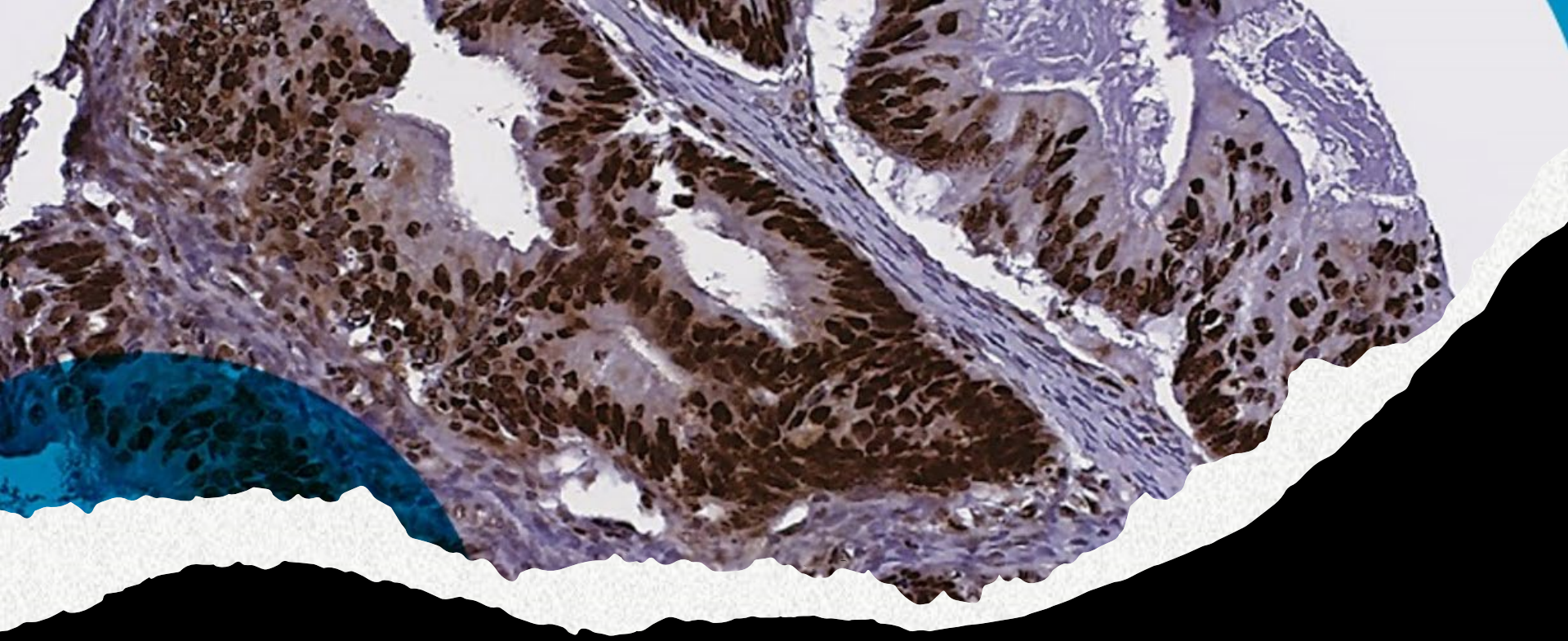
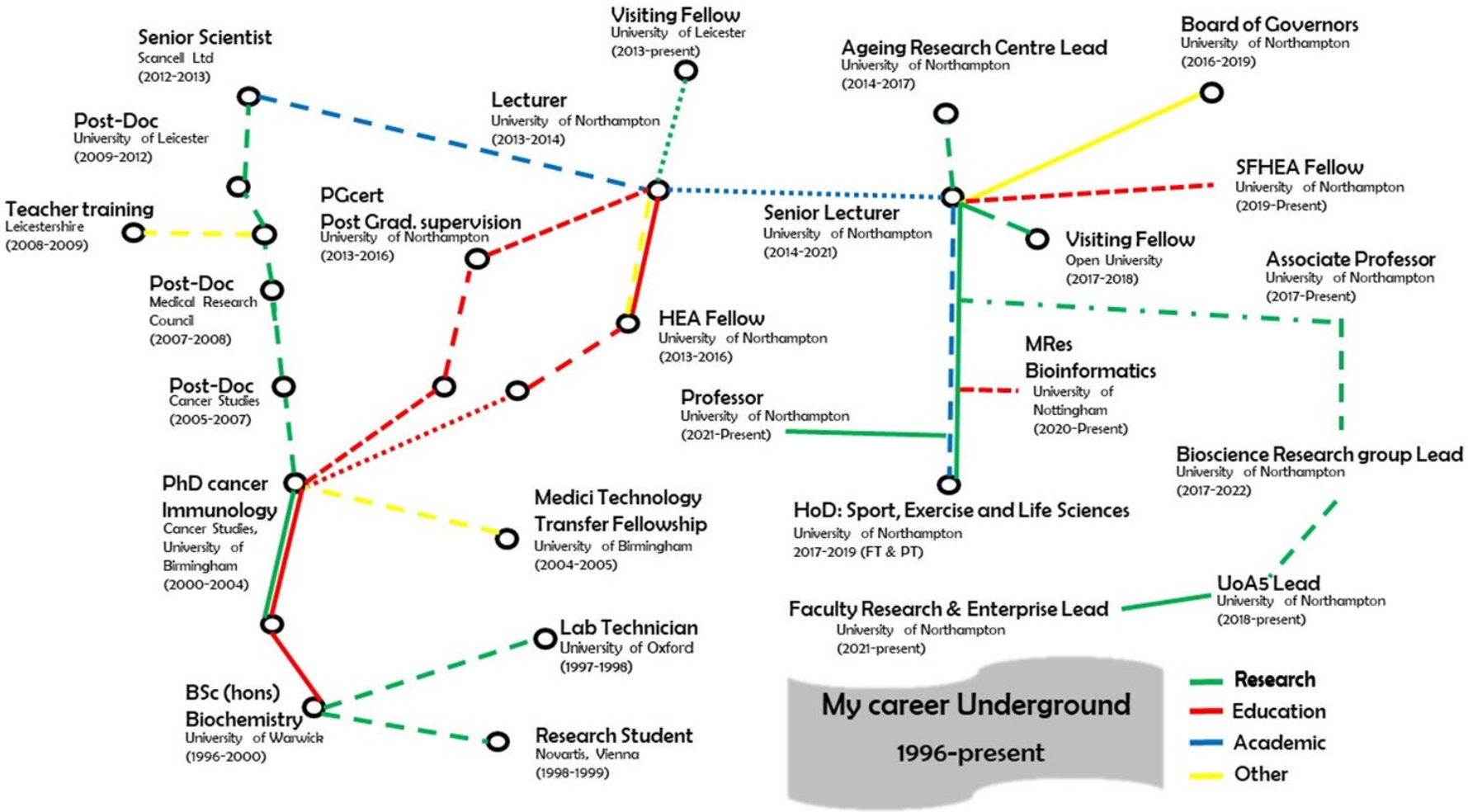


Exploring the landscape of cancer: The use of translational Bioinformatics

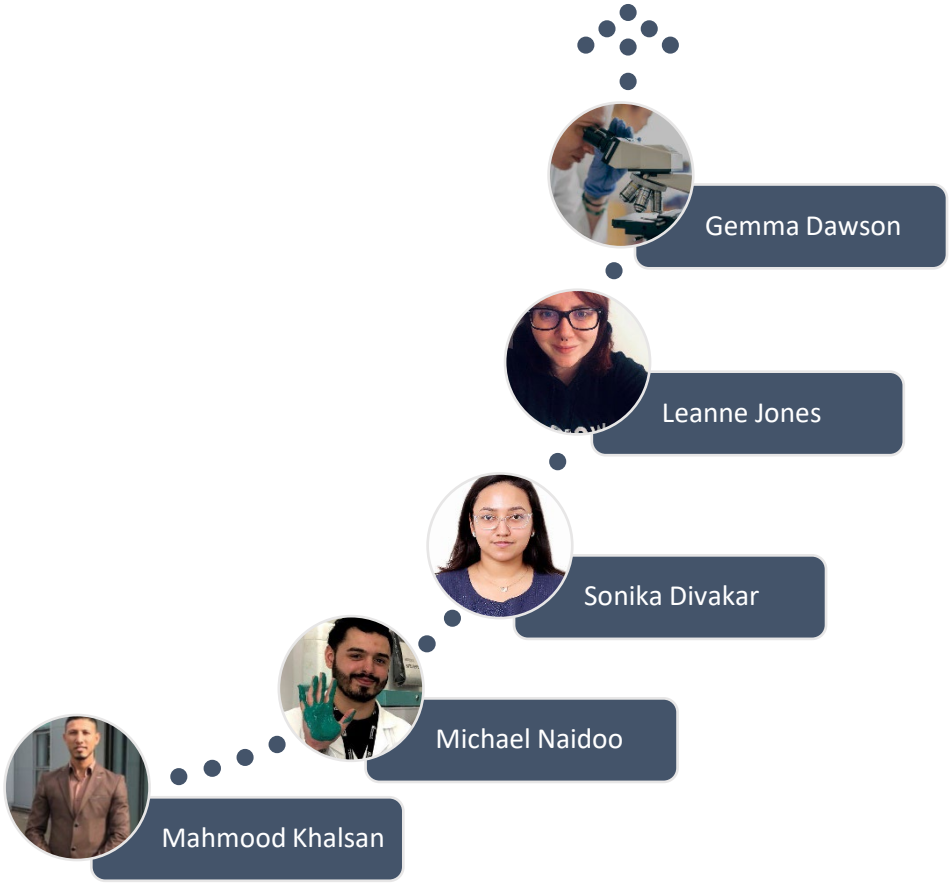
Prof. Lee Machado
@drleemachado

My career underground



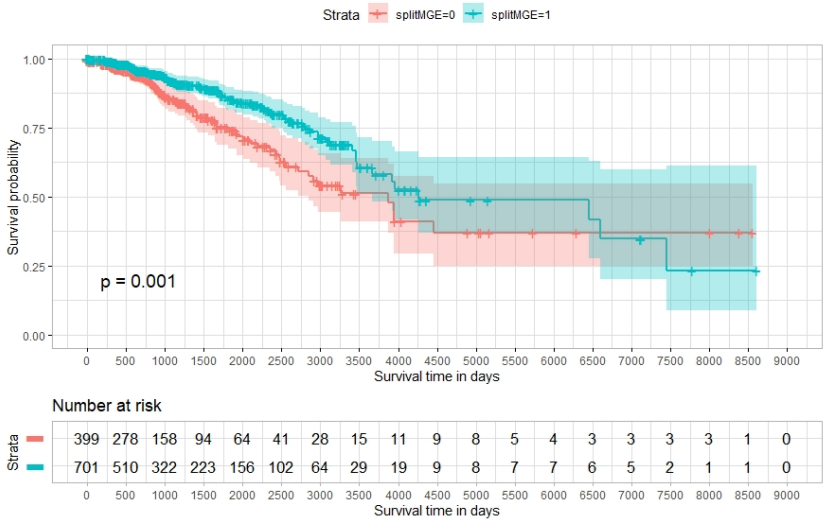
University of Northampton

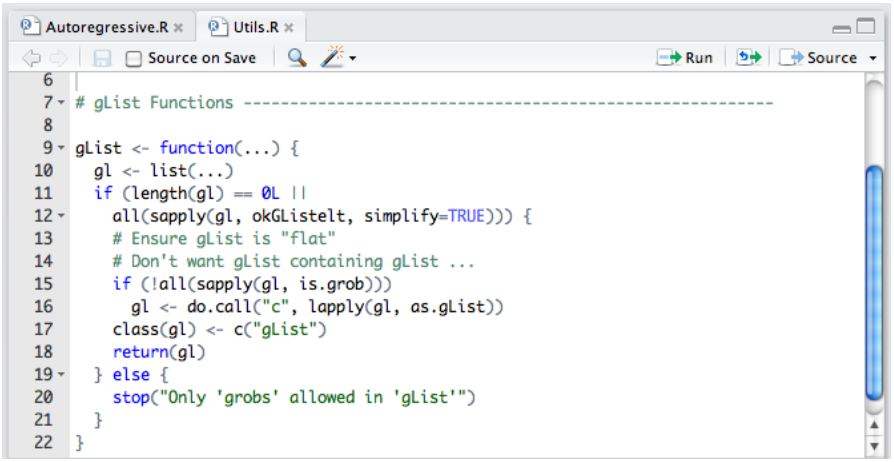
Colleagues working on Cancer projects



University of Northampton

How can we use new advances In big data to help understand the biology of cancer?





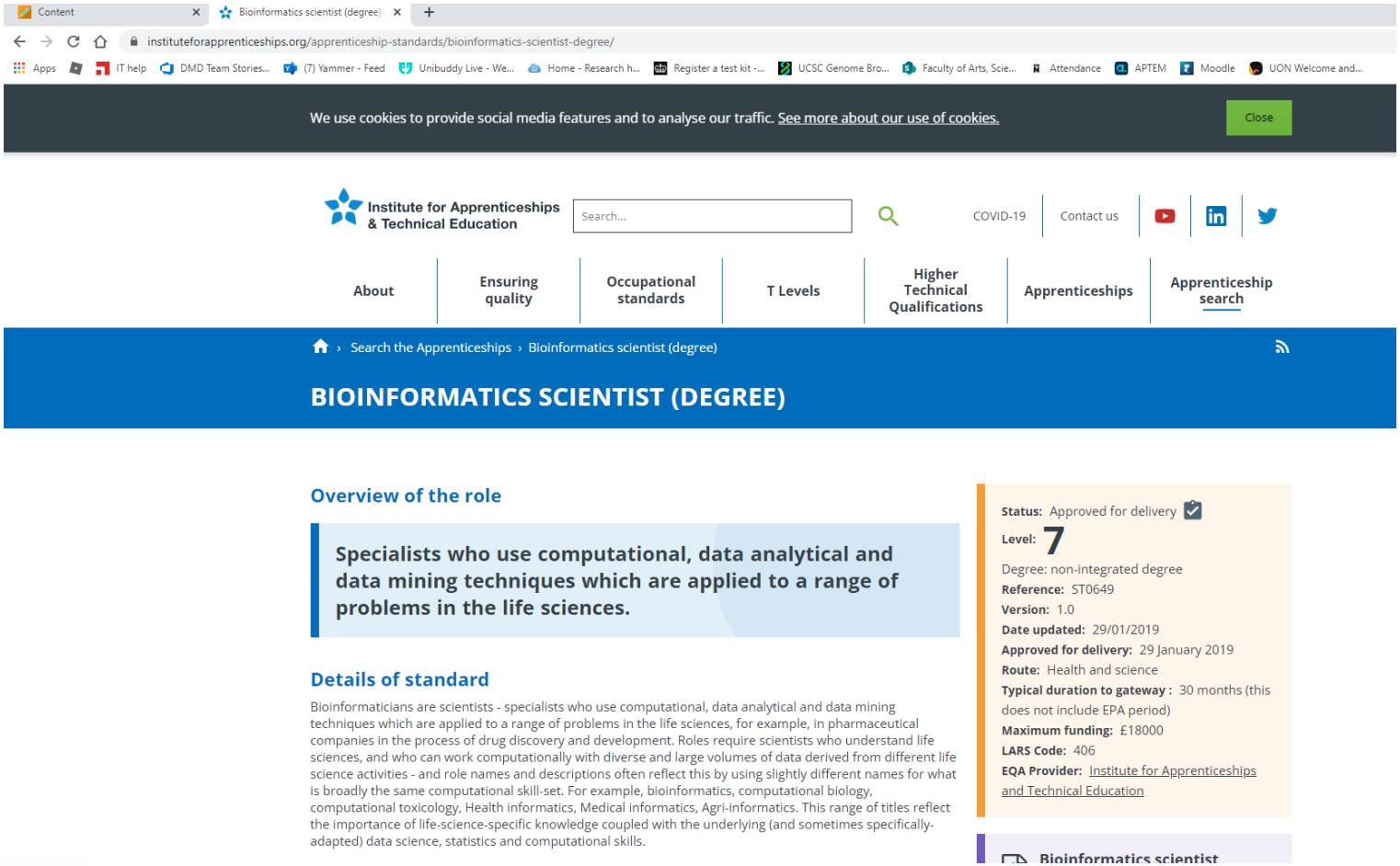
Dr Adam Blanchard

Prof Richard Emes

Prof David Brook

University of Northampton

MRes (level 7) Apprentice in Bioinformatics



University of Northampton

Research project -The *DMD* gene and Cancer

The Duchenne Muscular Dystrophy (*DMD)* gene has been investigated for a potential role in cancer

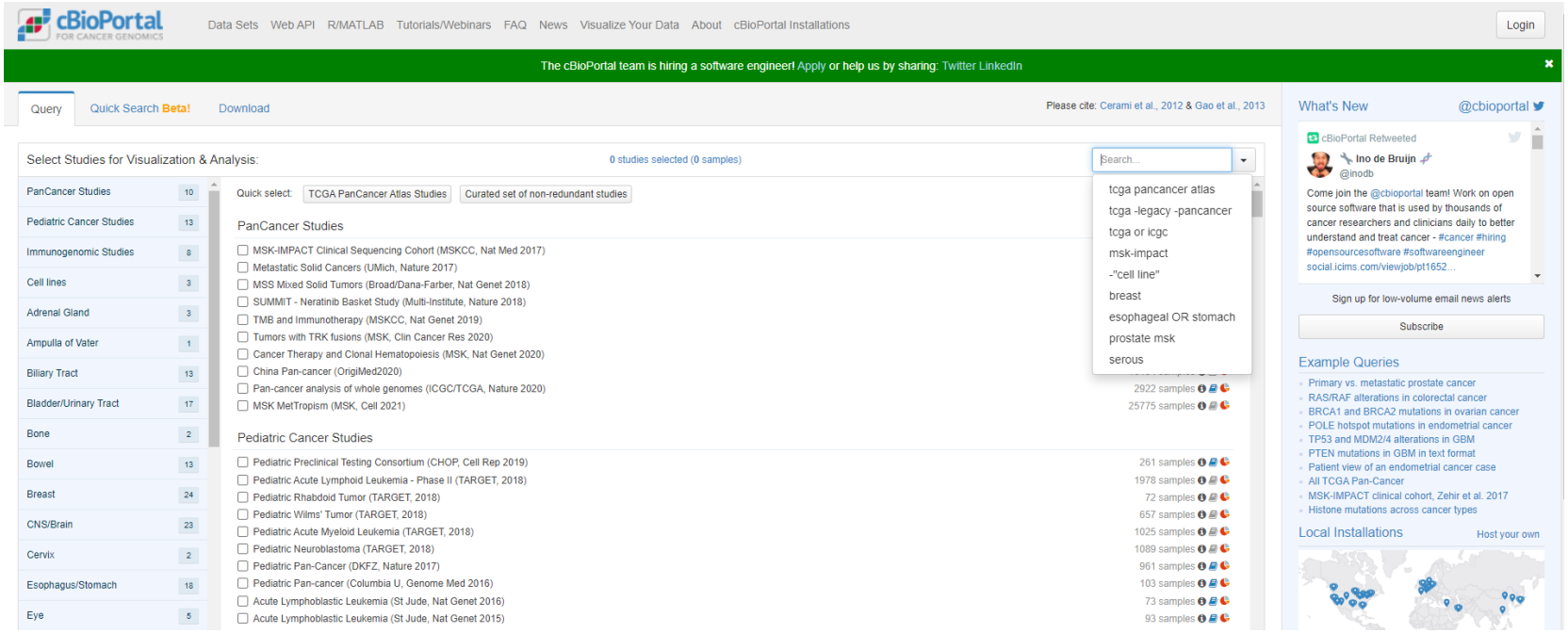
The *DMD* gene is the largest gene in the human genome and expresses several dystrophin proteins

High *DMD* expression associated with leukaemias, renal carcinomas, ependymomas and astrocytoma

Low *DMD* expression associated with melanoma, prostate and gastric cancers (Tan *et al* 2016, Luce *et al.* 2017)

University of Northampton

The cancer genome atlas (TCGA)



Quick activity - what is the frequency of mutation in DMD gene in Esophageal cancer?

University of Northampton

https://www.cbioportal.org/

The frequency of mutation in TCGA cancers

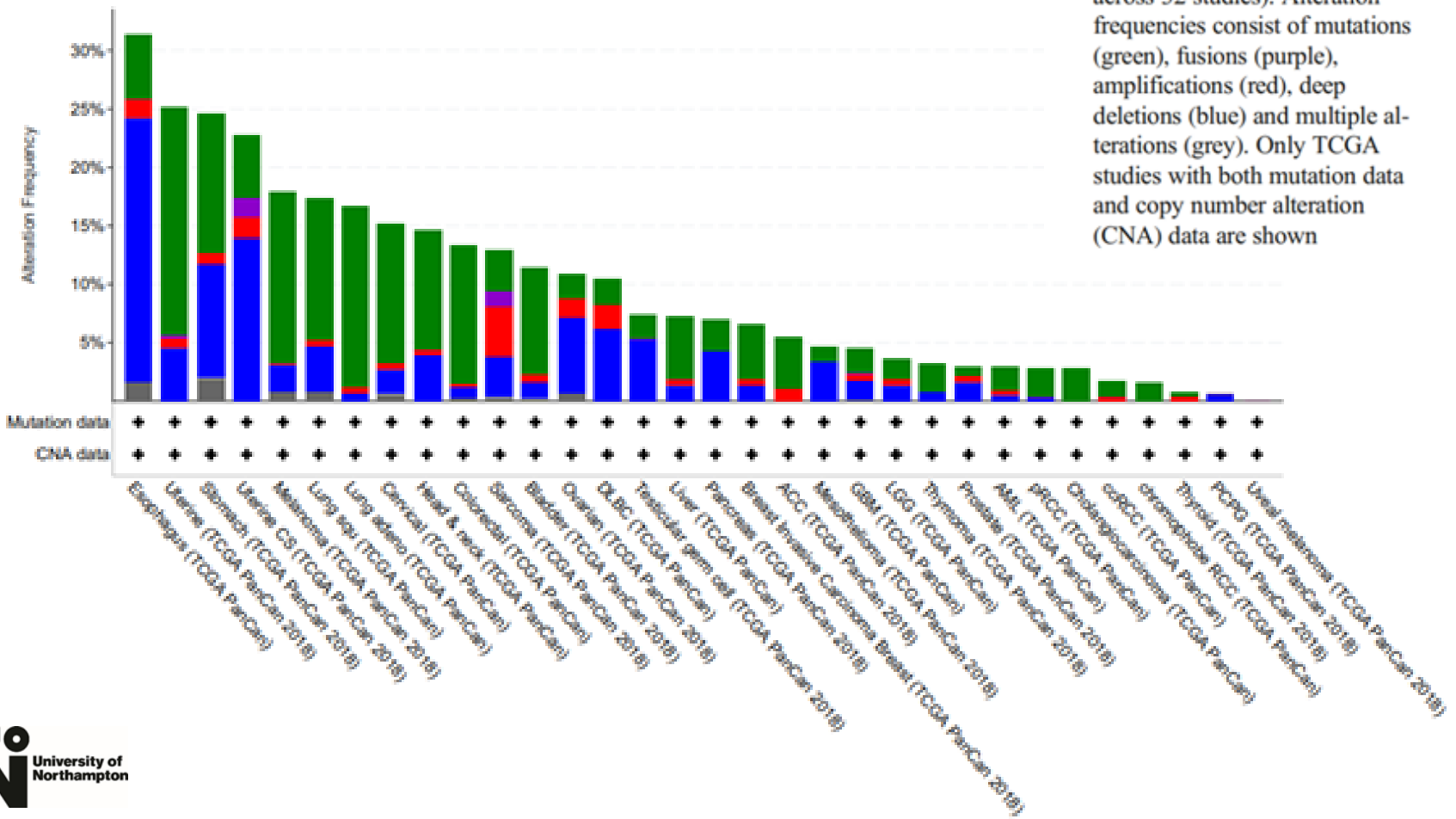


Fig. 2 cBioportal was used to rank *DMD* alteration frequencies across the TCGA PanCancer Atlas studies (10,953 patients across 32 studies). Alteration frequencies consist of mutations (green), fusions (purple), amplifications (red), deep deletions (blue) and multiple alterations (grey). Only TCGA studies with both mutation data and copy number alteration (CNA) data are shown

University of Northampton

TCGA has more than just mutation data!

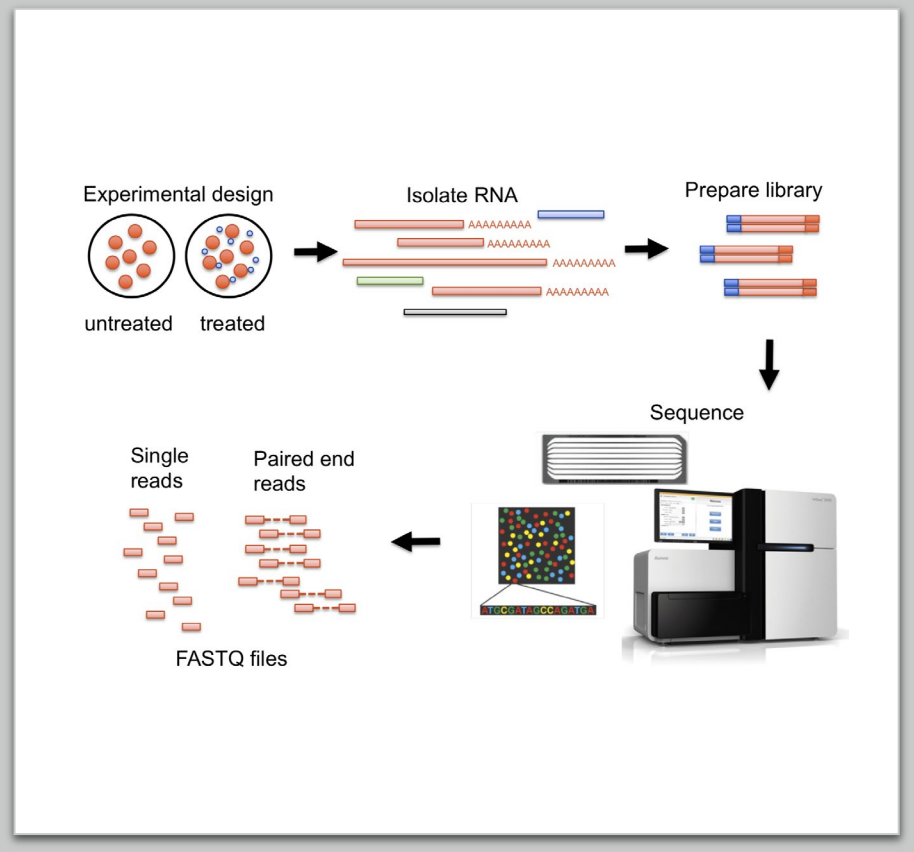
University of Northampton

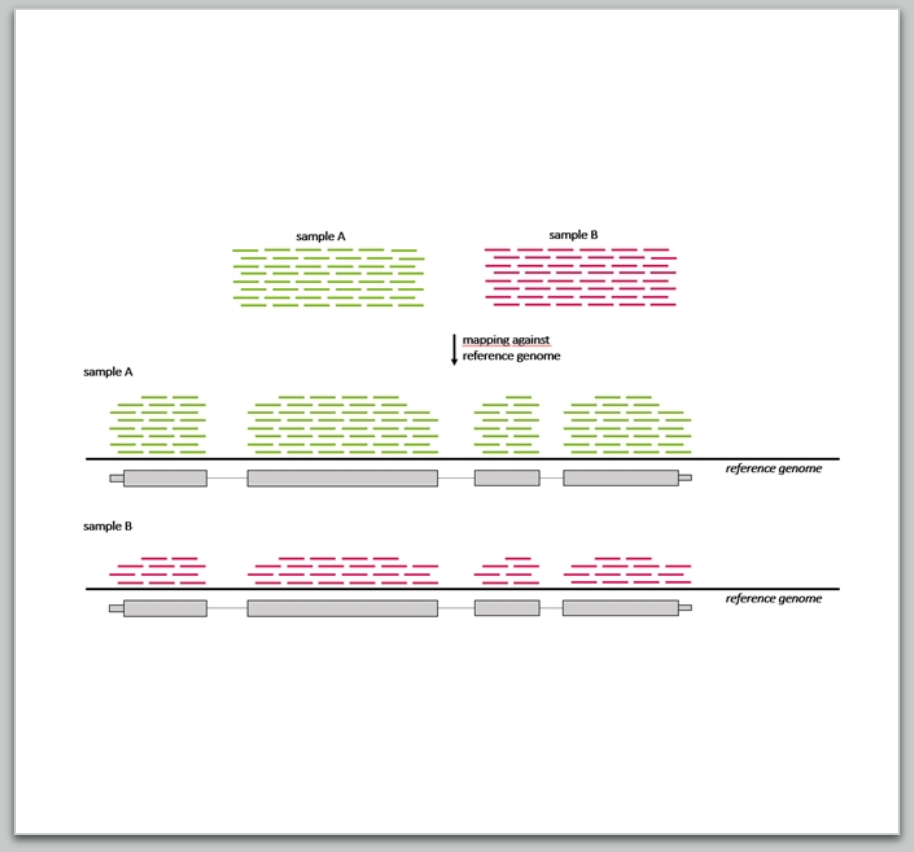
The cancer genome atlas (TCGA)

Table S1. Definition of TCGA cancer type abbreviations and mapping to GTEx tissues. Related to STAR Methods.

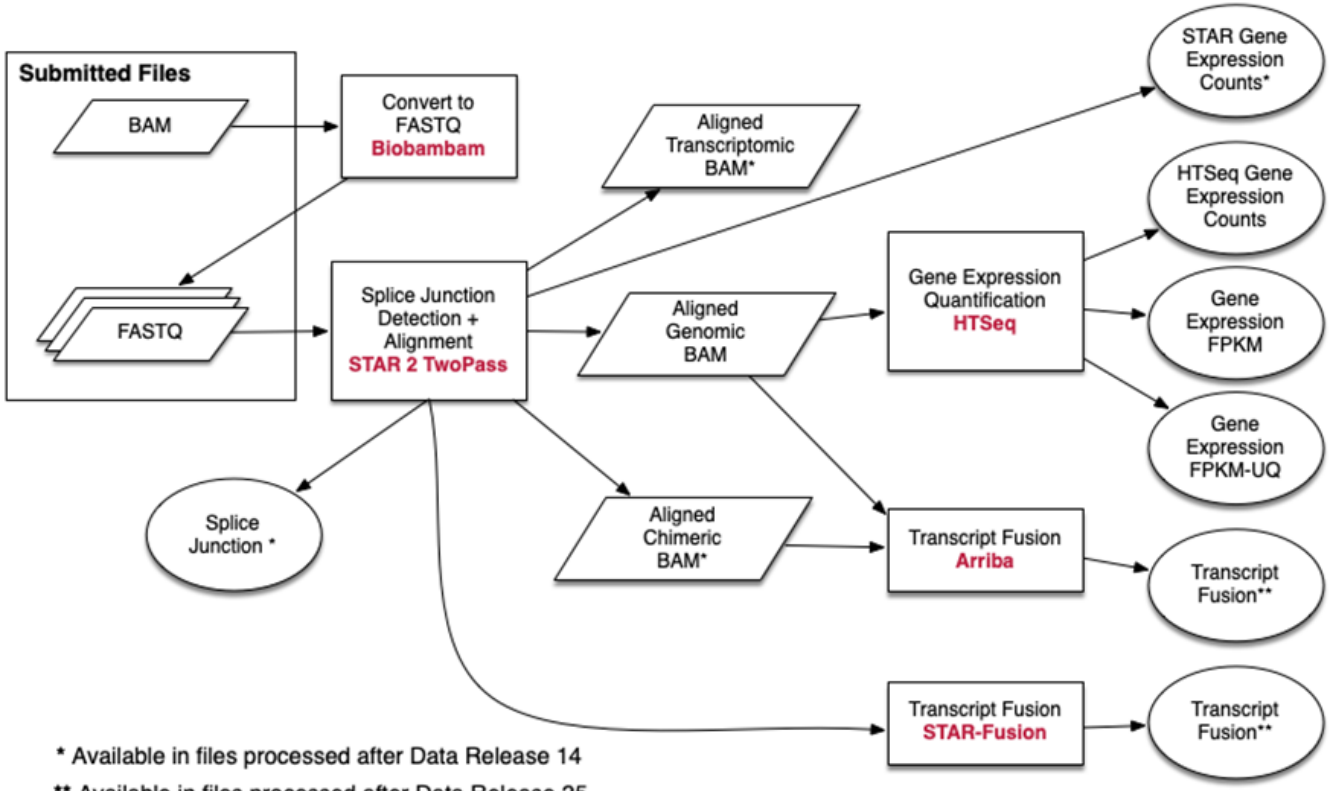
TCGA cancer type	Description	GTEx tissue type
ACC	adrenocortical carcinoma	adrenal gland
BLCA	bladder urothelial carcinoma	bladder
BRCA	breast invasive carcinoma	breast
CESC	cervical squamous cell carcinoma and endocervical adenocarcinoma	cervix uteri
CHOL	cholangiocarcinoma	liver
COAD	colon adenocarcinoma	colon
DLBC	lymphoid neoplasm diffuse large B-cell lymphoma	--
ESCA	esophageal carcinoma	esophagus
GBM	glioblastoma multiforme	brain
HNSC	head and neck squamous cell carcinoma	salivary gland
KICH	kidney chromophobe	kidney
KIRC	kidney renal clear cell carcinoma	kidney
KIRP	kidney renal papillary cell carcinoma	kidney
LGG	brain lower grade glioma	brain
LIHC	liver hepatocellular carcinoma	liver
LUAD	lung adenocarcinoma	lung
LUSC	lung squamous cell carcinoma	lung
MESO	mesothelioma	--
OV	ovarian serous cystadenocarcinoma	ovary
PAAD	pancreatic adenocarcinoma	pancreas
PCPG	pheochromocytoma and paraganglioma	adrenal gland
PRAD	prostate adenocarcinoma	prostate
READ	rectum adenocarcinoma	colon
SARC	sarcoma	--
SKCM	skin cutaneous melanoma	skin
STAD	stomach adenocarcinoma	stomach
TGCT	testicular germ cell tumors	testis
THCA	thyroid carcinoma	thyroid
THYM	thymoma	--
UCEC	uterine corpus endometrial carcinoma	uterus
UCS	uterine carcinosarcoma	uterus
UVM	uveal melanoma	--

An introduction to RNA-seq analysis





Recent harmonization of TCGA data



University of Northampton

Fragments Per Kilobase of transcript per Million mapped reads (FPKM)

- normalising for read depth and gene length
- Counts DNA fragments instead of reads
- Used in paired end RNA-seq experiments to avoid bias

Overview

FPKM-UQ is implemented at the GDC on gene-level read counts that are produced by HTSeq[1] and generated using custom scripts[2]. 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

Like HTSeq - count files, FPKM-UQ files are available as tab delimited files with the Ensembl gene IDs in the first column and the expression values in the second.

Notes

- The scalar (10^9) is added to normalize the values to "**kilo** base" and "**million** mapped reads."
- FPKM-UQ values tend to be much higher than FPKM values because of the large difference between the total mapped number of reads in an alignment and the mapped number of reads to one gene.

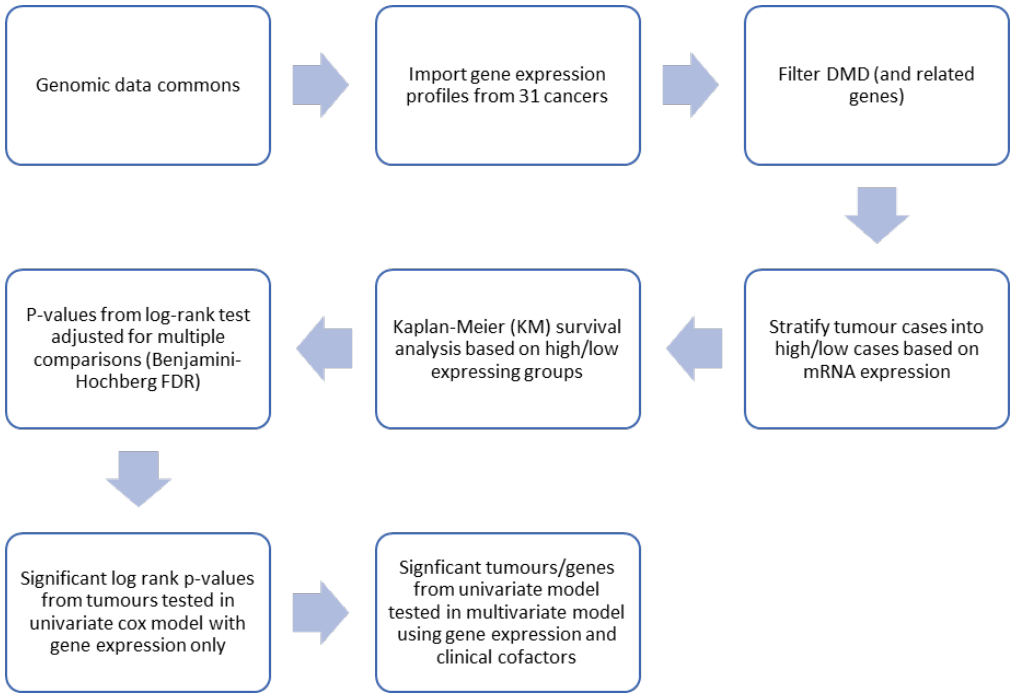
Tools

1. HTSeq Website

References

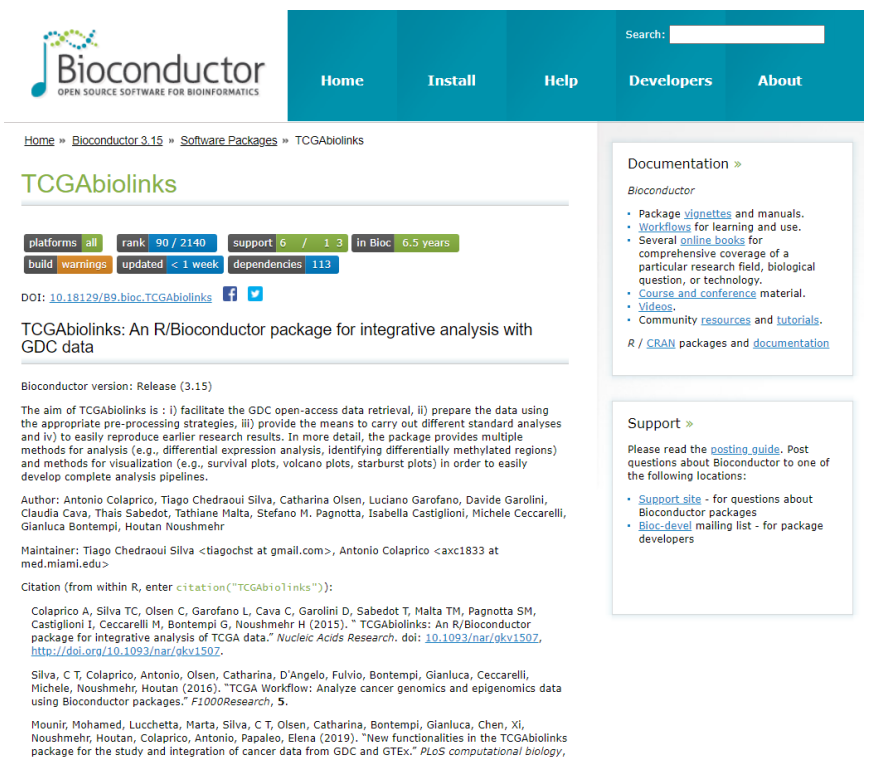
1. Anders, S., Pyl, P.T. and Huber, W., 2014. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics, p.btu638.
2. GDC mRNA-Seq Documentation

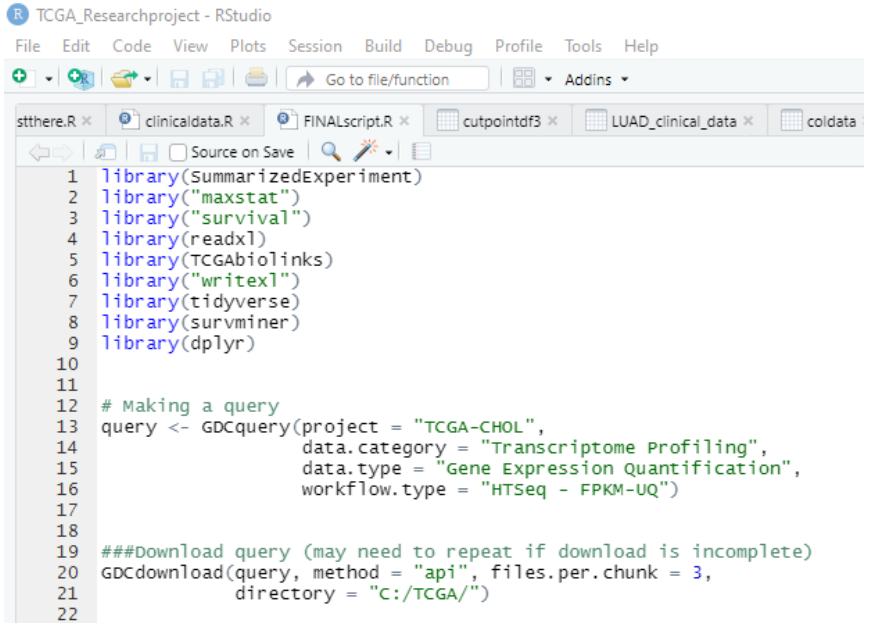
Bioinformatic pipeline in R for Biomarker discovery



University of Northampton

The use of R libraries to analyse TCGA data





University of Northampton

Dichotomisation of DMD gene expression

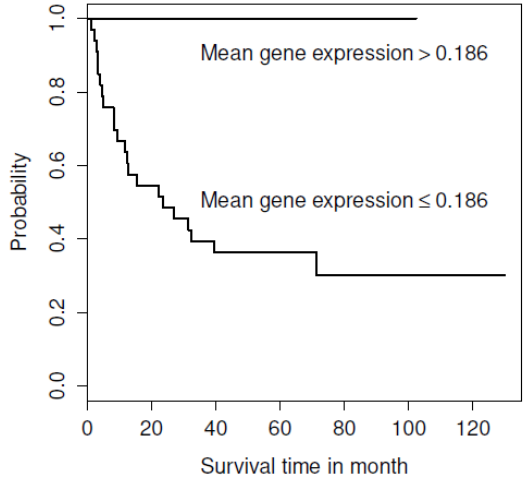


Figure 2: Kaplan-Meier curves of two groups of DLBCL patients separated by the cutpoint 0.186 mean gene expression.

4 Gene Expression Profiling

The distinction of two types of diffuse large B-cell lymphoma by gene expression profiling is studied by Alizadeh et al. (2000). Hothorn and Lausen (2003) suggest the mean gene expression (MGE) as quantitative factor for the discrimination between two groups of patients with respect to overall survival time. The dataset `DLBCL` is included in the package. The maximally selected log-rank statistic for cutpoints between the 10% and 90% quantile of MGE using the upper bound of the *P*-value by Hothorn and Lausen (2003) can be computed by

```
>library("maxstat")
>library("survival")
>data("DLBCL", package="maxstat")
>mtHL <- maxstat.test(Surv(time, cens) ~ MGE,
+         data=DLBCL, smethod="LogRank", pmethod="HL")
>mtHL

Maximally selected LogRank statistics using HL

data:  Surv(time, cens) by MGE
M = 3.171, p-value = 0.02218
sample estimates:
estimated cutpoint
         0.1860526
```

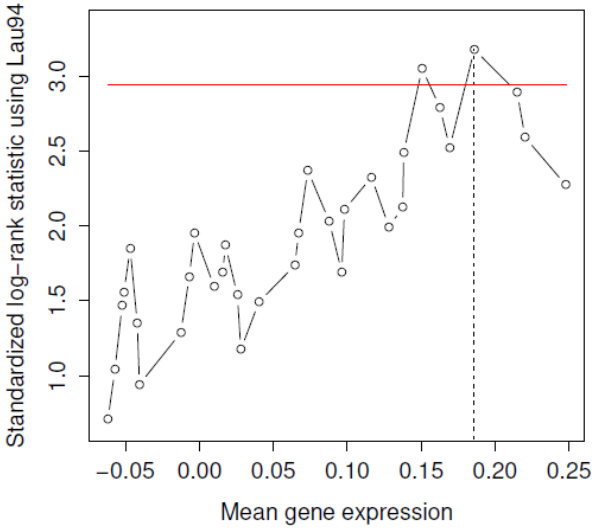
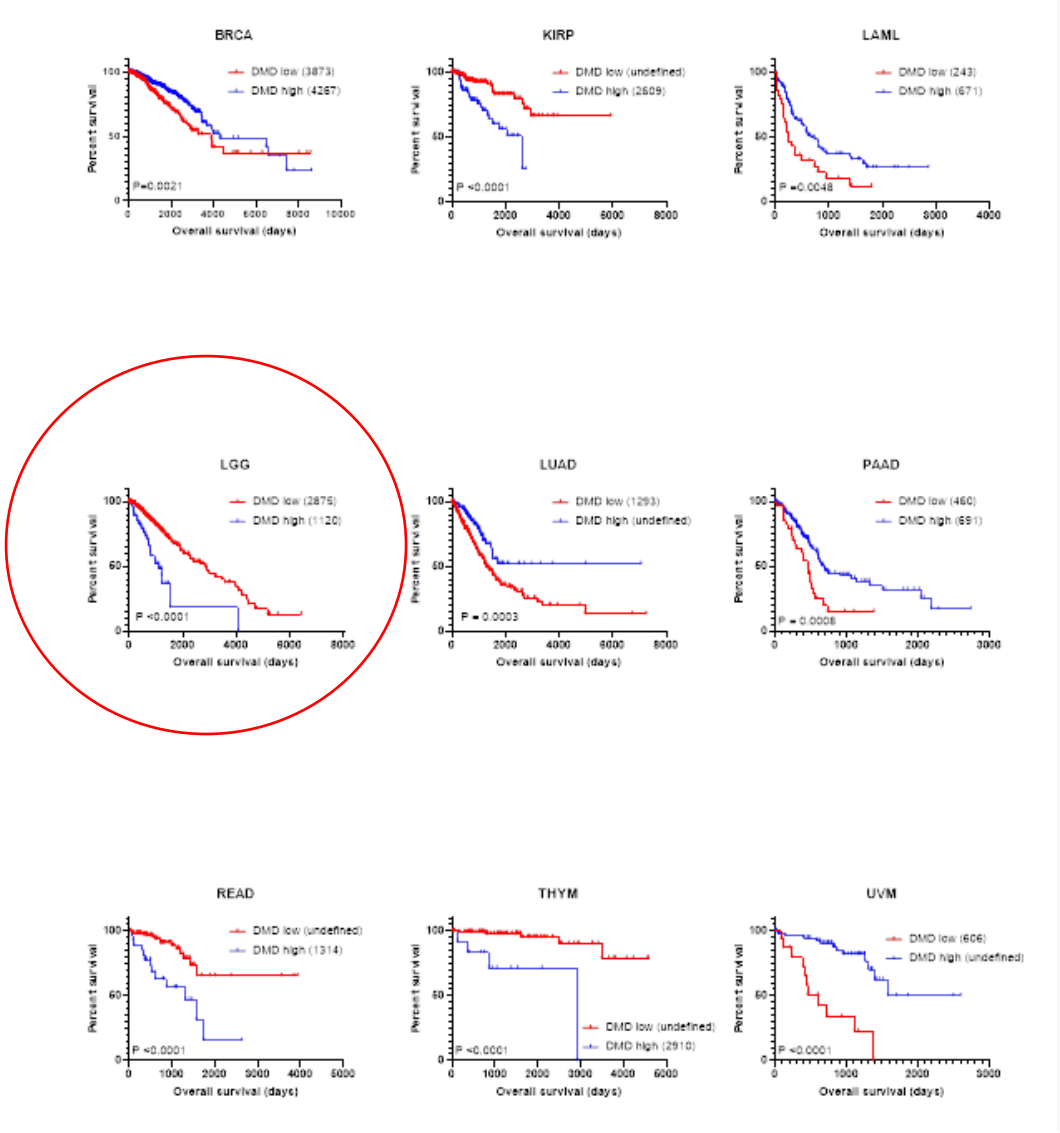


Figure 1: Absolute standardized log-rank statistics and significance bound based on the improved Bonferroni inequality.

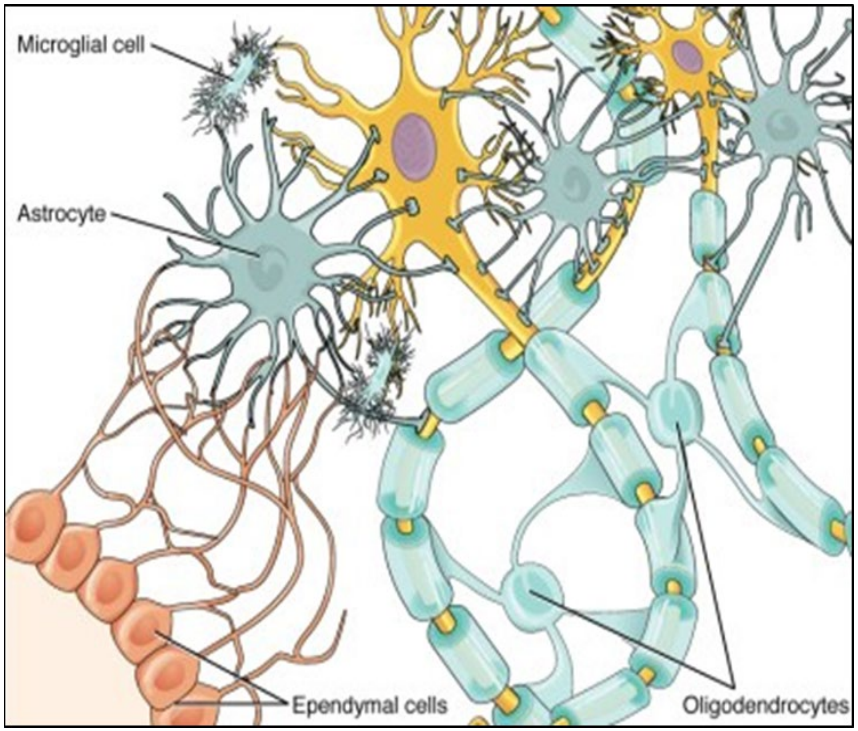
DMD Kaplan Meier survival curves distinguish patients with different survival outcomes



University of Northampton

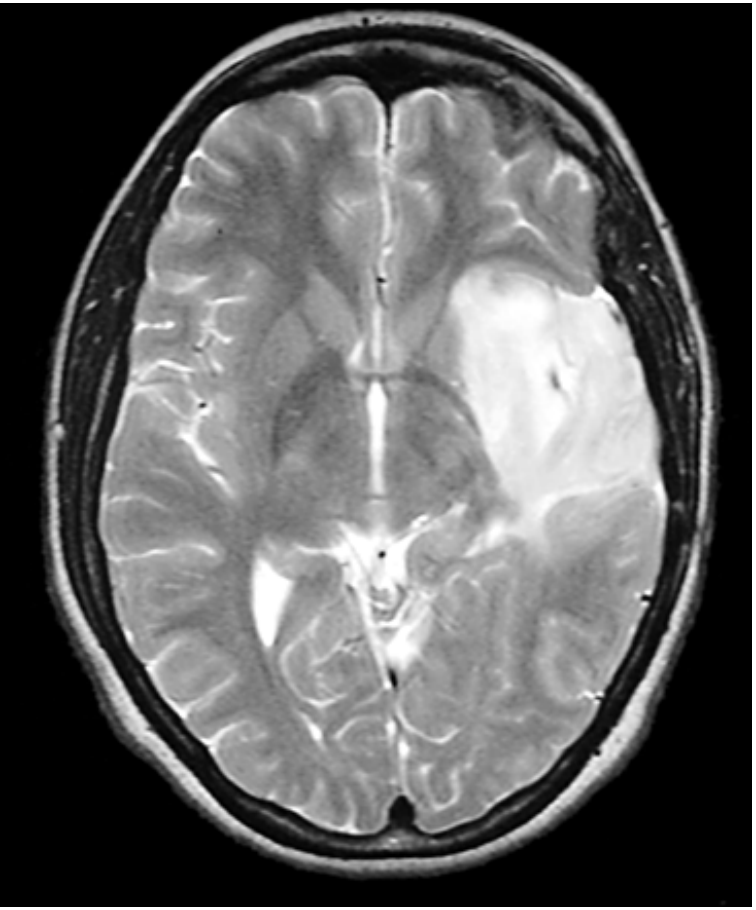
Glioma

- The most common CNS cancers
- Cancers that originate from Glial cells – cells that support the CNS
- Glial cells comprise of astrocytes, oligodendrocytes, ependymal cells and microglial cells
- 25.5% of all CNS tumours and 80.8% of malignant CNS tumours
- Glioma covers many histologies ranging from grades I-IV



The Glia Cells | Function, Role, Types, Characteristics & Summary. *The Human Memory*. doi: https://human-memory.net/the-glia-cells/

University of Northampton



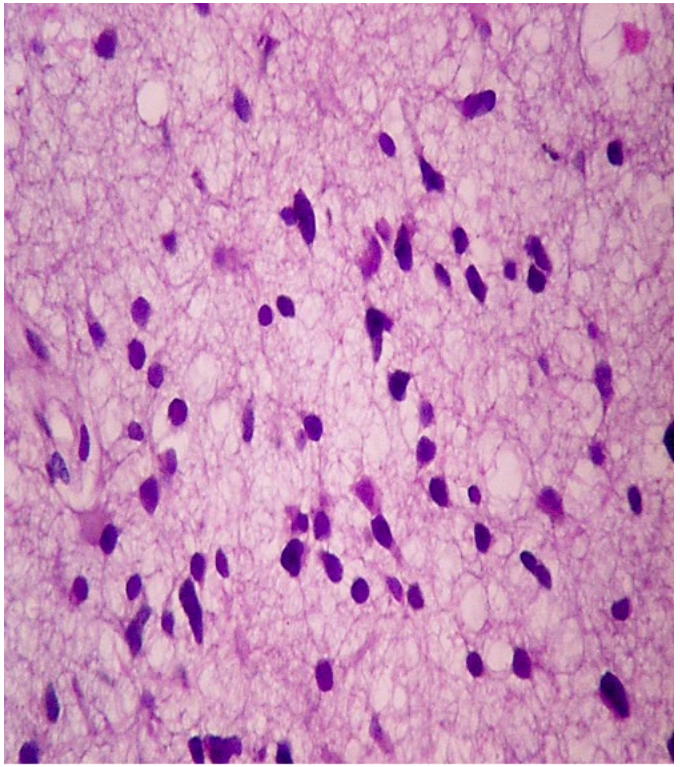
T2 weighted magnetic resonance image (MRI) scan of 28 year old woman with complex partial seizures

Diffuse Low-grade Glioma

- Diffuse low-grade gliomas (dLGG) are grade II gliomas with diffuse infiltrating properties
- Astrocytomas and oligodendrogliomas
- Represent 15% of all gliomas
- Average age at diagnosis is 41.4 years old
- Survival is between 5-8 years
- All eventually progress into high-grade lethal tumours

Grade II Diffuse Astrocytomas

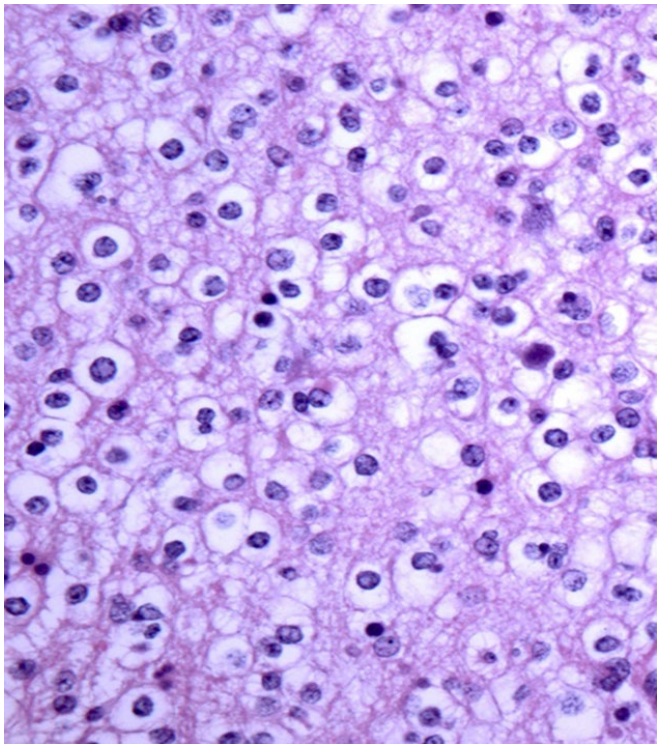
- Originate from astrocytes
- Most abundant cell in the brain
- Many roles, including neuron support, BBB, fluid/ion/pH/neurotransmitter homeostasis, energy and metabolism
- 50% 5-year survival after diagnosis



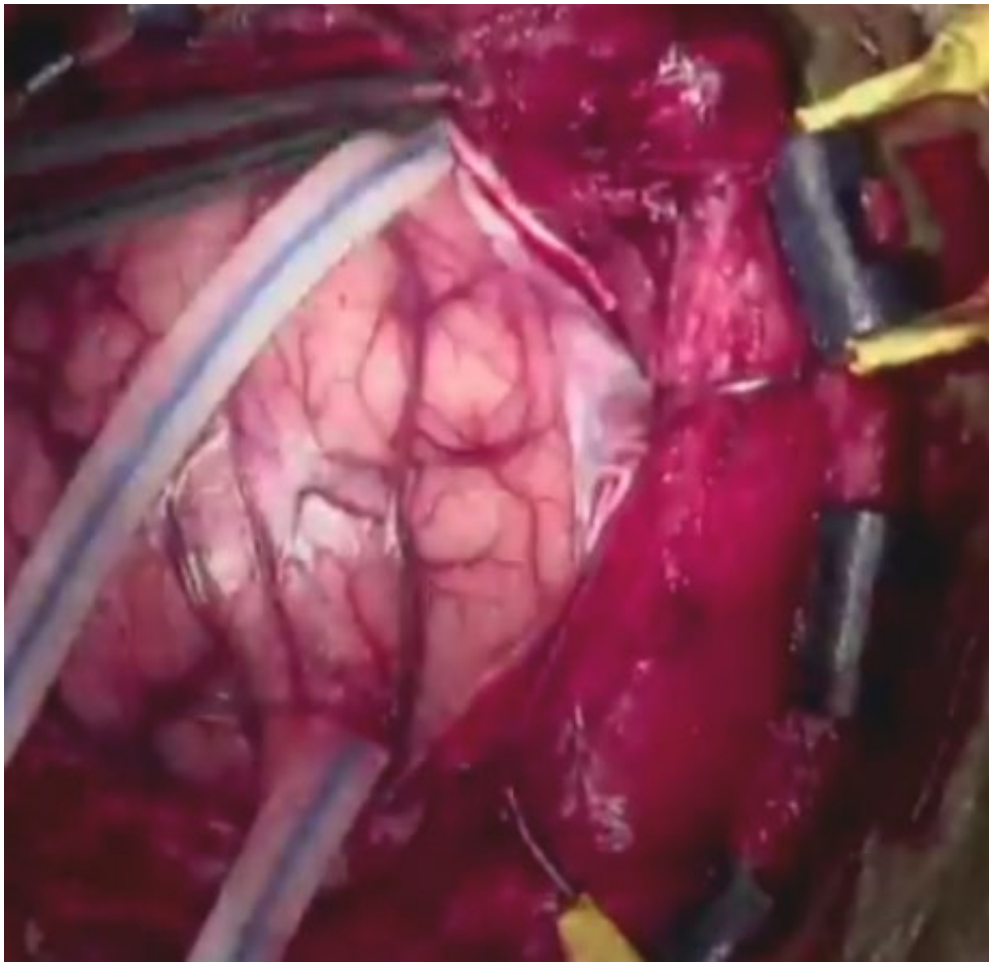
Astrocytoma, consisting of fibrillary neoplastic astrocytes on a loose tumour matrix background.

Grade II Oligodendrogliomas

- Originate from oligodendrocytes
- Form the myelin sheath on the neuronal axons in the CNS
- 40% 5-year survival after diagnosis



Oligodendrogliomas are characterised by uniform-appearing, infiltrating cells with perinuclear clearing in a honeycomb pattern



MIT doctoral student, Steven Keating

Treatments in dLGG

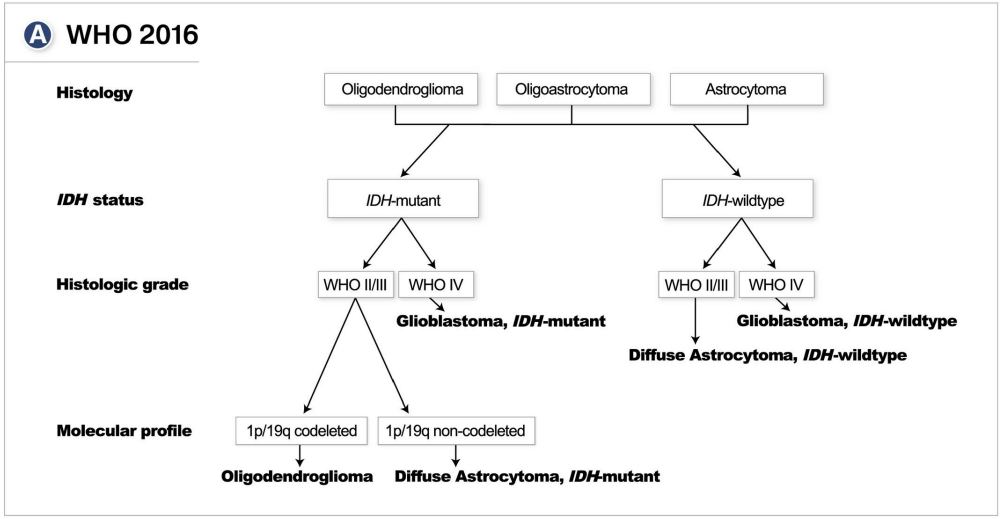
- Surgery - analysis, symptoms and prognosis
- Radiotherapy
- Chemotherapy - Temozolomide

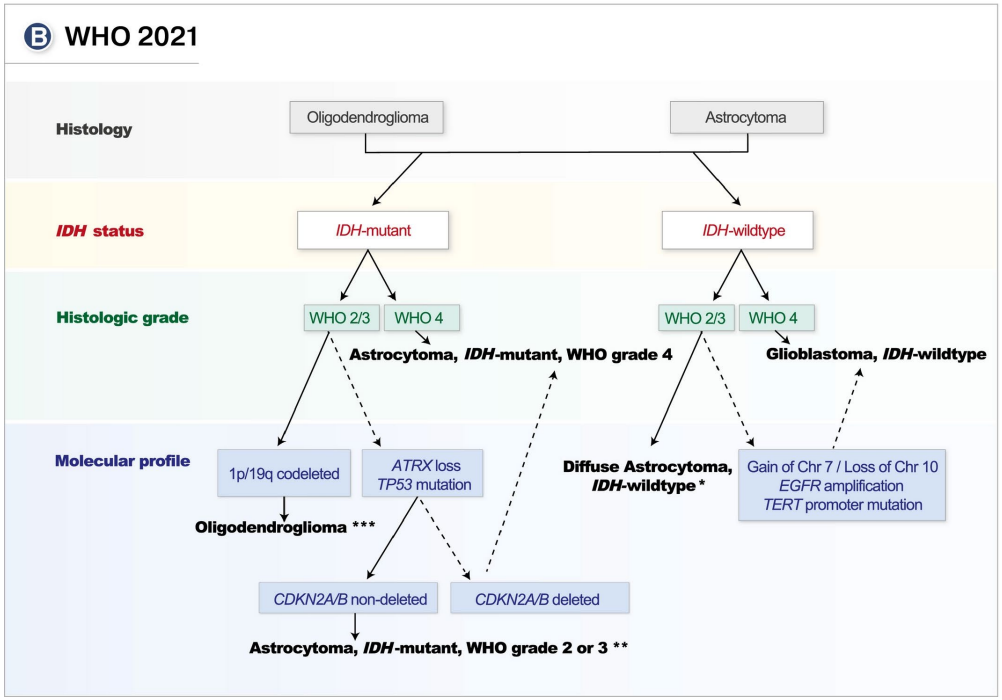
University of Northampton

Genetic biomarkers IDH1 & IDH2

- Isocitrate dehydrogenase – cellular metabolism
- Diagnostic and prognostic marker
- Mutation associated with better survival
- Common in low grade gliomas
- Rare in high grade

WHO Classification system for gliomas





www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

Duchenne muscular dystrophy gene expression is an independent prognostic marker for IDH mutant low-grade glioma

Michael Naidoo[1], Leanne Jones[1], Benjamin Conboy[1], Wael Hamarneh[2], Darwin D'Souza[3], Karen Anthony[1,5]✉ & Lee R. Machado[1,4,5]✉

Alterations in the expression of the Duchenne muscular dystrophy (DMD) gene have been associated with the development, progression and survival outcomes of numerous cancers including tumours of the central nervous system. We undertook a detailed bioinformatic analysis of low-grade glioma (LGG) bulk RNAseq data to characterise the association between *DMD* expression and LGG survival outcomes. High *DMD* expression was significantly associated with poor survival in LGG with a difference in median overall survival between high and low *DMD* groups of over 7 years ($P = < 0.0001$). In a multivariate model, *DMD* expression remained significant ($P = 0.02$) and was an independent prognostic marker for LGG. The effect of *DMD* expression on overall survival was only apparent in isocitrate dehydrogenase (IDH) mutant cases where non-1p/19q co-deleted LGG patients could be further stratified into high/low *DMD* groups. Patients in the high *DMD* group had a median overall

Michael Naidoo

University of Northampton

https://www.nature.com/articles/s41598-022-07223-2

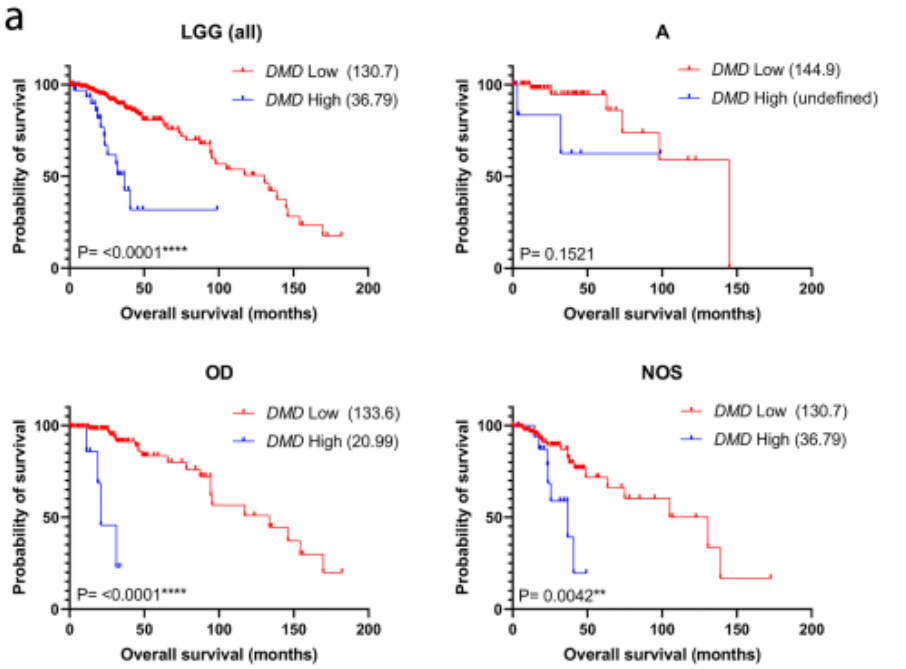
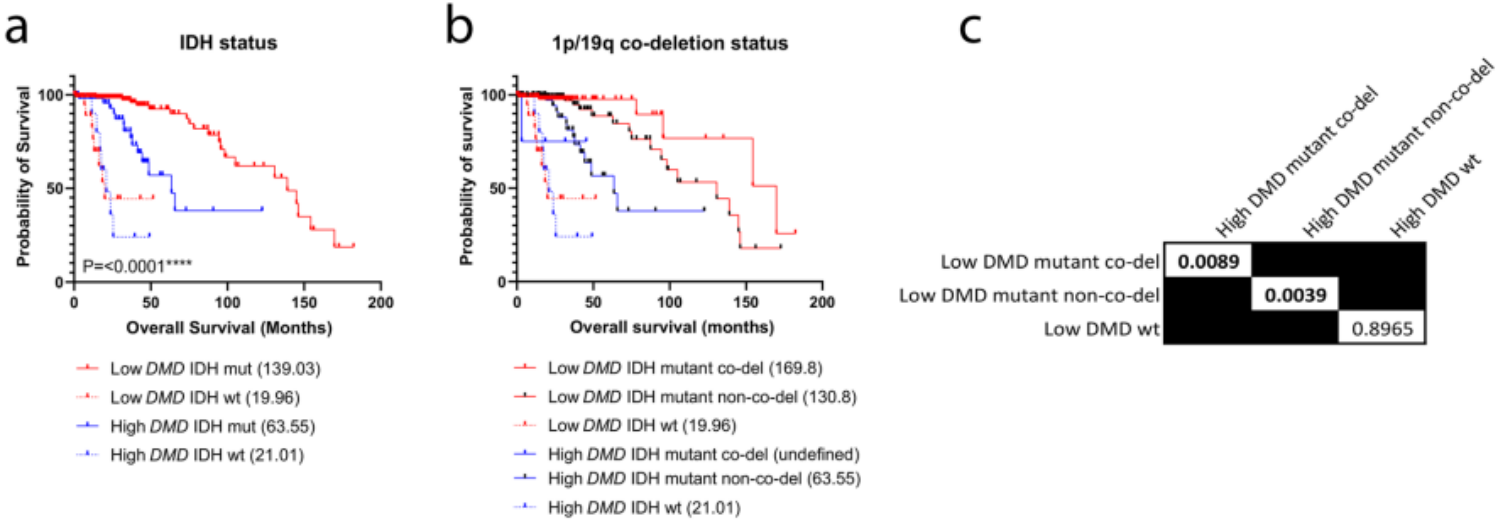


Figure 1. High *DMD* expression is significantly associated with poor survival in LGG. (**a**) TCGA RNAseq data from WHO grade II LGG cases was dichotomised into high (blue) and low (red) *DMD* expressing groups and survival analysis performed in GraphPad using the log-rank test. Tumour subtype analysis was also performed. Numbers in brackets are median overall survival times in months. (**b**) Forest plot revealing the log-rank hazard ratio with 95% confidence intervals and number of patients for each group. *A* astrocytoma, *OD* oligodendroglioma, *NOS* not otherwise specified.

High DMD expression is significantly associated with poor survival in LGG

DMD gene expression further stratifies IDH mutant LGG

- Low *DMD* expression offers a particularly significant survival advantage for IDH mutated patients regardless of 1p/19q co-deletion status compared with high *DMD*



University of Northampton

1p/19q genes are upregulated in patients with high *DMD* expression which worsens survival outcomes for these patients

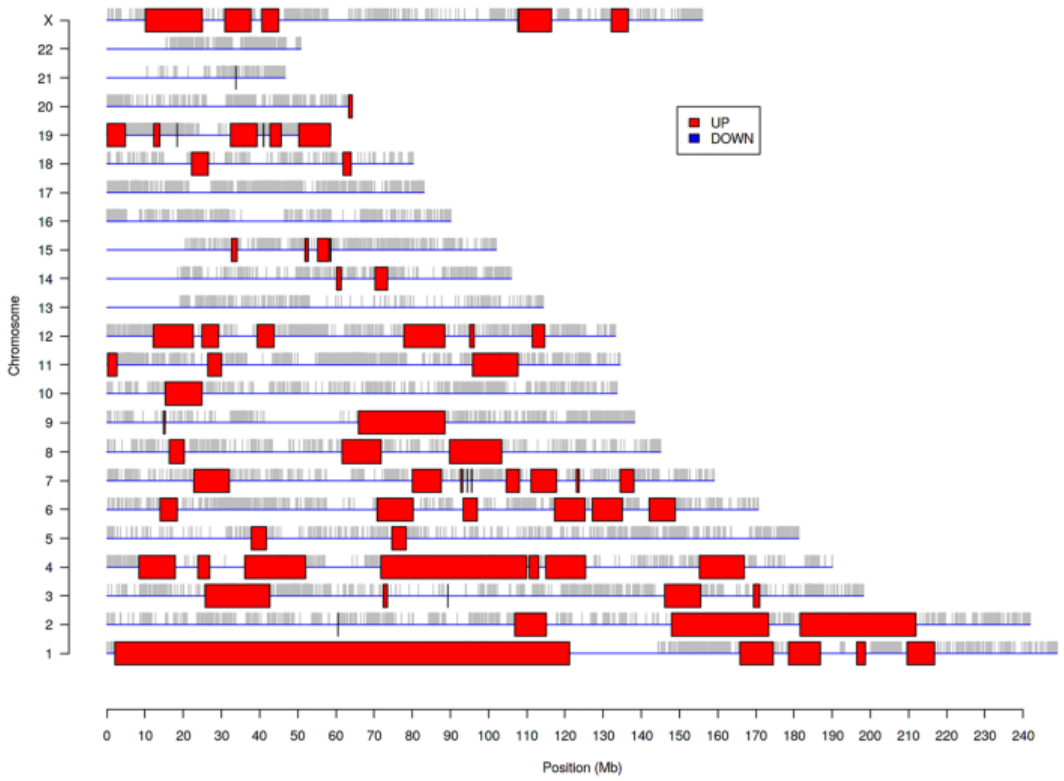


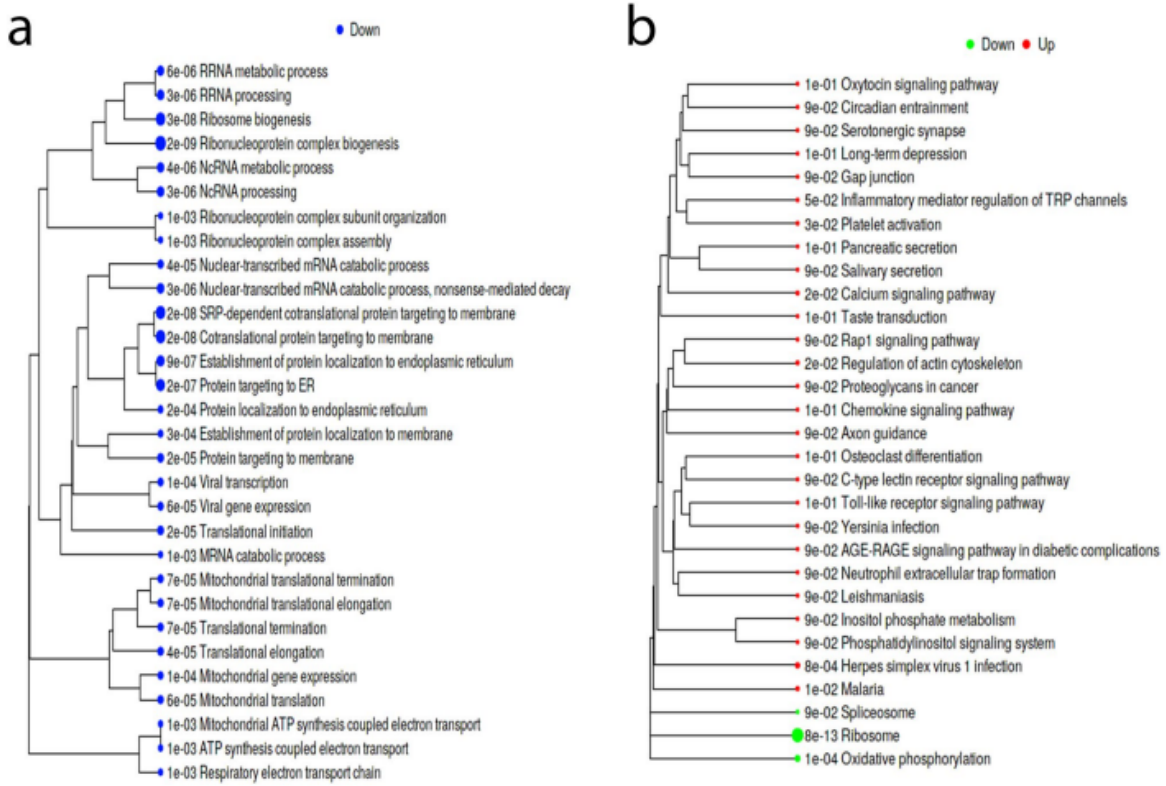
Figure 3. Genomic regions significantly dysregulated in WHO grade II LGG cases with high vs low *DMD* expression. Significantly altered regions are highlighted with coloured boxes.

University of Northampton

Position RElated Data Analysis (PREDA)

Network and pathway analysis of differentially expressed genes (DEGs) implicates development and ribosome biogenesis

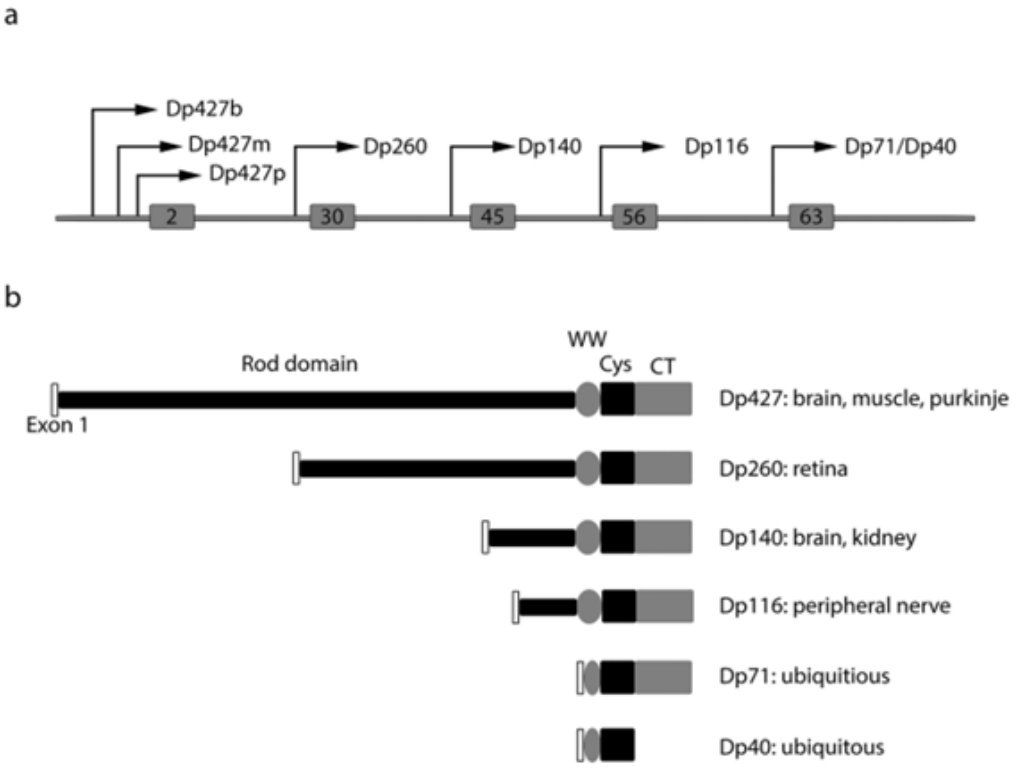
- In high vs low *DMD* cases, ribosome pathways are downregulated and selected DEGs are enriched in processes relating to development



Ge, Steven Xijin et al. “iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data.” *BMC bioinformatics* vol. 19,1 534. 19 Dec. 2018,

DMD gene products

Fig. 1 Structure of the *DMD* gene. **a** Schematic diagram showing the location of the seven independent promoters within the *DMD* gene. **b** Differential expression and domain structure of the different gene products, including unique first exons. Dp71 and Dp40 share a common promoter and first exon, Dp40 is produced through the use of an alternative polyadenylation site. Note that additional splice isoforms of Dp427p, Dp260, Dp140 and Dp71 are not shown. WW: WW domain; Cys: cysteine rich domain; CT: C-terminal domain



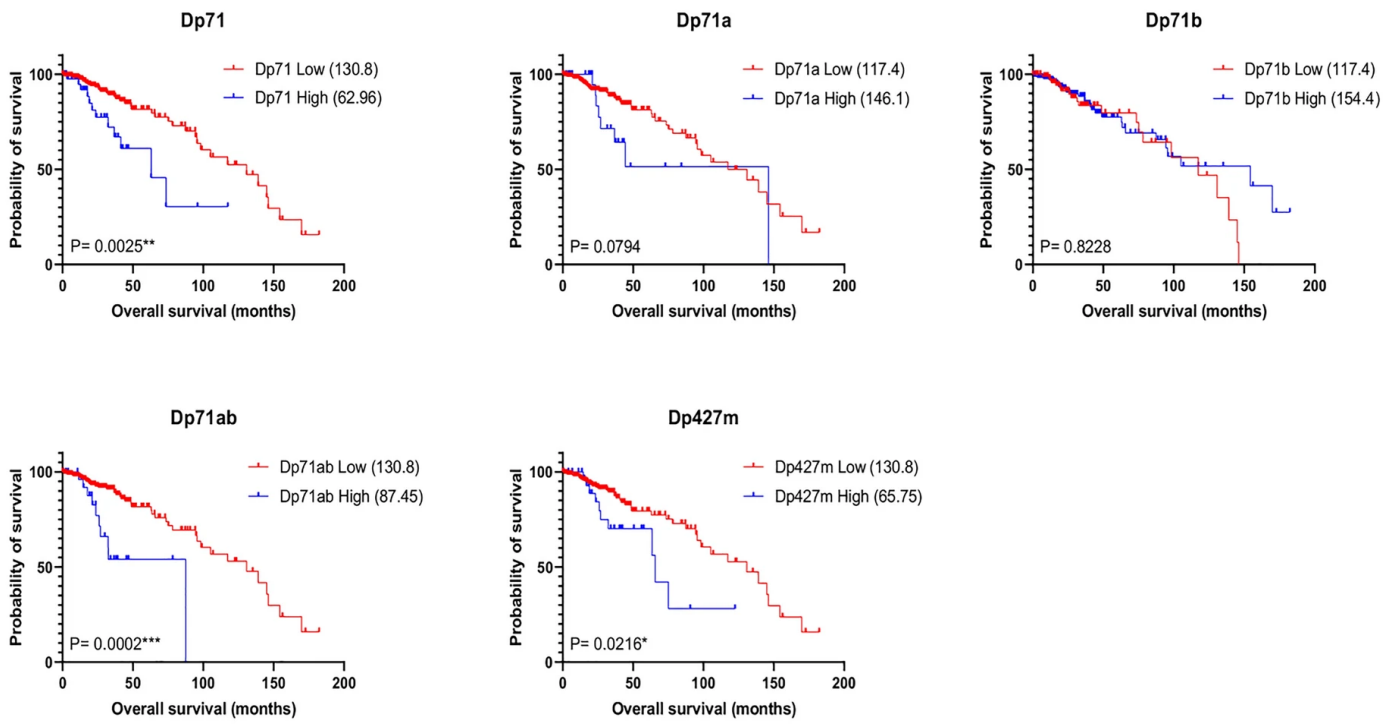
University of Northampton

The Dp71 protein

- Dp71 is a ubiquitously expressed protein product of the *DMD* gene
- Predominant *DMD* product in the brain
- It has 14 known spliced isoforms that have varying functions within the cell
- Cytoplasm, nucleus or both
- Potential roles in cell adhesion, cell division and nuclear structure
- Cancer cell line studies have found associations with high and low expression

The association of *DMD* gene expression with LGG survival is replicated by multiple *DMD* gene products

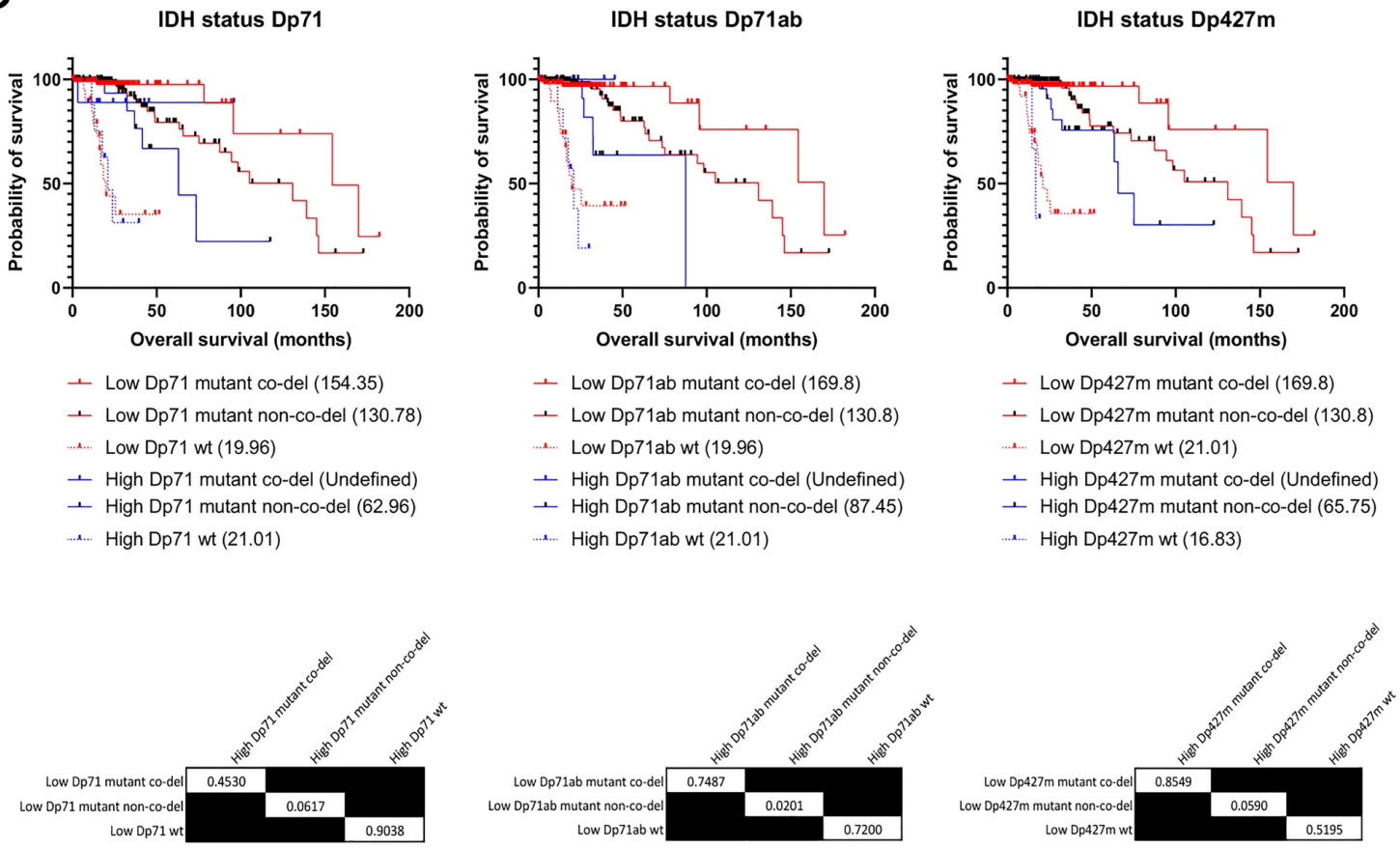
a



University of Northampton

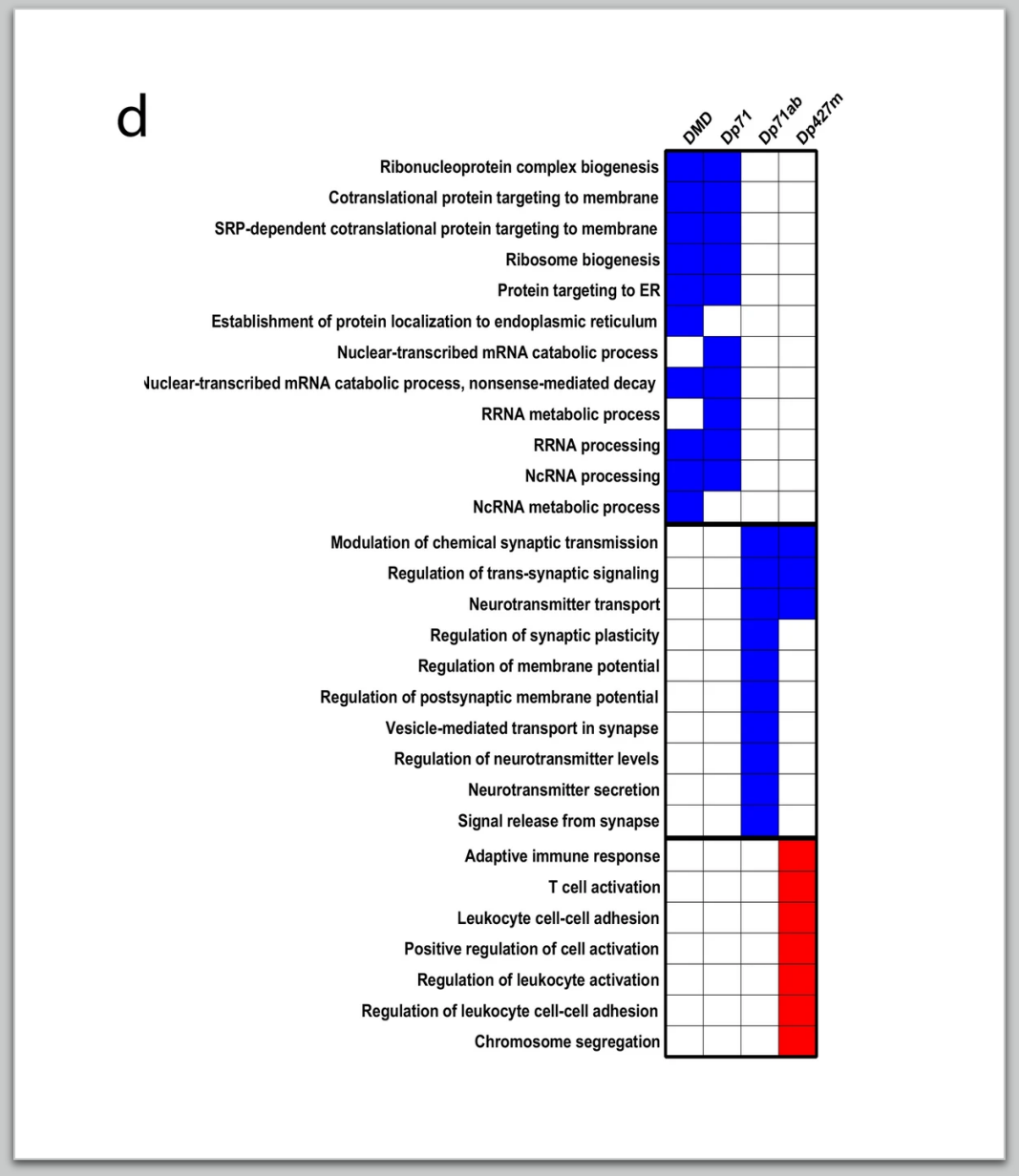
The association of *DMD* gene expression with LGG survival is replicated by multiple *DMD* gene products

b



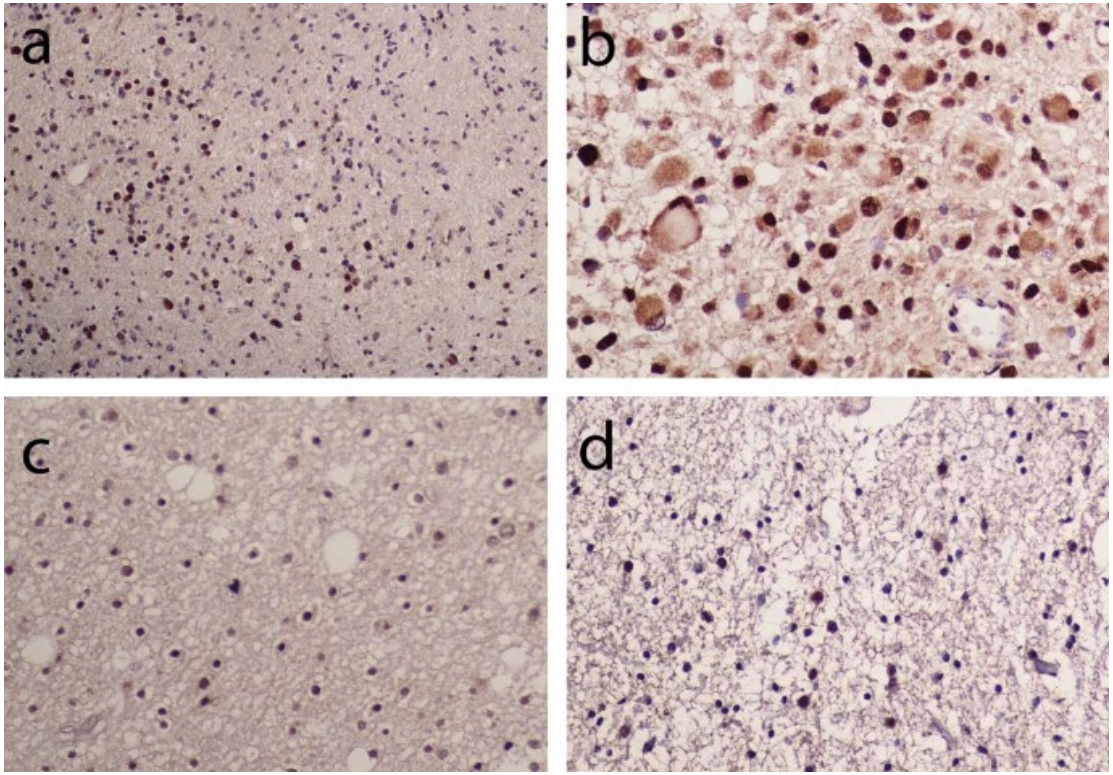
High expression of individual *DMD* gene products have differential biological effects relevant to the pathogenesis of LGG

- Pathway analysis using GO biological processes gene annotation.
- The heatmap was produced in GraphPad and displays the top 10 GO terms for high vs low *DMD*, Dp71, Dp71ab and Dp427m.
- Upregulated (red) and downregulated (blue) pathways are indicated.



University of Northampton

Dystrophin is expressed in the cytoplasm and nucleus of glial cells in LGG



Representative images of grade II astrocytoma (**a**), gemistocytic astrocytoma (**b**), oligodendroglioma (**c**) and NOS (**d**).

Images were taken at × 20 magnification except for image (**b**) which was taken at × 40 magnification to better visualise the gemistocytic staining.

Conclusion

We confirmed dystrophin protein is variably expressed in LGG tumour tissue by immunohistochemistry

We demonstrated that DMD expression has potential utility as an independent prognostic marker which can further stratify IDH mutant LGG to identify those at risk of poor survival.

This knowledge may improve risk stratification and management of LGG.

Other evidence for the role of DMD gene in cancer

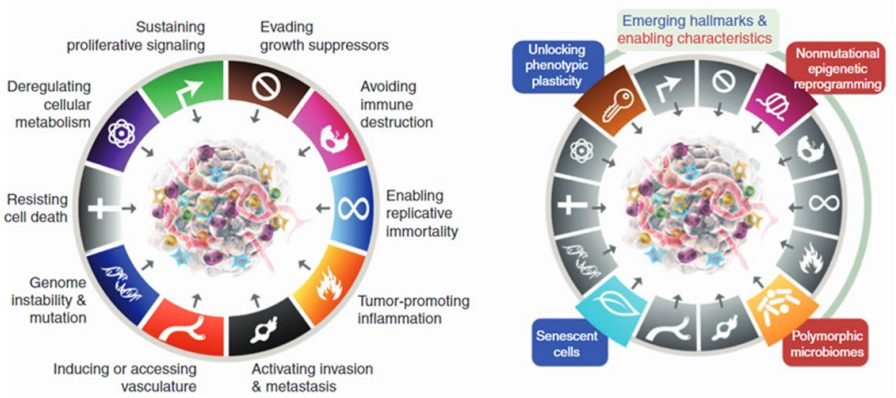
Table 2 Summary of evidence for a role of the *DMD* gene in cancer

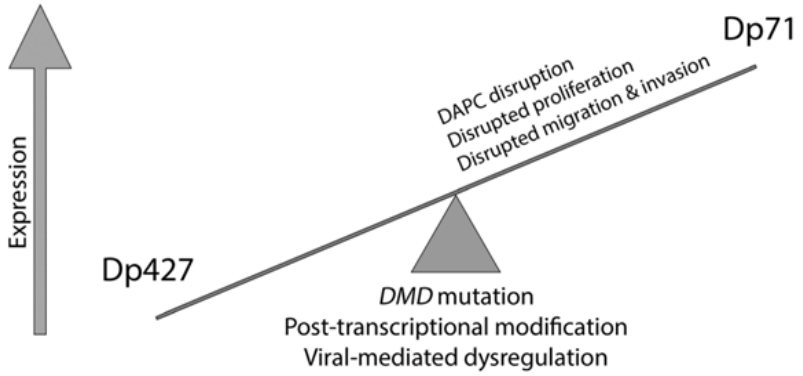
Tumour type	*DMD* mutation(s)	*DMD* expression	Oncogenic or tumour suppressive?	Strength of evidence	Ref(s)
STS	5' deletions affecting only Dp427. Intron retention.	Dp427 absent or severely reduced, Dp71 maintained	Tumour suppressive role for Dp427	+++	[18, 37, 38, 41]
Olfactory neuroblastoma	5' deletions	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

Jones, Leanne et al. "The Duchenne muscular dystrophy gene and cancer." *Cellular oncology (Dordrecht)* vol. 44,1 (2021): 19-32.

Proposed DMD-driven cancer development

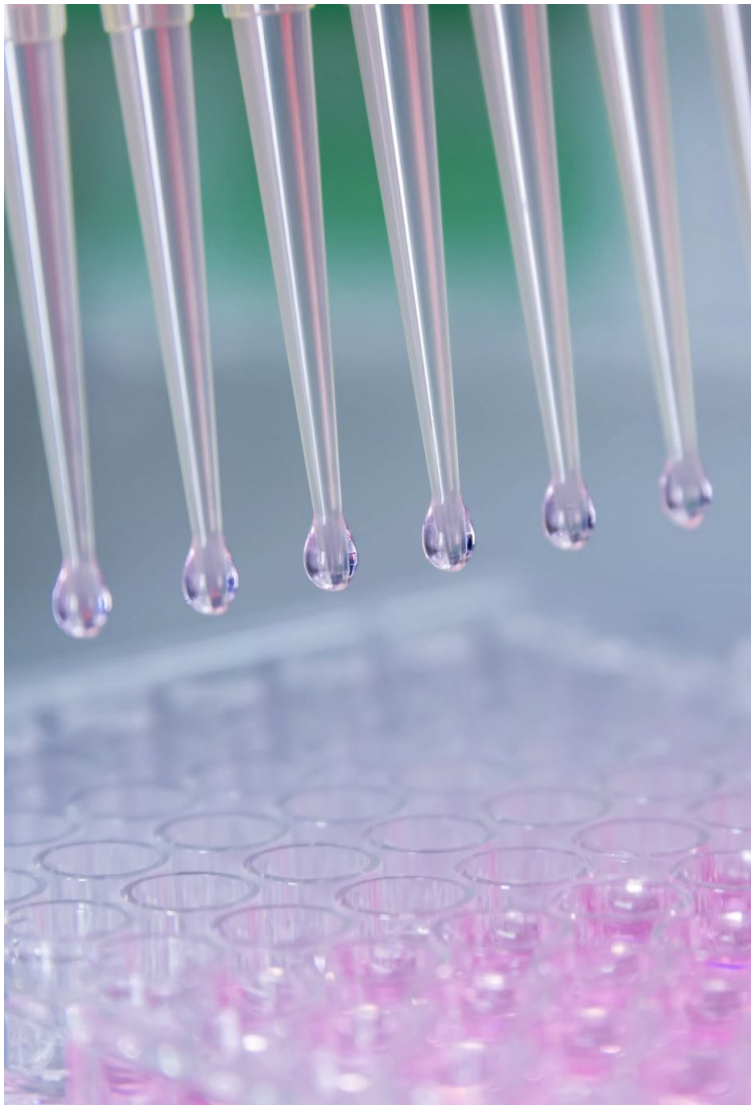




University of Northampton

What's next – wet lab ?

- Further analyse LLG patient brain tissues for Dystophin expression
- Source more glioma cell lines
- Overexpress Dp71
- Knock down the expression of Dp71

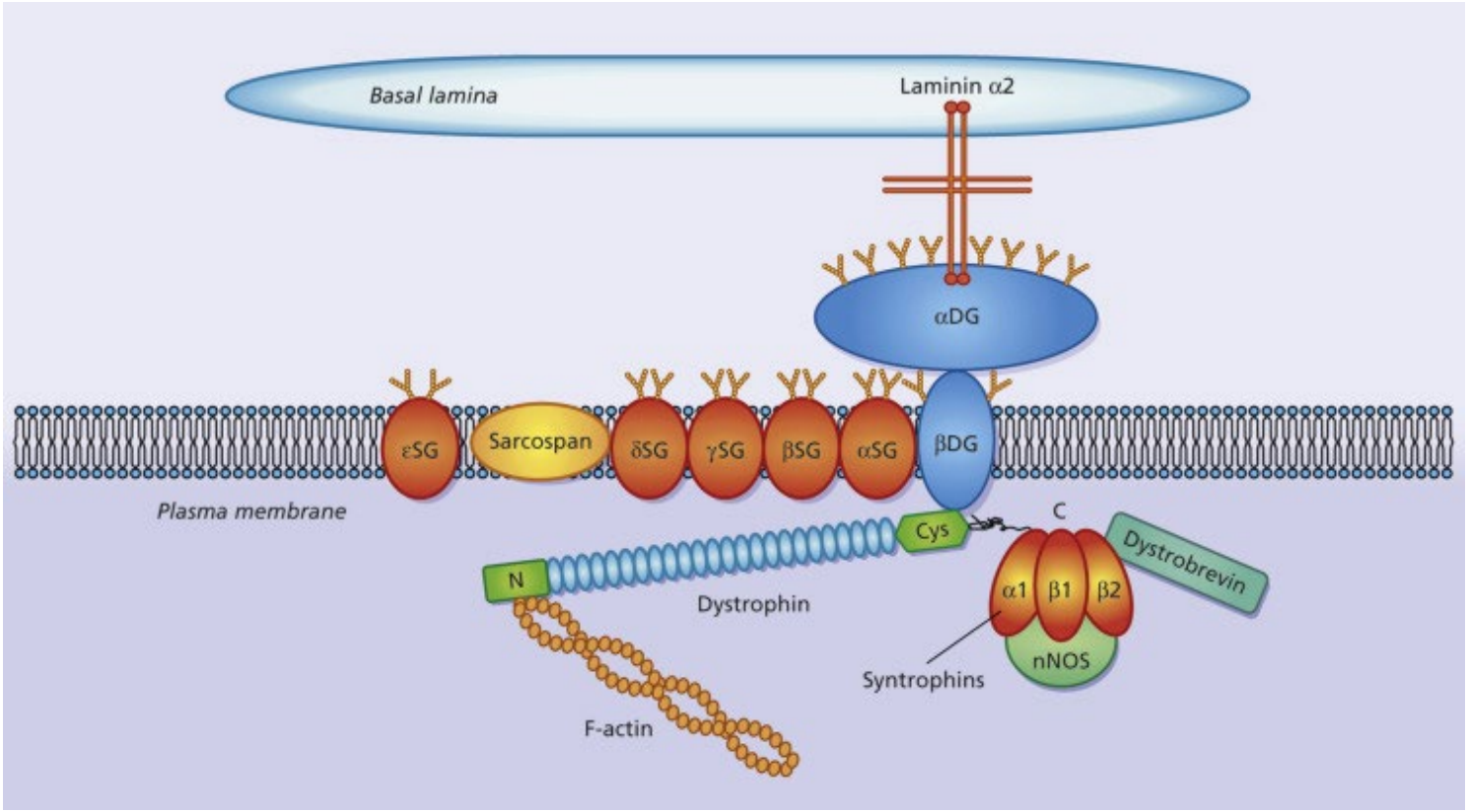


University of Northampton

- **Thanks & questions!**

Dystrophin in the Dystrophin associated protein complex (DAPC)

- Sarcolemma showing interaction of the dystrophin-associated protein complex with the actin cytoskeleton. (Note that the probable interaction of nNOS with the rod domain of dystrophin is not shown).



https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/dystrophin-associated-protein-complex