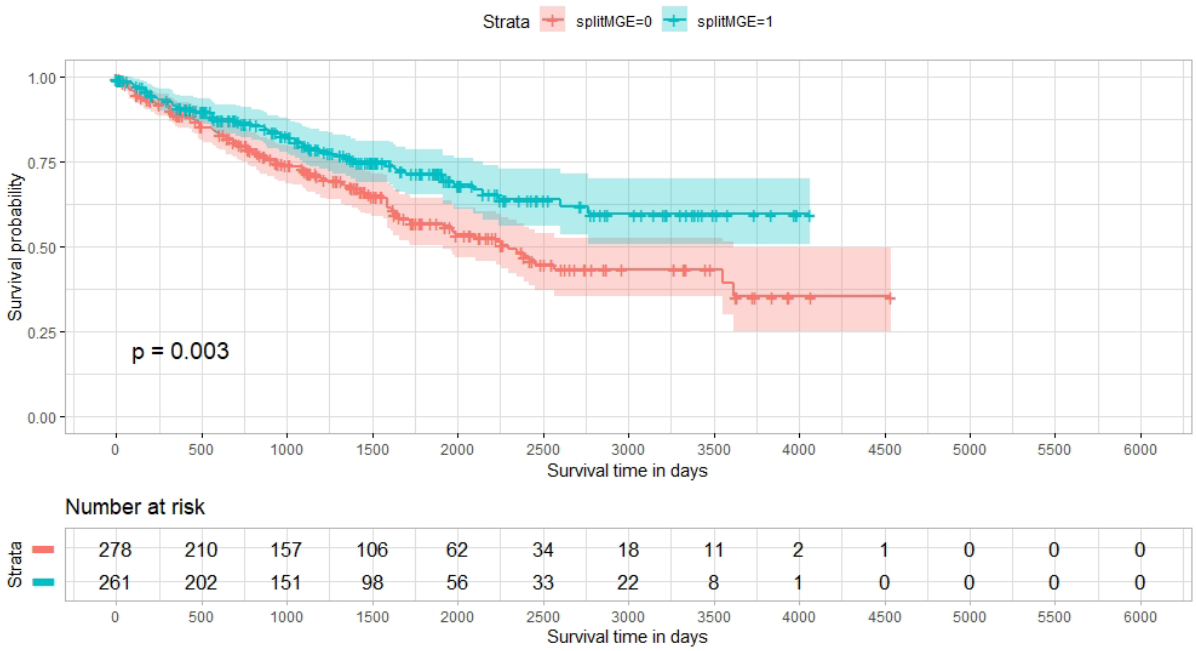
Supplementary figure 9. Kaplan-Meier survival curves for high vs. low *DMD* expression in GBM. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



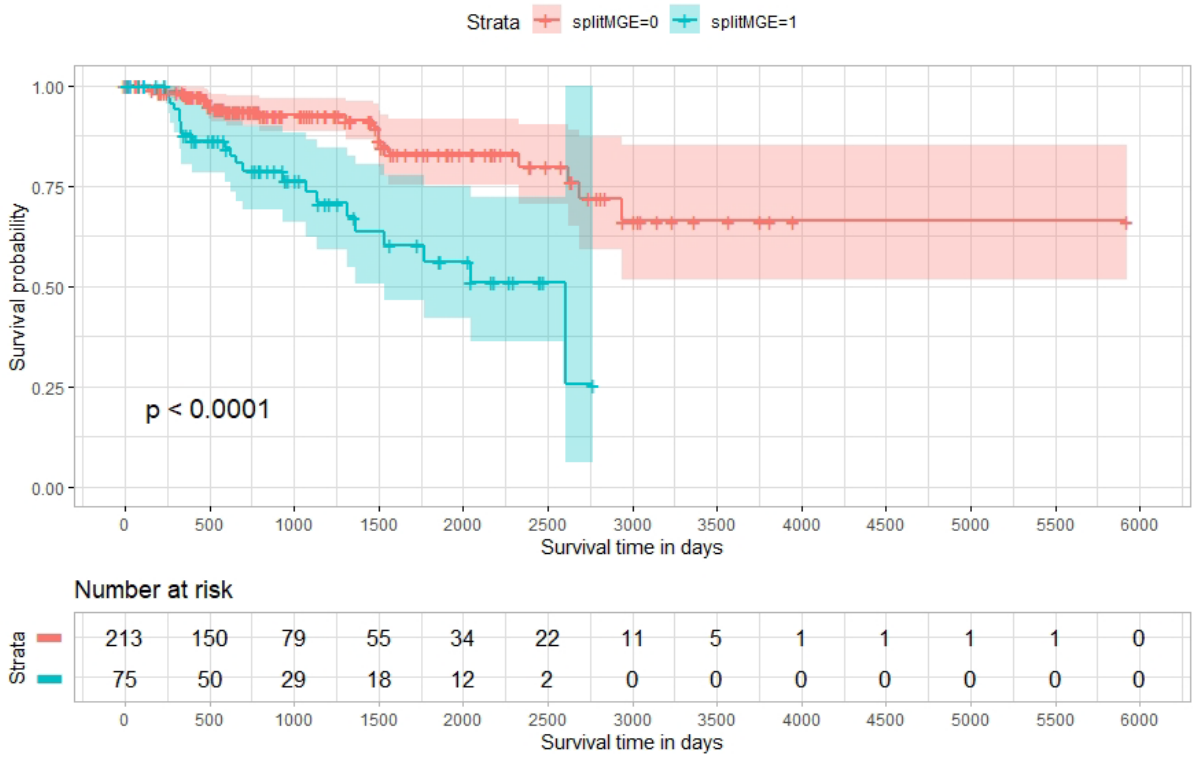
Supplementary figure 10. Kaplan-Meier survival curves for high vs. low *DMD* expression in HNSC. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



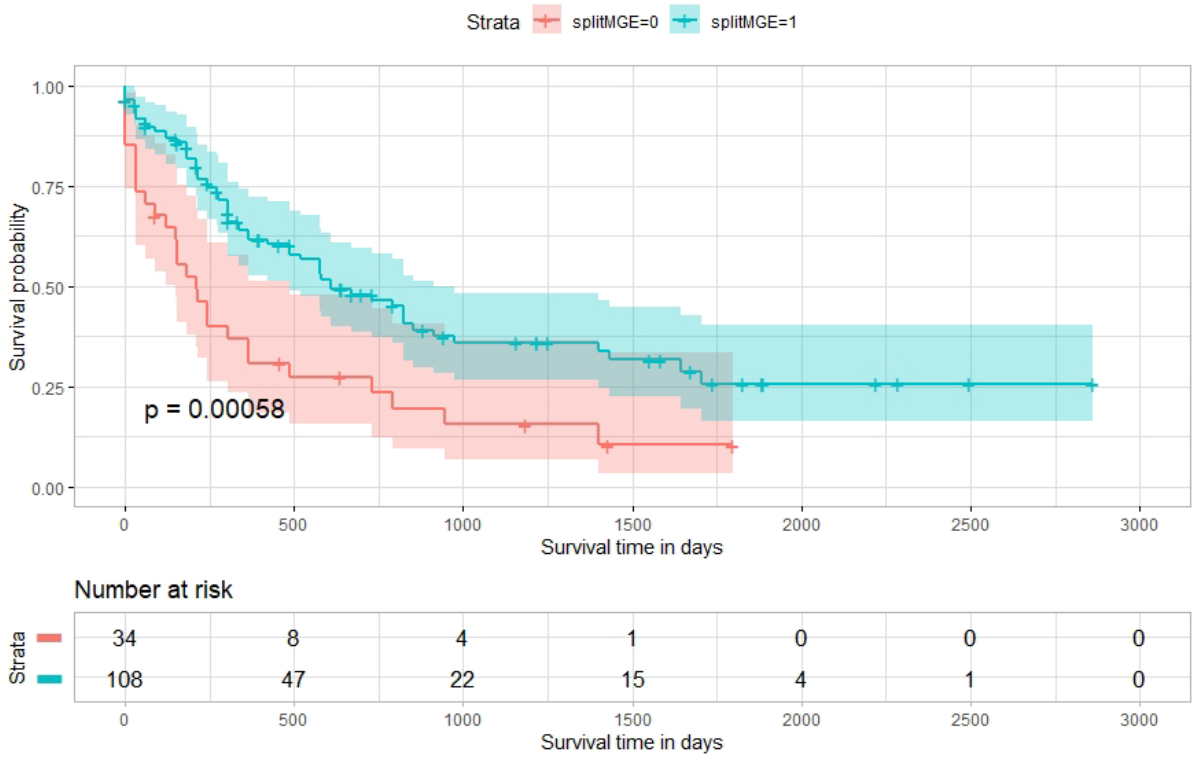
Supplementary figure 11. Kaplan-Meier survival curves for high vs. low *DMD* expression in KICH. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



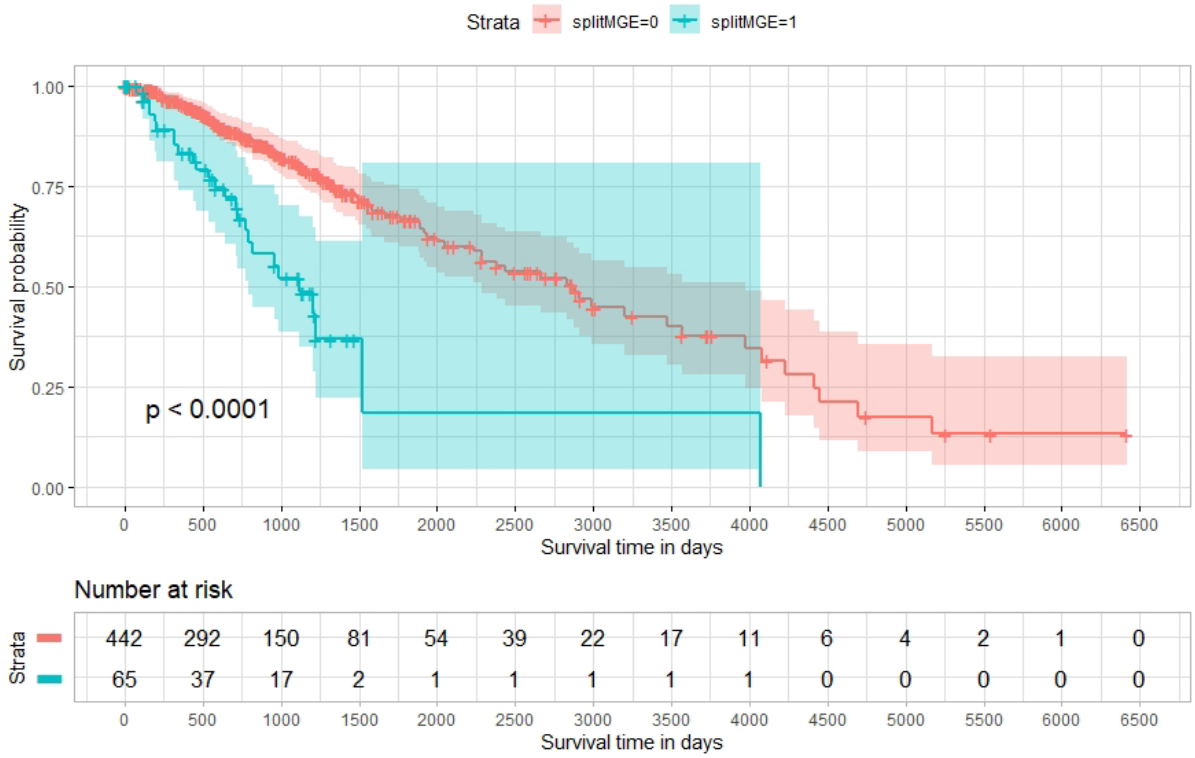
Supplementary figure 12. Kaplan-Meier survival curves for high vs. low *DMD* expression in KIRC. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



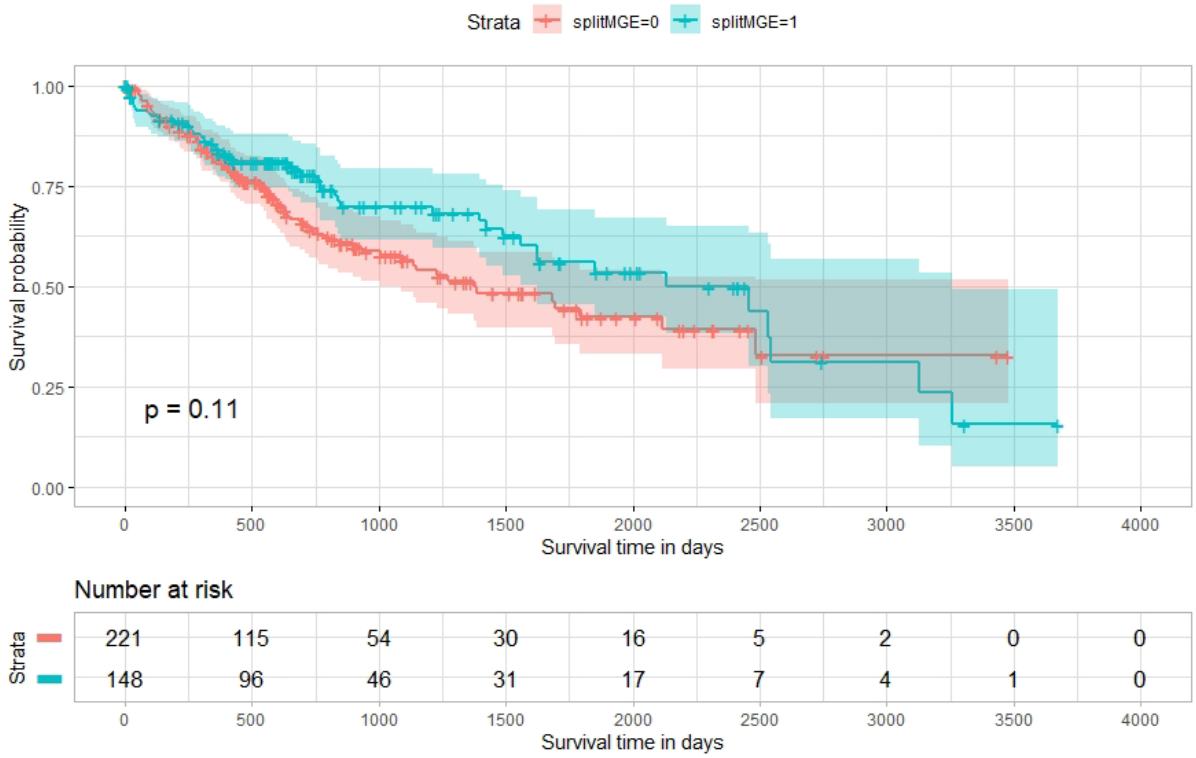
Supplementary figure 13. Kaplan-Meier survival curves for high vs. low *DMD* expression in KIRP. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



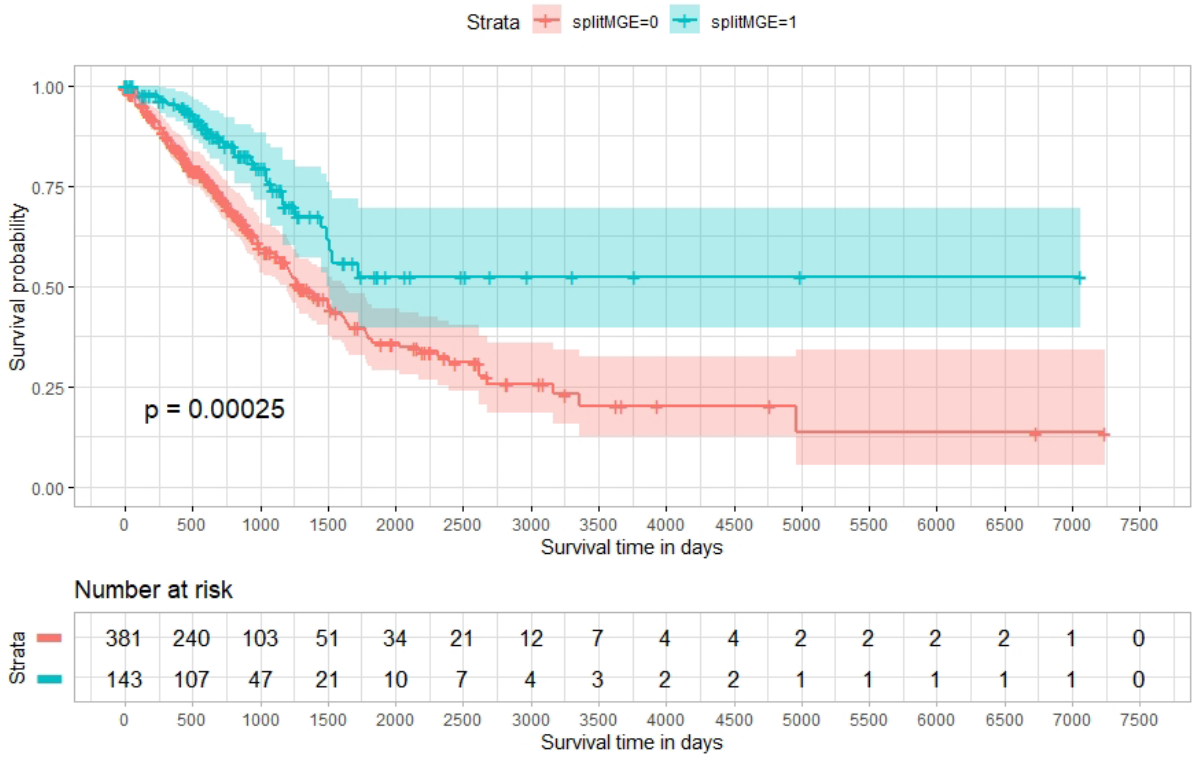
Supplementary figure 14. Kaplan-Meier survival curves for high vs. low *DMD* expression in LAML. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



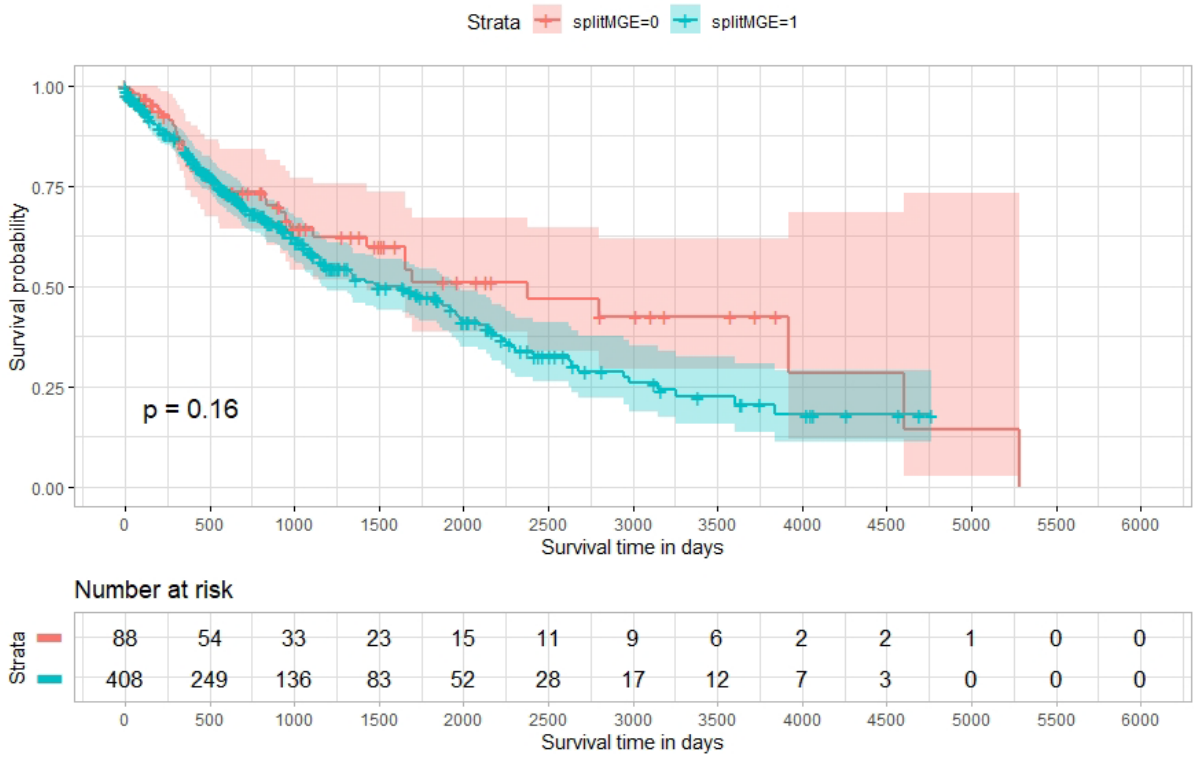
Supplementary figure 15. Kaplan-Meier survival curves for high vs. low *DMD* expression in LGG. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



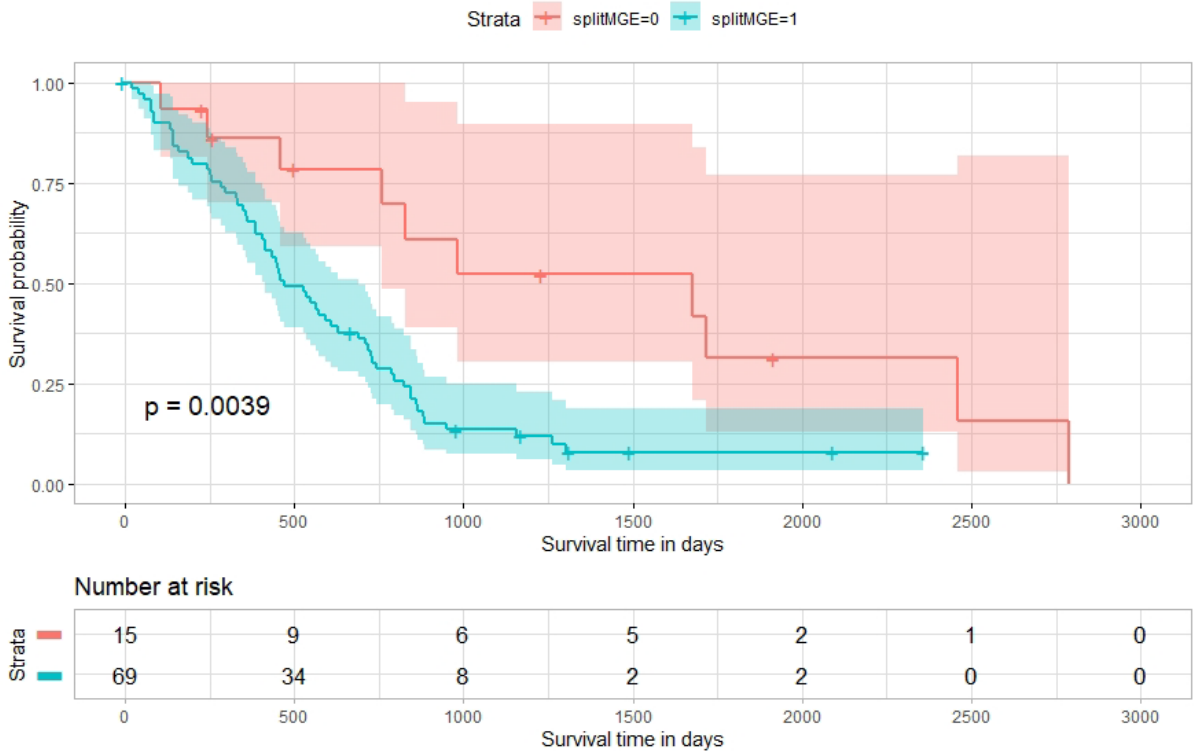
Supplementary figure 16. Kaplan-Meier survival curves for high vs. low *DMD* expression in LIHC. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



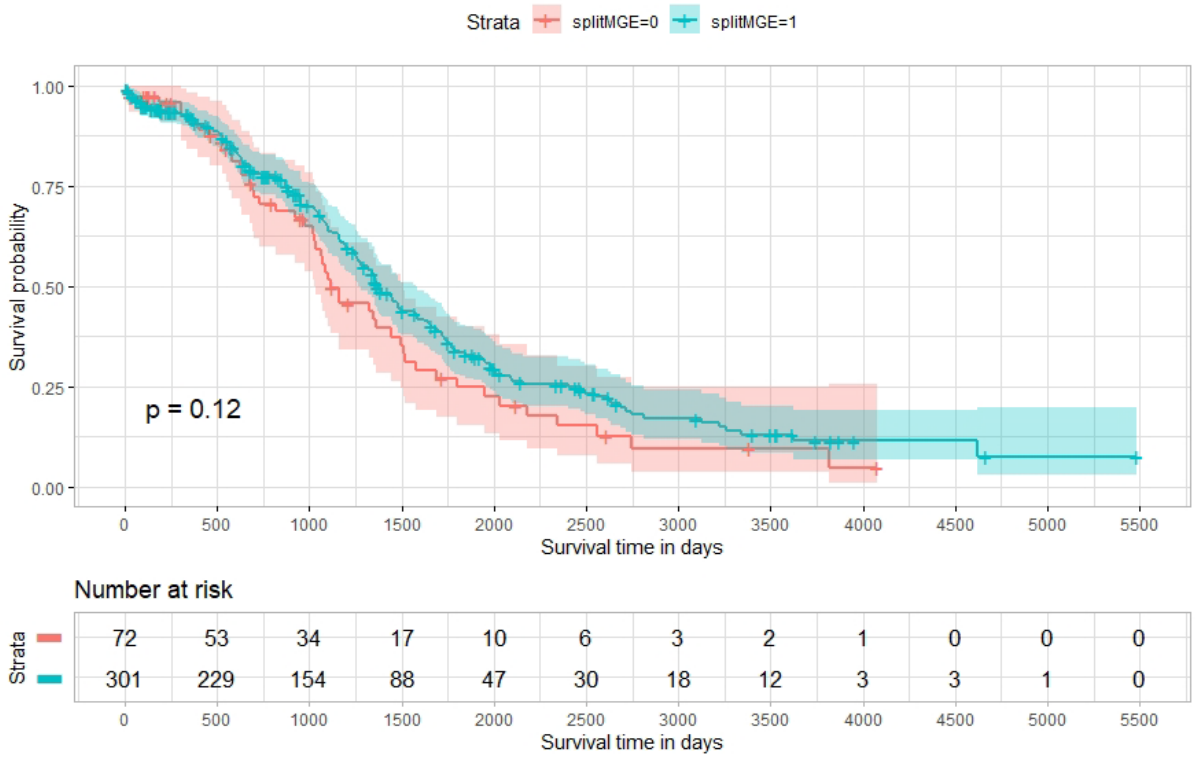
Supplementary figure 17. Kaplan-Meier survival curves for high vs. low *DMD* expression in LUAD. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



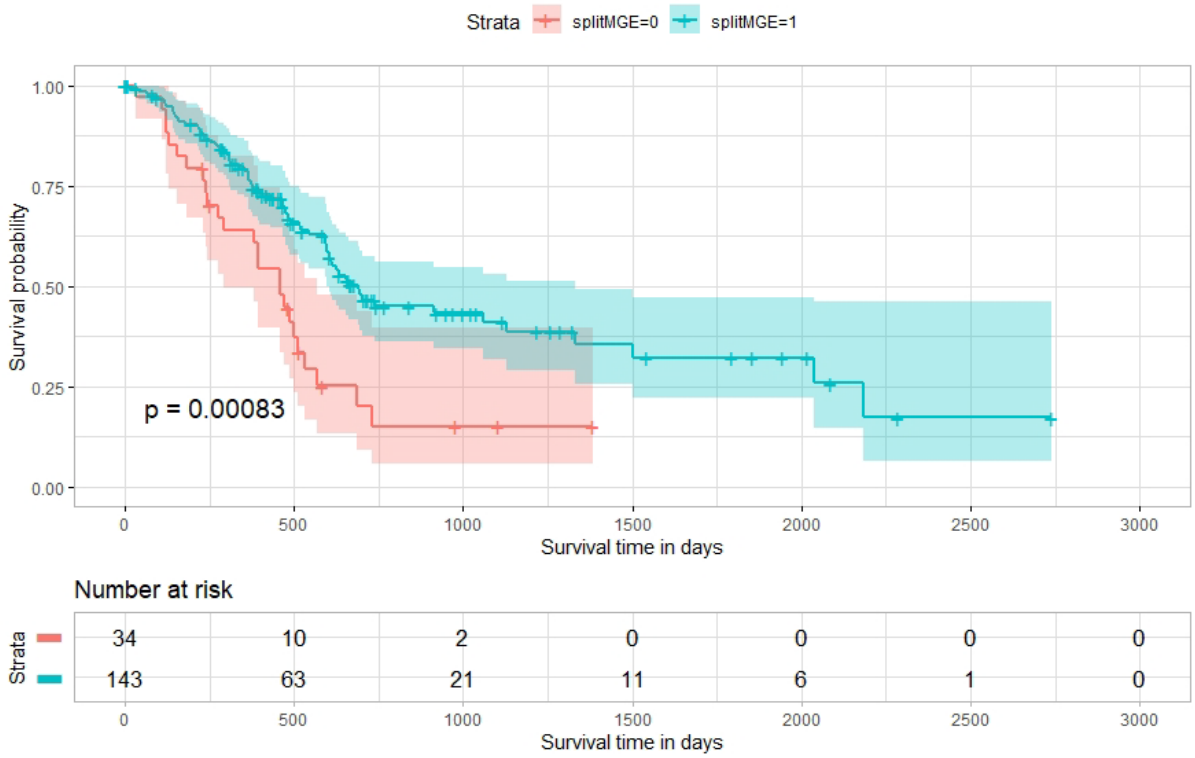
Supplementary figure 18. Kaplan-Meier survival curves for high vs. low *DMD* expression in LUSC. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



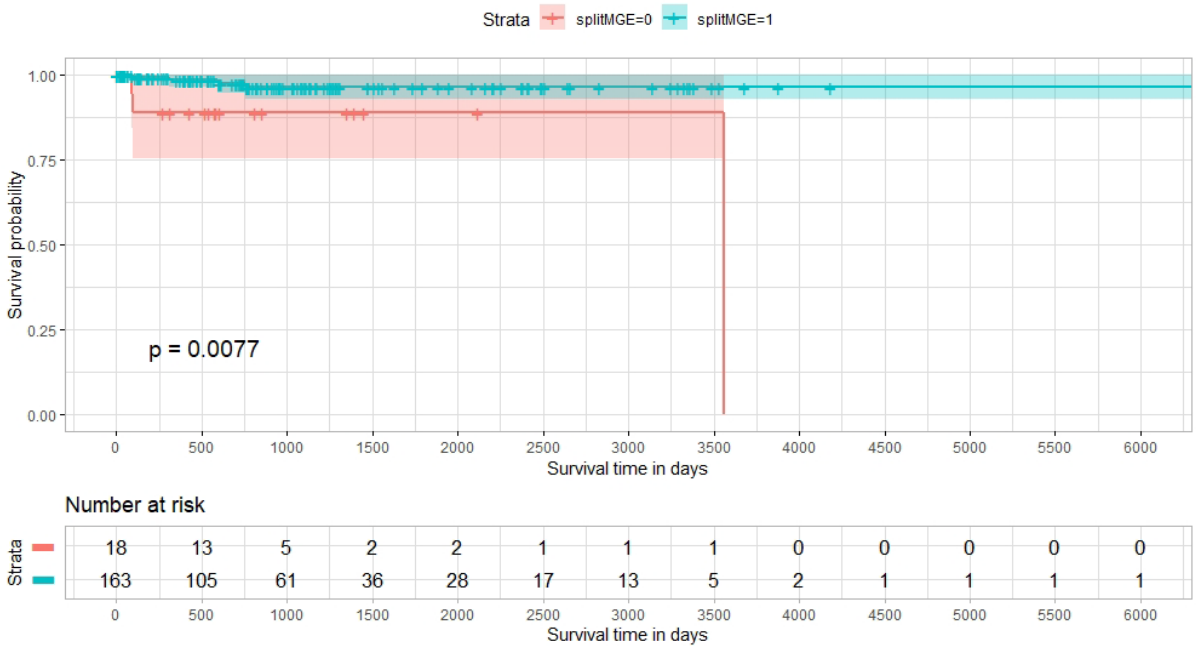
Supplementary figure 19. Kaplan-Meier survival curves for high vs. low *DMD* expression in MESO. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



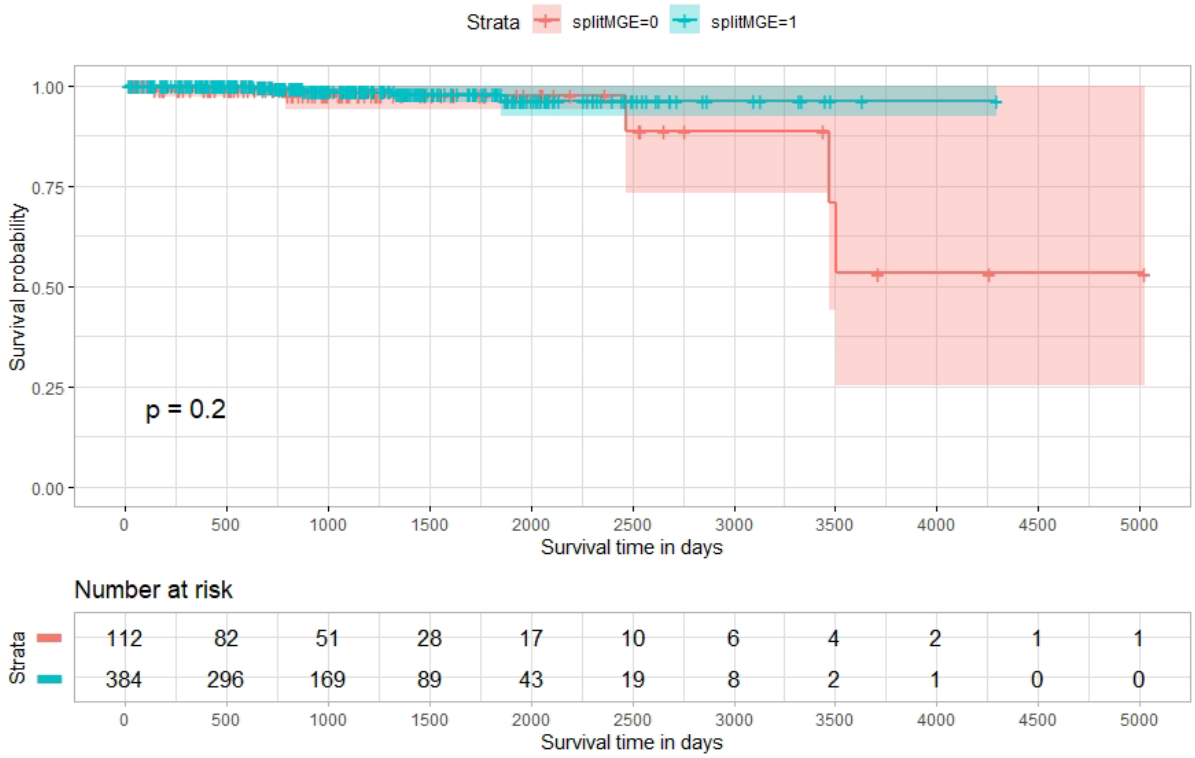
Supplementary figure 20. Kaplan-Meier survival curves for high vs. low *DMD* expression in OV. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



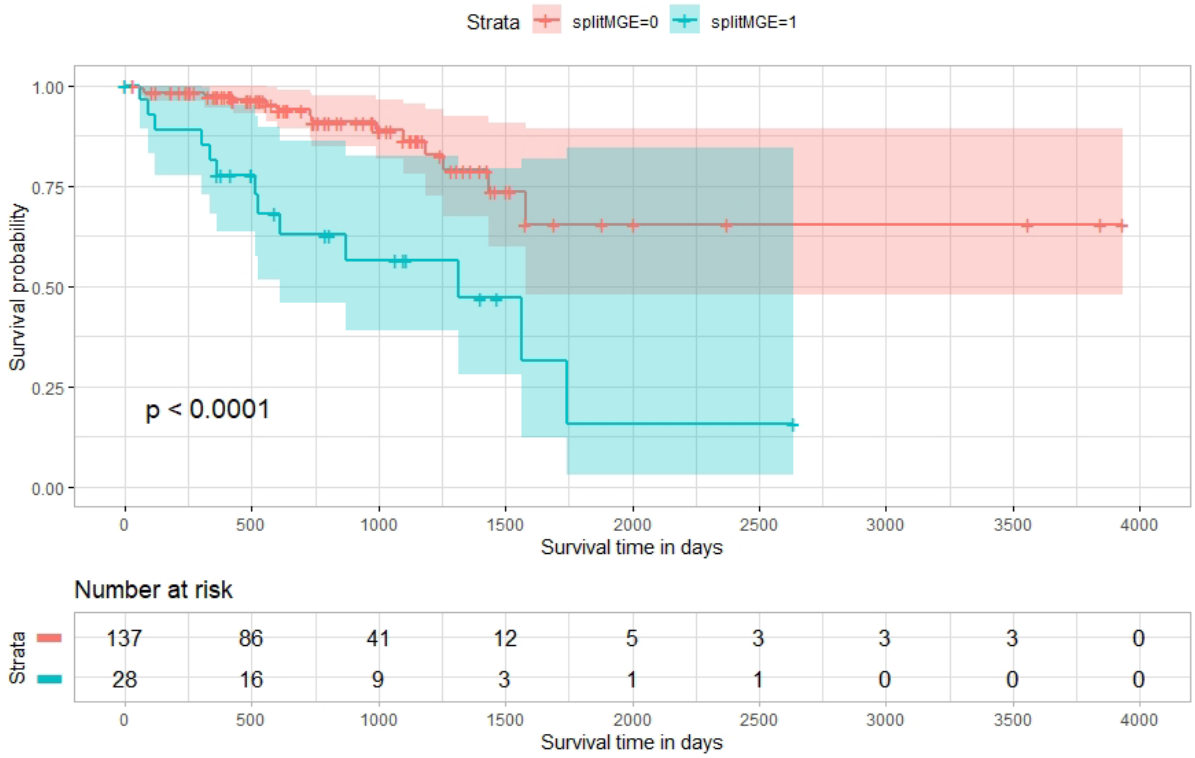
Supplementary figure 21. Kaplan-Meier survival curves for high vs. low *DMD* expression in PAAD. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



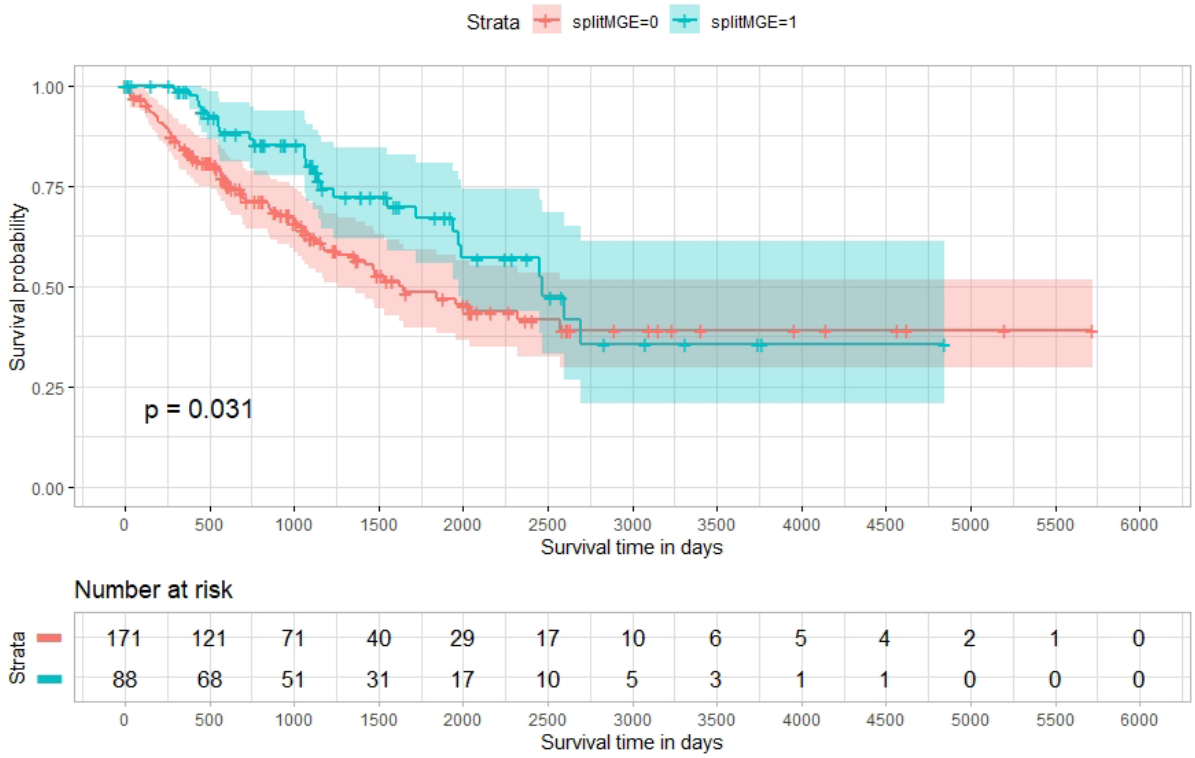
Supplementary figure 22. Kaplan-Meier survival curves for high vs. low *DMD* expression in PCPG. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



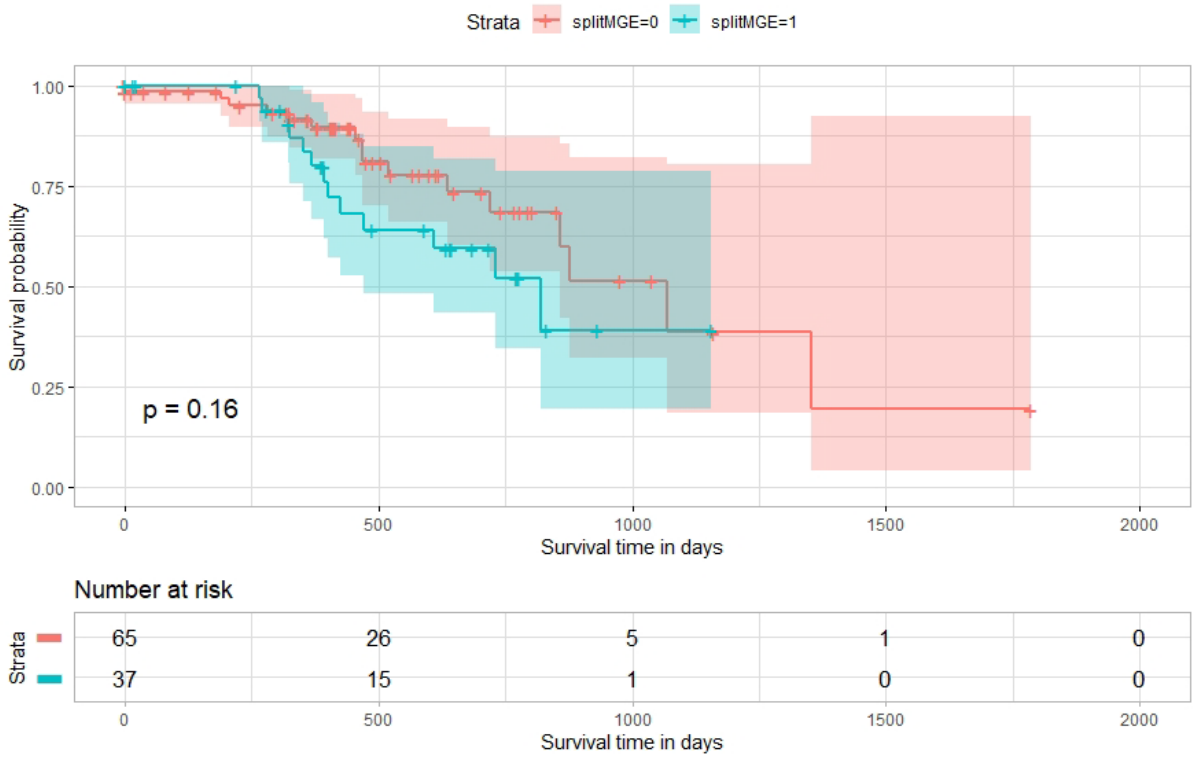
Supplementary figure 23. Kaplan-Meier survival curves for high vs. low *DMD* expression in PRAD. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



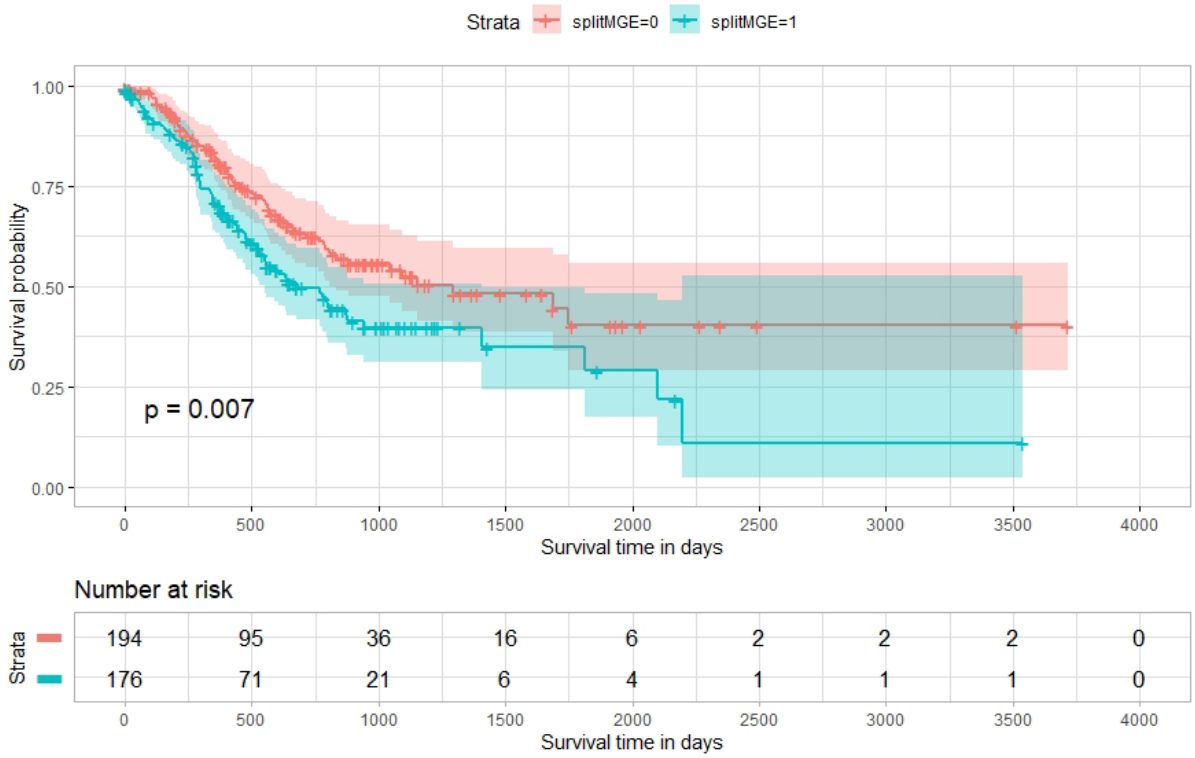
Supplementary figure 24. Kaplan-Meier survival curves for high vs. low *DMD* expression in READ. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



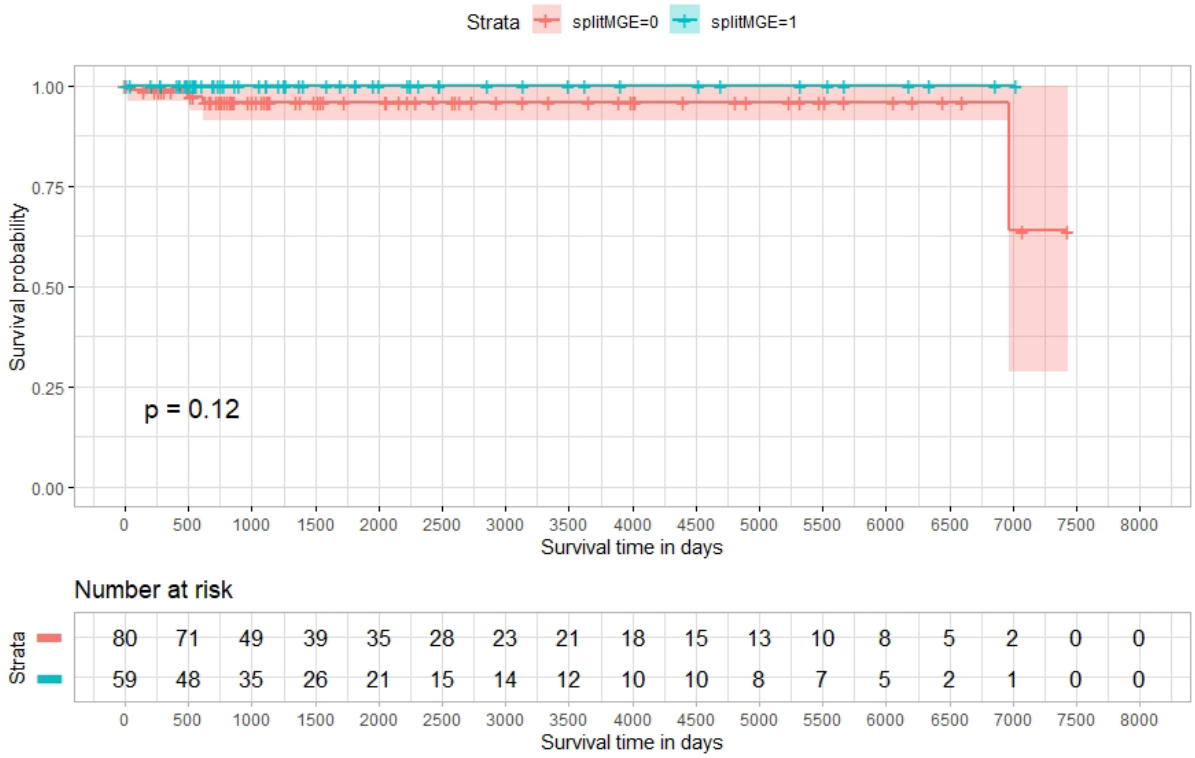
Supplementary figure 25. Kaplan-Meier survival curves for high vs. low *DMD* expression in SARC. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



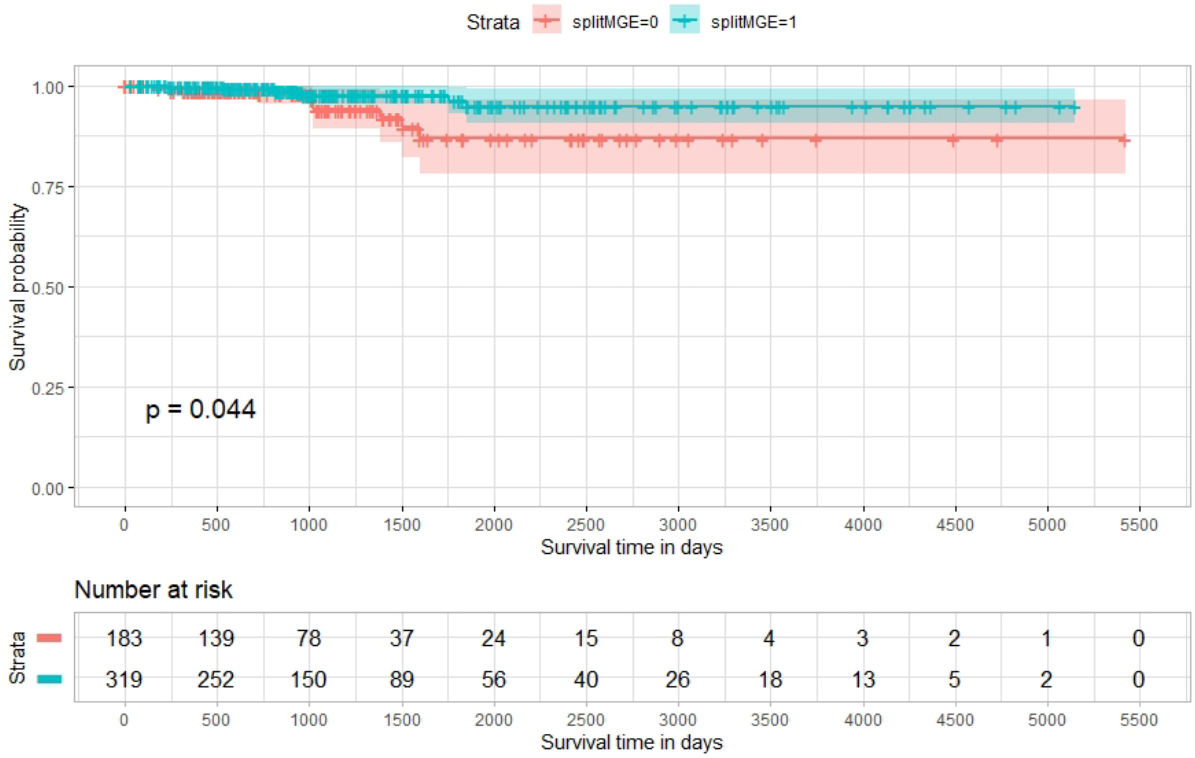
Supplementary figure 26. Kaplan-Meier survival curves for high vs. low *DMD* expression in SKCM. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



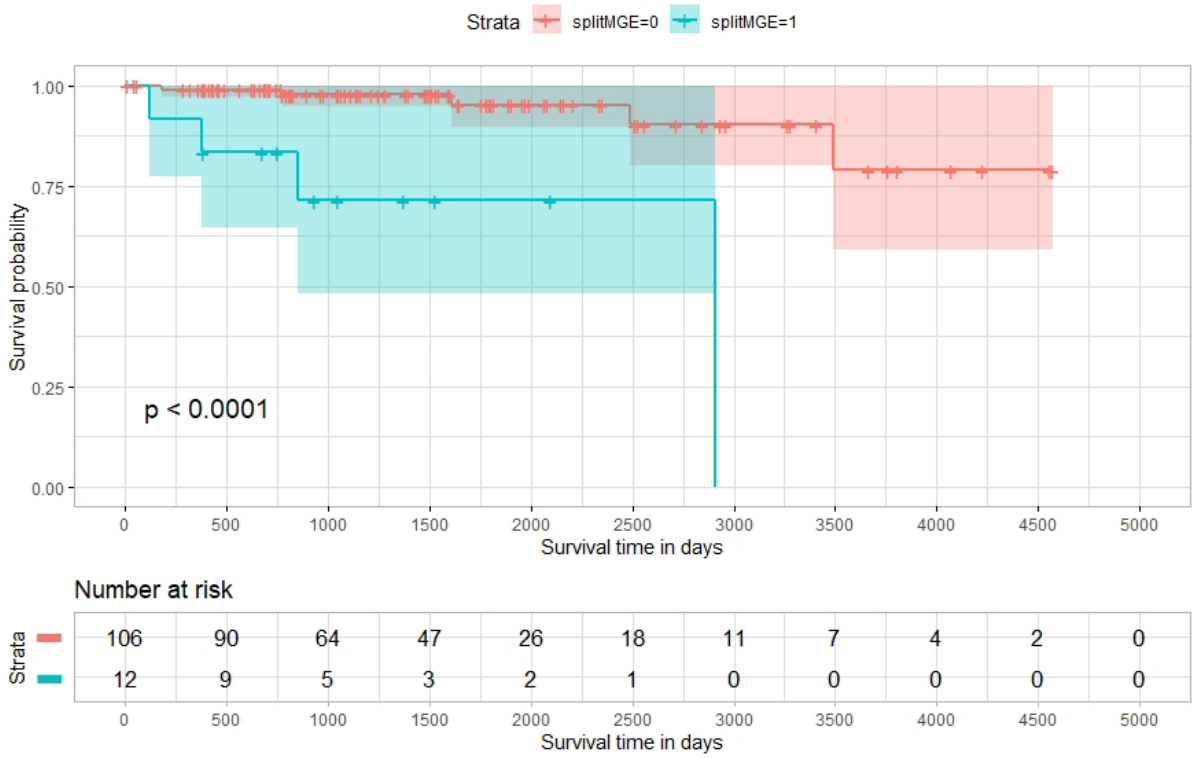
Supplementary figure 27. Kaplan-Meier survival curves for high vs. low *DMD* expression in STAD. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



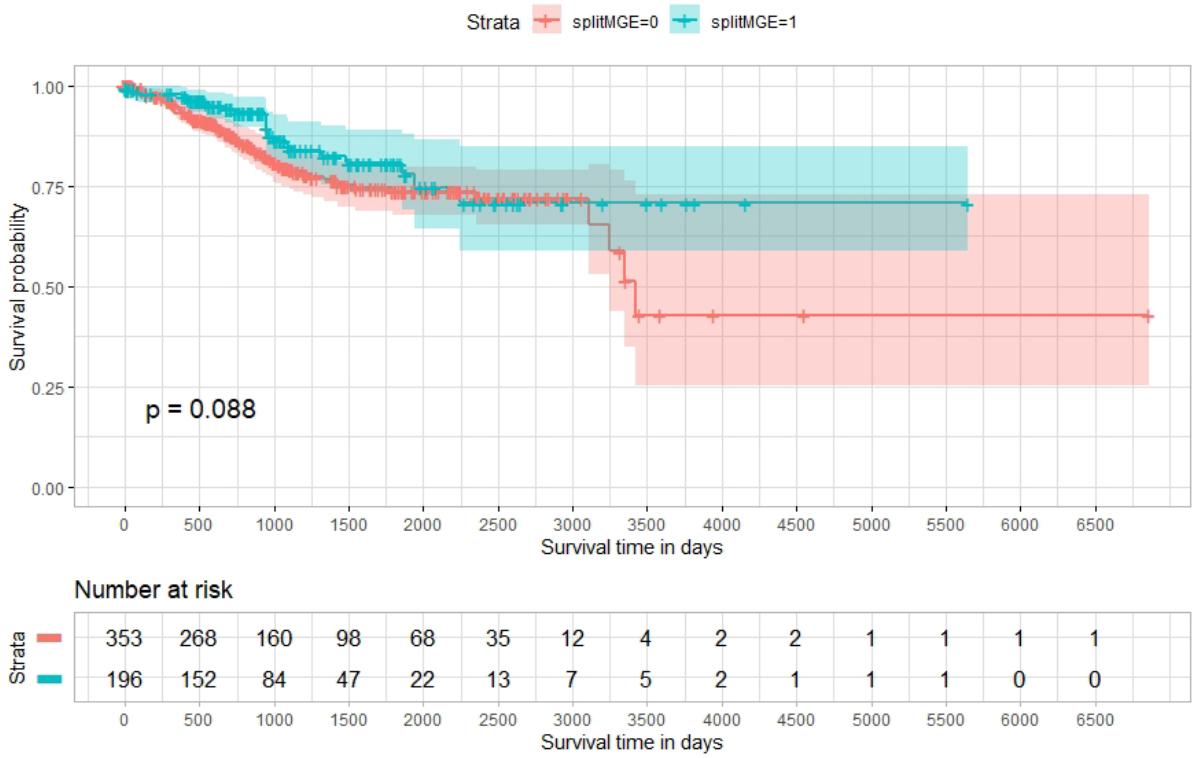
Supplementary figure 28. Kaplan-Meier survival curves for high vs. low *DMD* expression in TGCT. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



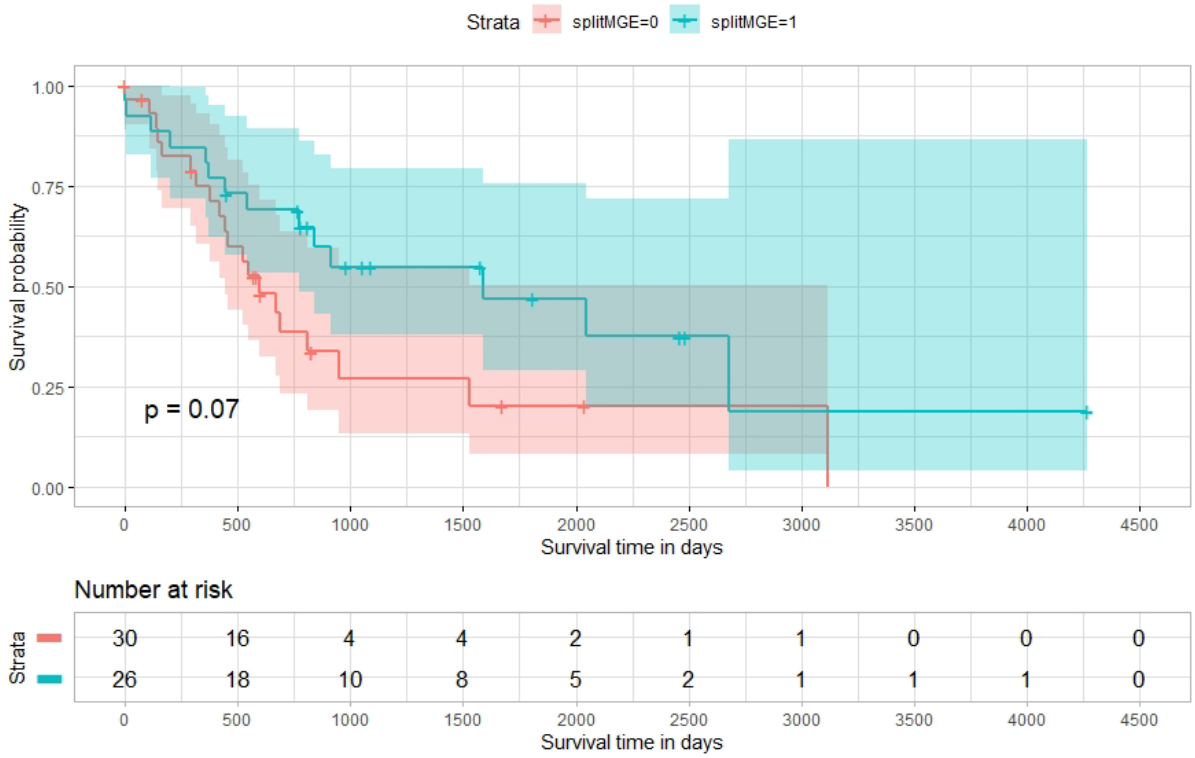
Supplementary figure 29. Kaplan-Meier survival curves for high vs. low *DMD* expression in THCA. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



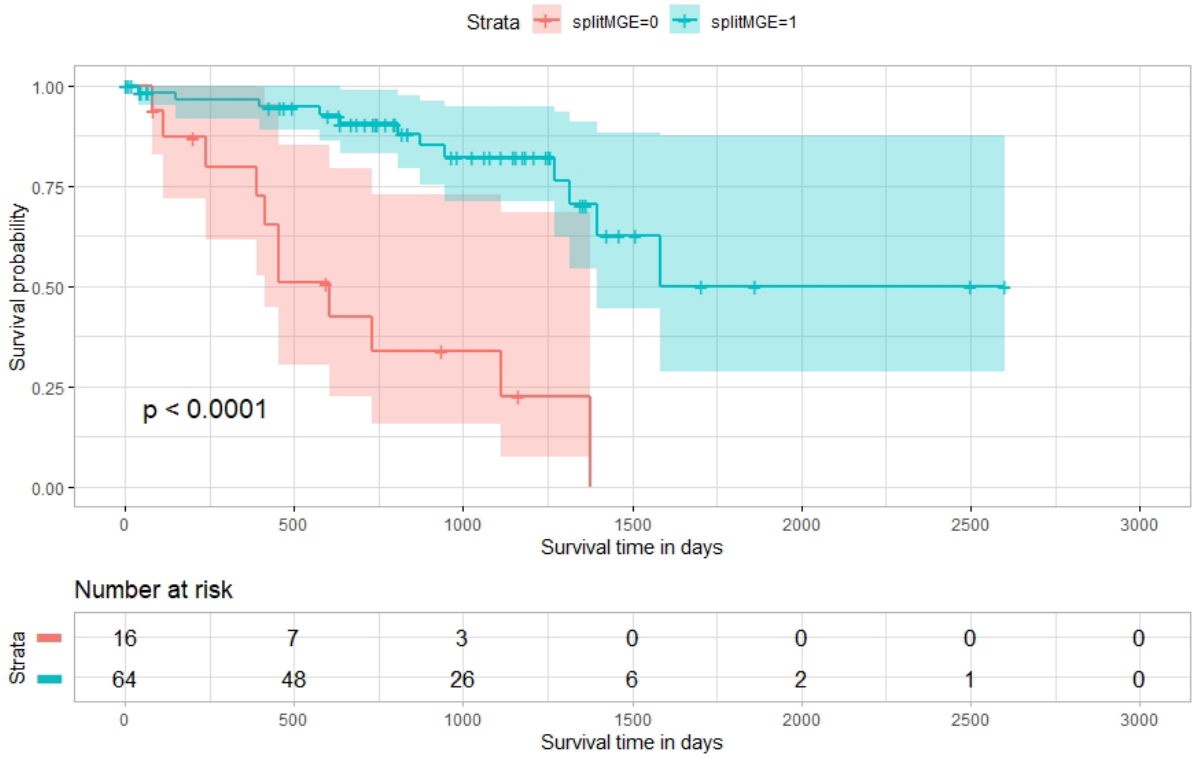
Supplementary figure 30. Kaplan-Meier survival curves for high vs. low *DMD* expression in THYM. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



Supplementary figure 31. Kaplan-Meier survival curves for high vs. low *DMD* expression in UCEC. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



Supplementary figure 32. Kaplan-Meier survival curves for high vs. low *DMD* expression in UCS. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.



Supplementary figure 33. Kaplan-Meier survival curves for high vs. low *DMD* expression in UVM. Red represents the low expression group and green represents the high expressing group. P value calculated using the log-rank test.