University of Southampton

Faculty of Medicine

School of Cancer Sciences

Identifying Novel Druggable Targets In Dedifferentiated Liposarcoma Using Biological Networks

Volume 1 of 1

Word count: 63,855

by

Ian Celyn Davies

BSc, MSc, PGCE

ORCiD: **0000-0003-2025-4561**

A Thesis for the degree of Doctor of Philosophy

08/01/2026

University of Southampton

Abstract

Faculty of Medicine

School of Cancer Sciences

Doctor of Philosophy

Identifying novel druggable targets in dedifferentiated liposarcoma using biological networks

By Ian Celyn Davies

Dedifferentiated liposarcoma (DDLPS) is a rare and aggressive adult soft tissue malignancy with limited treatment options and high relapse rates following surgical resection and radiotherapeutic interventions. DDLPS are characterised by amplifications of the *MDM2* and *CDK4* genes. Consequently, these have been targeted with small molecule inhibitors which have shown mixed results in clinical trials; DDLPS is in dire need of more targeted therapy options. This study applies a systems biology approach using gene co-expression network (GCN) analysis to identify novel therapeutic targets in DDLPS.

A GCN was constructed from DDLPS RNA-seq data (TCGA) and was sorted into modules using weighted gene co-expression network analysis (WGCNA) with optimised parameters. Modules of co-expressed genes were ranked by gene significance (GS) that describe disease characteristics, and sub-networks were inspected using a random walk with restart algorithm to identify hub genes. Integration with protein-protein interaction networks (PPIN) (STRING.db) and drug-targe databases (Therapeutic Target Database and Chemical Probes Portal) enabled drug identification.

UBE2C was identified as hub gene in the top-ranked module, with UBA1, acting upstream of UBE2C was found to be targeted by TAK-243 which is a small molecule inhibitor in phase 1 clinical trials. Beyond the identification of drug targets, it was identified that interferon signalling may contribute to a fibrotic tumour microenvironment (TME) and stromal heterogeneity through epigenetic mechanisms. Furthermore, vascular cells show gene expression patterns that indicate vascular mimicry and endothelial to mesenchymal transition. Lastly, enrichments for lipid metabolism, notably cholesterol efflux correlated to stem-cell like tumour features.

The integrative approach used here effectively identified genes associated with DDLPS biology. The ubiquitin-mediated proteasome was implicated through UBA1 targeting by TAK-243. Future work is needed to validate TAK-243 as a drug candidate in DDLPS.

Table of Contents

Table of Tables

Table of Figures

Research Thesis: Declaration of Authorship

Print name: IAN CELYN DAVIES

Title of thesis: **Identifying novel druggable targets in dedifferentiated liposarcoma using biological networks**

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission.

Signature:.. Date: ....

Acknowledgements

I would like to first thank my supervisors William Tapper, Zoë Walters, Stephen Thirdborough and Matthew Rose-Zerilli for their support, guidance and unwavering faith throughout the project. I have learnt much from them, even now as my life and career move in a new direction from when I first started this thesis. From the occasional lifting out of the matrix and into reality, to the shuddering of my frequent use of "tidy", and the use of whiteboard space for regurgitating networks, I thank you.

I am grateful to Sarcoma UK for providing the funding necessary to carry out this work. I would like to also extend my thanks to the entire Innovation for Translation Research Group (ITRG) and members of the Bioinformatics Club. I would like to offer thanks to the following individuals: Jack Harrington, Carmen Tse, Oliver Pickering, Christina Putnam, William Pratt, Suly Villa-Vasquez, Ben Sharpe, and Ian Redding, among others for their comradery, and for being proper tidy. I am grateful to the HPC community, particularly Alister Boags for putting up with my constant need for more storage resources and aiding in optimising script design for quicker computation.

I am eternally grateful, and awe struck over the support I have received from my family and friends. They know this journey has not been an easy one but have given me support that has given me an immense sense of pride in giving this my all. One, a very significant "one" I might add, my wife, Tansy, always the most wonderful flower in my eyes, through sickness and health she has been by my side. Without her encouragement I doubt I would not have had the strength to see this through to the end. *Diolch cariad*.

Lastly to my baby Anwen, tentatively named "bump", who has brought boundless joy. A perhaps entirely predictable transferable skill from the latter stages of this thesis is the completion of the practical module - "An introduction to sleep deprivation". This followed by my respectable repertoire of dad jokes, a substantial amount of which was obtained from my time at the ITRG, prepared me to be the best dad I can be.

Tansy and Anwen, I dedicate this to you.

Tansy ac Bump, ryw'n neilltuo hwn I chi

Definitions and Abbreviations

CDK4	Cylin Dependent Kinase 4
CDKi	Cyclin Dependent Kinase inhibitors
DDLPS	Dedifferentiated liposarcoma
DEG	Differentially Expressed Genes
DEP	DepMap Gene Effect Scores
DGEA	Differential Gene Expression Analysis
DTI	Drug-Target Interactions
DTN	Drug-Target Network
EC	Endothelial cells
EGN	Eigengene network
FE	Fisher's Exact
FRS2	Fibroblast Growth Factor Receptor Substrate 2
GCN	Gene Co-expression network
GEO	Gene Expression Omnibus
GO	Gene Ontology
GOBP	Geene Ontology Biological Processes
GS	Gene Significance
GSEA	Gene Set Enrichment Analysis
HMGA2	High Motility Hook Region A2
ICB	Immune Checkpoint Blockade
IFN	Interferon
IL	Interleukin
TEC	Tumour Endothelial Cell
IMC	Intramodular connectivity
kME	Gene module membership

Chapter 1

LASSO	Least Absolute Shrinkage and Selection Operator
LC	Louvain Community
UBE2C	Ubiquitin Conjugating Enzyme E2
LPS	Liposarcoma
MDM2	Mouse double minute 2 homolog
MDM2i	Mouse Double Minute 2 homolog inhibitor
ME	Module Eigengene
NCC	National Cancer Center Japan
NGS	Next Generation Sequencing
ORA	Overrepresentation Analysis
PH	Proportional Hazard
PPIN	Protein-Protein Interaction Network
PTPRN2	Protein Tyrosine Phosphatase Receptor Type N2
REAC	Reactome Pathways
RNA-seq	RNA sequencing
ROS	Reactive Oxygen Species
RWR	Random Walk with Restart
SCNA	Somatic Copy Number Alteration
scRNA-seq	Single Cell RNA-sequencing
SMI	Small Molecule Inhibitor
SNM	Sample Network Metrics
STS	Soft tissue sarcoma
TAM	Tumour Associated Macrophage
TCGA	The Cancer Genome Atlas
UBA1	Ubiquitin Activating Enzyme 1
UbiPS	Ubiquitin-proteasome system
WDLPS	Well-differentiated liposarcoma

WES	Whole-exome sequencing
WGCNA	Weighted Gene Co-expression Network Analysis
WGS	Whole genome sequencing

Chapter 1 Literature Review

1.1 Sarcomas

Sarcomas (meaning “fleshy lump”) are rare and heterogenous mesenchymal cancers, divided into two primary categories: bone sarcomas which arise in the skeleton and soft-tissue sarcomas (STS) arising in soft tissues across multiple body sites.[1,2] They are more prevalent in children, accounting for approximately ~10-15% of all paediatric cancers but are significantly rarer in adults, representing only ~1% of malignancies.[3]

The World Health Organisation (WHO) identifies over 130 sarcoma entities, with 44 considered malignant, 61 benign, 29 intermediate, and three classed as benign/intermediate.[1,2] Each of these malignancies are stratified into bone or soft tissue sarcoma (STS) and are further divided into sub-categories.[1,2] These categories are largely based on cytological and histological features according to the normal tissue they most closely resemble and/or arise in.[2,4] Although, this is not always the case, for example, synovial sarcomas do not arise in the synovium, they are thought to be of neuronal origin similar to malignant peripheral nerves sheath tumours (MPSNT).[5] STS is the most prevalent of the two major sarcoma types accounting for 90% of cases with the remainder of 10% involving bone.[6] STS sub-categories also contain heterogenous tumour subtypes which are largely based on further cytological, histological and adjunct genomic evidence.[7]

STS are split over eleven categories: (1) Adipocytic (malignant adipocytic tumours are liposarcoma), (2) Fibroblastic and Myofibroblastic, (3) Fibrohistiocytic (4) Vascular, (5) Pericytic, (6) Smooth muscle, (7) Skeletal Muscle, (8) Gastrointestinal stromal, (9) Chondro-osseous, (10) Peripheral Nerve Sheath Tumours, and (11) Uncertain differentiation. These tumours can occur anywhere in the body although there are preferential sites.

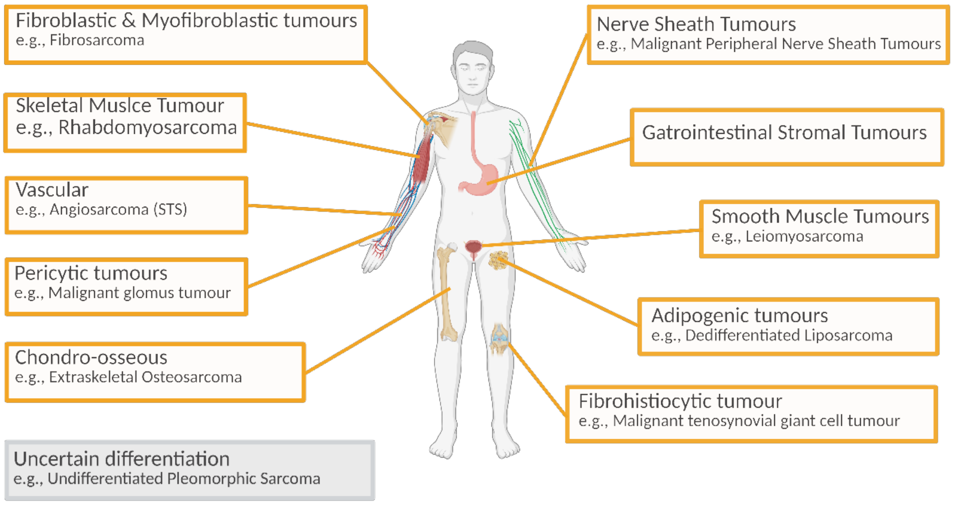

Fibroblastic & Myofibroblastic tumours
e.g., Fibrosarcoma
Skeletal Muslce Tumour
e.g., Rhabdomyosarcoma
Vascular
e.g., Angiosarcoma (STS)
Pericytic tumours
e.g., Malignant glomus tumour
Chondro-osseous
e.g., Extraskeletal Osteosarcoma
Uncertain differentiation
e.g., Undifferentiated Pleomorphic Sarcoma
Nerve Sheath Tumours
e.g., Malignant Peripheral Nerve Sheath Tumours
Gatrointestinal Stromal Tumours
Smooth Muscle Tumours
e.g., Leiomyosarcoma
Adipogenic tumours
e.g., Dedifferentiated Liposarcoma
Fibrohistiocytic tumour
e.g., Malignant tenosynovial giant cell tumour

Figure 1.1: Overview of human sarcoma categories by normal tissue type. Figure created in BioRender.com.

The National Cancer Registration and Analysis Service in England reported 4,500 STS diagnoses a year.[8] The five most common subtypes (***Figure 1.2A***) were found to be (in decreasing order) Gastrointestinal stromal tumours (GIST), Leiomyosarcoma (LMS), Liposarcoma (LPS), Undifferentiated sarcoma (US), and Myofibroblastic sarcomas (MFS). Similar results were found in national studies of other countries including the United States[9], Australian[10], and Europe.[11] In the UK – England the age standardised incidence rate (ASR) is 78.36 per one million for STS, rising slightly for males to 84.91 and decreasing for females at 71.80.[8] It is across studies generally regarded that LPS and LMS are consistently the most commonly diagnosed STS.

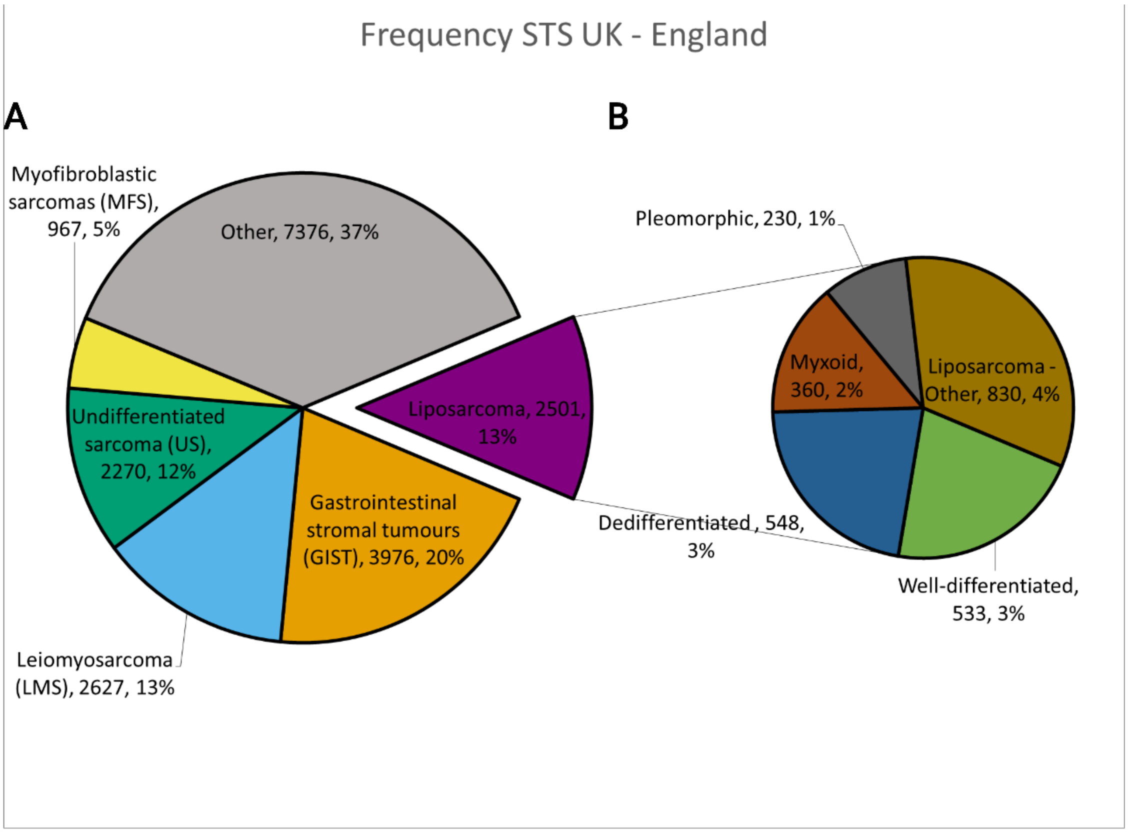


Figure 1.2: Pie charts of common STS tumours observed in a recent UK England study showing **A**: The percentage of the top five most common STS subtypes with total number of cases per subtype. **B:** The percentage and number of cases of WHO recognised LPS subtypes. Numbers of cases and percentages are those reported in Bacon et al and spanned four years (2013-2017).[8]

1.2 Liposarcoma

LPS are lipomatous STS malignancies that are categorised by the WHO into five main subtypes; well-differentiated liposarcoma/atypical lipomatous tumours (WDLPS/ALT – referred to hence force as WDLPS unless specified), dedifferentiated liposarcoma (DDLPS), myxoid liposarcoma (MLPS), pleomorphic liposarcoma (PLPS) and myxoid-pleomorphic liposarcoma (MPLPS).[1] Further heterogeneity arises from the presence of distinct histological variants. For instance, WDLPS presents across four recognised variants – lipoma-like, sclerosing, inflammatory and spindle cell whereas MLPS presents across two – the myxoid and round cell.[12,13]

1.2.1 Epidemiology

LPS accounts for around 20% of all STS cases.[8,10,14-19] The most common liposarcoma subtype is WDLPS occurring approximately 35-40% of the time, followed by MLPS (including the high-grade round cell variant) at 25-30%, DDLPS at 15-20%, PLPS at 5-10% with MPLPS being exceedingly rare.[18,20-25] In the recent UK – England STS population study, DDLPS was the most common (548 cases, 33%), followed by WDLPS (533 cases, 32%), then MLPS (360 cases, 21%) and then PLPS (230 cases, 14%) (***Figure 1.2B***).[8] It is generally regarded that WD/DDLPS tumour entities when taken together account for nearly 10% of all adult sarcomas occurring in extremities and truncal regions.[26] They most commonly occur in adults and very rarely (0.7%) occur in paediatric patients (<16 years old).[27] Some literature reviews also refer to WDLPS/DDLPS accounting for nearly three-quarters of LPS cases, although this may be an over-estimate.[23] LPS usually occurs in the aging population, with a peak incidence in the sixth or seventh decades.[14-16] A younger onset is noted in MLPS where incidence of those aged ≥50 and those aged ≤ 50 is near equal whereas in the other LPS subtypes, 80% of patients are aged ≥ 50.[28]

1.2.2 Clinical characteristics

LPS typically present as a large (>5cm) painless mass which is investigated by ultrasound and/or a magnetic resonance imaging (MRI)/computed tomography (CT) scan for tumours in deep-seated tissue.[3,29] Guidelines typically recommend that any lump >5cm and/or is increasing in size are to be treated as suspected malignancies until a benign diagnosis can be proven.[29]

LPS can occur anywhere in the body where there is soft-tissue mass where 40% of primary LPS are extremital, 20% are retroperitoneal and 40% occur in other sites.[16,18,30,31] The thigh (~60% of cases) is the most common extremital site of LPS tumours.[30] LPS is the most commonly occurring retroperitoneal sarcoma.[24,32] Although LPS tumours can present in any anatomical location, LPS subtypes occur at preferential anatomical sites.

MLPS and PLPS are predominant in the extremity. Primary MLPS commonly occurs in the low extremity 40% of the time where it is rare for MLPS to be found in retroperitoneal sites in 2% of cases.[30,33] PLPS tumours occur in the extremity 53% of the time, other sites include the retroperitoneum/pelvic cavity in 13% of cases and thoracic cavity in 6%.[34] Tumours occurred in diverse anatomical sites including breast tissue, stomach, uterus, groin, neck, and scrotum.[34] WDLPS commonly occurs at extremital, truncal and retroperitoneal/intra-abdominal sites, and infrequently in the thorax and head and neck.[35] They are generally considered to be equally distributed among retroperitoneal and extremital sites, although a recent analysis indicates that WDLPS is more common in the extremital sites in ~36% of cases, followed by the trunk in 30%, and then the retroperitoneum in 20%.[28] The most common site for DDLPS is the retroperitoneum or abdomen in ~60% of cases, followed by pelvic in 17%, extremity in 14%, trunk and thoracic in 6%, and rarely in the head and neck regions.[36] When located on an extremity or at the trunk, WDLPS are often called atypical lipomatous tumour (ALT), due to marginal resection often being curative.[2,35,37,38]

Histological and morphological review is conducted for subtyping. Fundamentally the approach to this is based on tumour histomorphology and immunoreactivity features.[2,39,40] This is conducted on samples that are typically retrieved via core-needle biopsy, unless the tumour is small in which case an excision biopsy is considered.[3,41] Samples obtained from biopsy are then pathologically reviewed for a histological diagnosis which involves a microscopic assessment of cell cytology looking for the shape, growth patterns, background and vascular structures within the cell.[42]

LPS exhibits great heterogeneity in histology showing biphasic or mixed phenotypes, so much so that it is very rare to identify a consistent singular (or “pure”) cell morphology or histological pattern in the same tumour.[43,44] In MLPS there is a myxoid (mucus/gelatinous) extracellular matrix (ECM) with hypocellular spindle cells, with pools of mucin.[39,45,46] MLPS can contain hypercellular round cells where if there is >5% coverage then it is classed as high-grade MLPS (previously referred to as round cell).[46-48]

WDLPS is histologically the best defined LPS tumour subtype. Its most typical (lipoma-like/adipocytic) presentation is characterised by mature adipocytes that display both typical and atypical features, including variable cell and nucleus sizes.[44] There are three other histological variants; sclerosing, inflammatory, and spindle cell.[12,49] The main differences between these are the phenotypes of stromal cells, density of collagen, prominence of primitive mesenchymal cells, degree of inflammatory infiltrates, and infrequently myxoid-like stroma.[12,44]

PLPS is identified by the presence of pleomorphic cells and lipoblasts, which are immature fat cells resembling normal preadipocytes.[50] PLPS exhibits a mixed histology with all cases being lipoblast-rich with notable histological pattern variation.[44,51] These patterns commonly resemble an UPS and sometimes with myxofibrosarcoma-like hypocellular regions of myxoid stroma.[51]

DDLPS is defined as a biphasic tumour containing undifferentiated pleomorphic or spindle cell regions juxtaposed to lipogenic regions of WDLPS. The non-lipogenic portion is typically high grade with morphology that is like PLPS and UPS.[42,44] These cells lack features of differentiated cells resembling primitive mesenchymal cells. The transition between these morphological features is most frequently abrupt but can also, in rare instances, be intermingled.[4]

In DDLPS it is generally considered that the majority, often cited as 90%, of DDLPS arise *de novo* defined by the observation as an occurrence of DDLPS within an ALT/WDLPS tumour where there is no known previous event of ALT/WDLPS or DDLPS in the patients clinical history.[4] [4] Such DDLPS are typically regarded as primary DDLPS. The remaining DDLPS occur upon a local recurrence of ALT/WDLPS where the observed DDLPS component was not present (or discovered) in the original ALT/WDLPS tumour with a mean interval of occurrence of 7.7 years.[4] Secondary differentiation occurs more frequently in repeated local recurrences of ALT/WDLPS. It has been reported that up to 28%[52] occur in the first recurrence of WDLPS, rising to 44% on the second recurrence.[53] It is unknown whether the risk of secondary DDLPS increases due to the number of local recurrences or the intervals between them.[54]

Dedifferentiation has been applied to several sarcomas that show biphasic morphology where histologically well-defined regions are adjacent to undifferentiated zones that show cellular, pleomorphic morphology.[43] Differing from undifferentiated tumours with the criteria of typically possessing differentiated (well-defined) elements in the same tumour.[43] The term of dedifferentiation was applied to LPS in 1979 to describe DDLPS.[43] Dedifferentiation has typically described entities that are high grade (according to the grading criteria) although low grade DDLPS has also been described.[4]

Diagnosing DDLPS can be challenging, as biopsy samples may miss the dedifferentiated non-lipogenic components and may be mistaken for WDLPS or normal adipose tissue. [3,55] Increasing the sampling rate during biopsies can partially mitigate these diagnostic limitations.[3,55]

In rare instances, DDLPS exhibits homologous lipogenic differentiation where lipoblasts are interspersed within the high-grade component.[56,57] Conversely, heterologous differentiation is seen in 5-10% of cases, most often in the form of myogenic differentiation (e.g., being leiomyosarcoma-like or rhabdomyosarcoma-like), or less frequently, as osteogenic features resembling an osteosarcoma.[44,58-60] Both DDLPS and PLPS present with increased mitotic activity, with PLPS exhibiting the highest levels. Both tumours also display moderate-to-extensive necrosis.[44]

Molecular testing primarily through immunohistochemical (IHC) staining for immunoreactivity of proteins pertinent to histological subtypes are also commonly leveraged. Adjunct analysis may also be used which can include the results from next-generation sequencing (NGS) or fluorescent in-situ hybridisation.[3,29,61]

Notable diagnostic molecular testing for LPS primarily includes the identification of supernumerary ring or giant rod marker chromosomes in WD/DDLPS.[37,62-64] These extra chromosomal structures include repeated amplifications of chromosomal region 12q13-15.[37,62-64] Notable genes within this region for WD/DDLPS are *mouse double-minute 2 homolog (MDM2),* and *cyclin dependent kinase 4 (CDK4)*. These can be identified upon karyotyping of the tumour or through positive stains during IHC assessment.[29] MDM2 and CDK4 staining along with cycling-dependent kinase inhibitor (P16 - CDKN2) are typically used to differentiate WD/DDLPS from benign lesions and other malignancies.[65,66]

Grading of LPS is conducted according to the Federation Nationale des Centres de Lutte Contre Le Cancer (FNCLCC) “French” grading system based on mitotic rate, necrosis, and differentiation status.[3,61,67] DDLPS and PLPS are typically assigned higher grades compared to WDLPS and MLPS due to their dedifferentiation score, which generally receives a score of 3. Consequently, DDLPS and PLPS tumours are most often classified grades 2 or 3.[22] However they are sometimes assigned a grade 1 if low scores are also given in the mitotic rate and necrosis categories.[43] WDLPS and MLPS show less abrogated normal morphology and hence are typically assigned lower grade values.[22] The majority of LPS are assigned a low-grade, likely due to ALT/WDLPS being the most common subtype which in the case of ALT are invariably classified as FNCLCC grade 1.[14]

The tumour grade with the size of the tumour, anatomical site of the tumour, metastasis and nodal involvement is then used to provide a tumour stage according to the American Joint Committee on Cancer (AJCC) staging system.[3,29] A separate staging system is used according to different anatomical sites. Tumour grade and size is used to classify tumours into stages I-III, with any metastasis a tumour is assigned to stage IV.

1.2.3 LPS General Treatment Modalities & Surveillance

Treatment for LPS follows the guidance as set out for STS and is largely divided into two main routes, those that are resectable and can be surgically removed, and those that are unresectable or are Stage IV (metastatic) classified as advanced disease.[3,41,68] The mainstay treatment is surgical resection particularly for localised disease where the standard procedure is an en-bloc resection with the aim to achieve tumour negative margins, removing the tumour entirely leaving a normal tissue boundary around the tumour.[3,29] It is typical to see surgical resection in up to 90-95% of cases.[69] Generally, resection with wide margins for low-grade and less aggressive STS can achieve excellent rates of local disease control (>90%).[70] However, large and/or deep-seated tumours, that are higher-grade, can make achieving a tumour-free margin challenging due to size and proximity to organs, which may require consideration for limb/organ removal (to retain tumour negative margins) or use of incomplete resection margins around the tumour.[3,71]

Tumour margins are commonly classified according to the R classification system corresponding to the presence of tumour at the resection boundary; completely resected tumour with tumour negative margins (R0) or tumour positive on microscopic inspection (R1), and tumour positive present at the resection margin in macroscopic inspection (R2). Cases that cannot be classified are typically assigned the Rx category.[69] As high as 62% of resections require resection of neighbouring organs however, an R0 margin is achieved 51% of the time.[72] As explored in section **1.1.2.1.5.3** complete or incomplete resection margins are debated to be prognostic of patient survival and in locoregional or metastatic recurrence. Histology is the greatest determinant of post-surgery relapse in retroperitoneal STS where DDLPS was found to show a greater risk of recurrence compared to WDLPS. Furthermore, there was limited benefit for a second surgery upon recurrence.[73] In a study by Zhao et al[69] R0 margins were possible in over half of cases (54%), R1 in 35% and R2 in just 11%.

LPS typically shows moderate sensitivity to radiotherapy (RT) (depending on when it is given in the course of treatment) where MLPS (including the high-grade round cell variant) shows the highest sensitivity of all LPS subtypes.[74] MLPS benefits from RT irrespective of site of occurrence and whether it is given pre or post-operatively, being effective in reducing tumour size for surgical intervention.[7,75-77] Adjuvant RT for retroperitoneal LPS shows mixed results; it may reduce recurrence rates, although impact on survival is not clear. Neo-adjuvant RT is typically not beneficial to LPS although has been shown reduce tumour size for resection in MLPS. The STRASS trial found no overall benefit of pre-operative RT, but WDLPS showed improved recurrence-free survival (RFS).

For advanced/metastatic disease, treatment modalities are primarily to provide palliative resolution.[3,29] Surgical resection in this setting is not effective in controlling disease. In advanced DDLPS/WDLPS occurring in the intra-abdominal space the combination of doxorubicin and ifosfamide was shown to give the best overall response rate (ORR) and is currently the most widely adopted frontline treatment.[78] Second-line chemotherapy includes use of trabectedin, eribulin, along with gemcitabine plus docetaxel.[79] Combinations of chemotherapeutic drugs (e.g., doxorubicin plus ifosfamide or trabectedin) can achieve better progression free survival (PFS) versus single agent, however, this comes at the cost of increased toxicity. [79] The excision of tumours (primary, recurrent and metastatic) in the advanced setting is considered based on palliative benefit to the patient.[3,29,80,81]

LPS has a high risk of post-surgical recurrence, varying by subtype and location, where retroperitoneal and high-grade LPS show particularly poor RFS.[14,82-86] Clear surgical margins is the most important factor in reducing recurrence although the impact on overall survival is debated.[16,38,87-89] Overall, current treatments are inadequate for high-grade LPS, particularly DDLPS. Presently, there is no standard procedure for disease follow-up, where different centres have shown varying protocols on the intervals and duration of follow-up and whether imaging is used.[3,81]

1.2.4 Survival patterns and prognostic factors for LPS

The survival of LPS, unstratified by subtype, is good overall with a 80% (82% in a recent UK England study[8]) five-year (5yr) overall survival (OS – defined as the time survived from either diagnosis or treatment of cancer to death) and a ten-year OS of 65%.[16] This compares favourably to the average STS 5yr OS of 55%.[3] However, LPS consists of subtypes that show distinct clinicopathological behaviour.[16,90] The most common LPS subtypes (WDLPS and MLPS) have a 5yr OS of 80-90%.[18,90]

More aggressive LPS subtypes including high-grade MLPS, PLPS and DDLPS have significantly lower survival. High-grade MLPS has shown a 5yr DSS of 72%.[90] PLPS and DDLPS perform similarly poor, with a 5yr survival that is below 50%.[90] For these high-grade, whilst the 5yr metrics are still moderate, this rapidly declines for ten-year survival rates.[22]

For individual prognostication in the UK for STS, the Sarculator is a commonly used nomogram available as an online tool.[91] Factors that can predict poor outcome for patients with STS tumours are broadly well understood and include stage (along with its derivates of grade, size and anatomical location), histology, age and gender.[3,80,92] The highest determinant for poor outcome in LPS is tumour grade which also accounts for histological subtype, as a poorly differentiated subtypes are graded higher where predictably FNCLCC grade 3 tumours perform the worst (**section 1.1.2.1.5.1**).[14,69,93] Grade is also a determinant in tumour recurrence following surgical excision, where the resection margin is also well-documented as being a shared determinant.[53] Although other factors are also significant, notably age where an age ≥ 55 predicts recurrence within the first year of treatment in retroperitoneal LPS.[83]

Extremital sites in LPS show better survival outlook versus other sites.[94] This is observed across STS where extremital tumours often perform better than deep-seated and retroperitoneal tumours. Extremital tumours are commonly (25%) low-grade liposarcomas.[95] WDLPS tumours at retroperitoneal sites have a poorer outcome with lower rates of 5yr OS and disease specific survival (DSS).[35] In DDLPS tumours located at deep tissue sites such as the retroperitoneum tend to have a poorer prognosis.[4,14,16] WD/DDLPS tend to recur at greater frequencies when they originate in the retroperitoneum as compared to other sites.[37]

Patient age and sex are also typically found to be common variables that predict poorer patient outcome, where older patients and males associate with a poorer prognosis.[8,96,97] increasing age corresponds to a deterioration of survival probabilities.[28,69] There is no defined cut-off for age, but thresholds of ~≥50 years can be common.[16,28,69] Patients are noted to be younger in MLPS.[90] For PLPS, an age of ≥ 65 was associated with poorer OS but could not be reproduced when using DSS.[98] Other studies have shown that an age ≥ 35 is significantly associated with decreased OS in WDLPS and MLPS.[22] Both gender (male) and age (unclear definition) are noted to be important factors at time of diagnosis that predicts both a poor OS and RFS outcome.[8,22,24,99] DDLPS is predominantly observed in male patients (65%) and most frequently observed in older adults, where 65.4% of patients are aged between 51 and 75 years of age.[36] Better outcomes were observed in patients aged between 26 to 50 years, and in tumours that were smaller than 10cm.[36]

Whilst initially thought to be equal in gender distribution, in larger studies LPS is shown to occur 60% in male and have a poorer prognosis where notably DDLPS has been shown to have a male-female ratio of 2:1.[22,28,100] For retroperitoneal LPS, 5yr OS rates for males was found to be 51% increasing to 60% for females following a primary resection.[101] Female patients in intermediate to high-grade tumours are found to have poorer survival compared to males.[102]

Locoregional recurrence and the propensity to metastasis to other tissues are also factors that can significantly impact the survival in LPS subtypes, although WDLPS does not metastasis.[3,103-106] local recurrence, tumour grade, resection margin, and tumour size are important risk factors for metastatic recurrence.[107] Subtype histology also plays a role in MLPS where high-grade (round cell) morphology predicts a greater risk of metastasis.[47] In DDLPS local recurrence is the most consistent risk factor for tumour metastasis.[107,108] Metastasis occurs in 30% of DDLPS patients with 20.5% of these being present at diagnosis.[36,104,108] The three most common metastatic sites being the lung (75%), subcutaneous/intramuscular tissue (52%) and lymph nodes (34%).[108] The median time to metastasis for patients progressing from localised to metastatic disease is 8 months.[108]

Tumour size in WDLPS provides a greater risk of poor outcome for DSS, OS and relapse-free survival (RFS) metrics.[49] WDLPS can be locally aggressive and show moderately high local recurrence (up to 52%) particularly at deep-tissue sites such as the retroperitoneum where recurrence is more common.[23,88,109]

The level of dedifferentiation in DDLPS has mixed prognostic significance. While a lower level of dedifferentiation was not found to be prognostic of better outcome.[4] In addition, the threshold by which dedifferentiation becomes prognostic of patient outcome is presently unknown. However, it has been shown that increasing degrees of dedifferentiated morphology with high-grading is prognostic to high local recurrence risk.[110]

DDLPS is in dire need of improved treatment strategies with unmet clinical need.

1.2.5 Genomic landscape of DDLPS/WDLPS

DDLPS and WDLPS are cytogenetically characterised by the presence of supernumerary ring chromosomes or giant marker chromosomes which frequently contain amplifications involving chromosome region 12q13-15.[37,62-64] Somatic copy number alterations (SCNA) dominate the mutational landscape observed in WD/DDLPS.[37,64] It is noted that DDLPS and WDLPS, like other STS tumours show low mutational burden in regard to somatic mutations.[64] Amplification in mouse double-minute 2 homolog *(MDM2)*, cyclin dependent kinase 4 (*CDK4)* and high-mobility group AT-hook 2 (*HMGA2)* are the most discussed frequent amplifications in DDLPS through increased copies of 12q13-15.[111-113] Other frequently amplified genes within 12q-13-15 have been observed in *carboxypeptidase M (CPM), fibroblast growth factor receptor* *substrate 2* (*FRS2*), *tetraspanin 31 (TSPAN31),* and *YEATS domain containing 4* (*YEATS4*).[37,114,115] The shared genomic alterations between WDLPS and DDLPS are thought to be the genomic events which can drive tumorigenesis.[114]

MDM2 is an E3 ubiquitin ligase that negatively regulates TP53 activity by targeting it for proteasomal degradation, preventing interactions with its transcriptional co-activators, and exporting TP53 out of the nucleus.[116,117] The TP53 protein plays a critical role in cell maintaining genomic stability through cell cycle control and DNA repair, and is a known tumour suppressor.[118] Dysfunction of this pathway has been implicated in most cancers, either through *TP53* mutation or downregulation.[119] In WD/DDLPS, dysfunction of TP53 is predominantly through MDM2 amplification.[120] *MDM2* (12q15) is amplified and expressed in 100% of WD/DDLPS tumours, and is considered the main driver of tumorigenesis in these tumours.[37,113,116] Levels of MDM2 amplification follow a log-normal distribution and the level of MDM2 amplification has been found to predict poor outcome in DDLPS.[121]

HMGA2 is a regulator of multiple core molecular processes such as replication, recombination, DNA repair and transcription.[122] Most studies suggest an oncogenic role for HMGA2.[122] However, the role of HMGA2 as an oncogene maybe context specific as it has been suggested that HMGA2 may contribute to repression of proliferation-associated genes to promote senescence in fibroblasts.[123,124] HMGA2 promotes adipogenesis and mesenchymal differentiation, and hence an inference maybe that in DDLPS there should be lower levels of HMGA2 activity (being antagonistic to dedifferentiation).[125,126] However, counterintuitively, *HMGA2* shows higher levels of non-damaging genomic rearrangements and overexpression in DDLPS compared to WDLPS.[113,125] It has been suggested that a high log gain ratio (>2) of MDM2 and HMGA2, displayed prognostic value, where the higher ratios were indicative of a shorter OS.[124]

CPM may have a role in adipogenesis, as mRNA expression is increased in pre-adipocyte differentiation.[127,128] However, CPM is over-expressed in both DDLPS and WDLPS samples compared to benign lipoma and normal fat tissue.[129] Additionally, *CPM* gene knockouts in two DDLPS cell lines (LPS141 and FU-DDLS-1) inhibited cell proliferation.[129] CPM has been shown to increase epidermal growth factor receptor (EGFR) signalling.[129] Dysregulated EGFR signalling is associated with many cancers, including breast and lung, with downstream signalling oncogenic signalling pathways.[130] These findings may suggest an oncogenic role for CPM through EGFR signalling which was associated with increased clinical stage and histological grade in STS.[131]

YEATS4, which has been suggested to be involved in repression of TP53. This prevents activation of two genes involved in apoptosis and cell cycle regulation (*CDKN2A (p14)* and

CDKN1A (p21)).[113,132,133] YEATS4 in DDLPS may work to support MDM2 in its TP53 repressive activities.[134,135] Although not statistically significant, *YEATS4* has been found to show a trend towards decreased RFS in DDLPS.[136]

FRS2 encodes a central adaptor protein within the fibroblast growth factor receptor (FGFR) pathway, necessary for signal transduction.[137] Aberrant signalling of the FGFR signalling pathway is known to have a role in angiogenesis, tumorigenesis and metastasis.[137,138] It has been found that in WDLPS/DDLPS the FRS2 protein is overexpressed with higher phosphorylation suggesting that signal transduction by FRS2 is active, and hence is signalling through the FGFR pathway.[115,137]

CDK4 is thought play a major role in DDLPS tumorigenesis and has been associated with poor disease prognosis.[126,136,139] CDK4 phosphorylates Retinoblastoma protein 1 (RB1), a tumour suppressor that regulates cell cycle processes through inactivating the transcription factor E2F.[140] Upon phosphorylation, the Rb-mediated inhibition of E2F is stopped, and E2F can continue to commit the cell to the G1/S transition of the cell cycle.[140]

DDLPS were shown to have higher mean ratios of amplification of *TSPAN31* compared to WDLPS.[141] TPSAN31 is the natural anti-sense transcript to CDK4, and has been shown to regulate its mRNA and protein expression, where silencing of *TSPAN31* increased *CDK4* mRNA and protein expression in hepatocellular carcinoma cells.[142] The role of TSPAN31 in DDLPS and its interactions with CDK4 in this context is unknown.

DDLPS is thought to diverge from WDLPS in early tumour progenitor cells.[143] The exact mechanism by which this occurs is unknown but it is thought that augmenting of the existing mutational landscape, and the acquisition of new SCNAs plays a crucial role.[114,144] The gain of 1p32.1 and 6q23, and loss of 11q23 regions have been described as being enriched alterations in DDLPS.[145,146] Of note, 1p32.1 is correlated with poor disease specific survival and contains the *JUN* oncogene.[64,114,126] In addition, *JUN* amplifications have been found to lead to downstream blocking of adipogenic differentiation.[126,147,148] The protein product of *JUN*, c-Jun forms the Activator Protein-1 (AP-1) transcription factor complex together with the c-Fos protein.[149] AP-1 is involved in many cellular processes including apoptosis and cell proliferation.[147] In DDLPS, c-Jun expression was detected 91% of the time compared to just 27% in WDLPS with no histological signs of DDLPS regions.[150]

The amplification of 6q23, includes the Mitogen-activated protein 3 kinase 5 (*MAP3K5)* gene (also known as ASK1), which activates Jun N-terminal kinase (JNK) and p38 mitogen-activated protein kinases in the mitogen-activated protein kinase pathway.[151] MAP3K5 is activated in response to cellular stress such reactive oxygen species (ROS) and inflammatory cytokines.[152] Activation of JNK leads to downstream activation of JUN and inactivation of peroxisome Proliferator-activated receptor gamma (PPAR-γ) which is required for adipocyte differentiation.[147,153,154] In absence of PPAR-γ expression, the cell line LPS141 could not be forced to differentiate when cultured in differentiation media.[148] This implicates c-Jun and MAP3K5 in the process of blocking adipocytic differentiation.[114] *JUN* and *MAP3K5* both show an amplification frequency in DDPLPS of 32% and 22% respectively.[64,155]

The low frequency of exclusive SCNAs may suggest other aberrant factors are likely involved in the dysregulation of adipocytic differentiation in DDLPS. The core region of 12q13-15 is pivotal to the formation of the ring structure observed and the neo chromosome is often the largest chromosome in WD/DDLPS tumour cells.[156,157] These chromosomal structures are formed from an initial chromosomal breakage and edge fusion to form a ring which undergoes a series of break-fusion-bridge cycles where the ring is replicated and fused across multiple cycles to produce the number of amplifications observed.[158] It is likely that these vast amounts of amplifications of oncogenes promotes the survival of WD/DDLPS initiating cells.[156,157]

To support this observation evidence for an exact process by which the adipogenic signature in WDLPS could be reversed were inspected. This was largely driven by the observation of *JUN/ASK1* amplifications although evidence was conflicting on their definitive roles in DDLPS, particularly where only a proportion of cases contained such amplifications.[147,148,150,151]

An integrated transcriptomic and genomic study identified further alterations (fusions and amplifications) that were hypothesised to correspond to a malignant transformation of WDLPS into DDLPS.[114] It is also hypothesised that DDLPS and WDLPS share a progenitor cell that initiates tumorigenesis containing the characteristic amplifications which then diverge into WD and DD progenitor cells.[143,159] A recent single-cell RNA-sequencing study highlighted molecular differences between a population of possible progenitor cells identified in WD and DD components.[143] DDLPS stem cells showed similarity to multipotent adipocytic progenitors and tumour stem cells containing 12q amplifications were conserved in both WD and DD components. Furthermore, bespoke alterations were identified between populations identified in WD and DD components.[143] DDLPS retaining its stem cell like properties is consistent with microscopic observations and the ability of DDLPS to undergo heterologous differentiation (e.g., to differentiate along osteoblastic lineages to resemble an osteosarcoma).[58,143] It was noted that DD components contain high-levels of *Transforming growth factor β* (*TGFβ*) and was hypothesised to be a contributing factor to the formation of DDLPS tumour cells.[143]

1.2.6 The emergence of targeted therapies for LPS

Presently, there are no FDA approved targeted therapies available for treatment of DDLPS, although, there are promising treatments that are in active phase2/3 clinical trials (**table 1.1.**). The most promising targeted therapies to date have targeted MDM2. Inhibitors against MDM2 include milademetan (DS-3032b), alrizomadlin, and brigimdalin (BI 907828). Milademetan failed to meet its primary endpoint of significantly improved PFS over trabecedtin in a phase 3 clinical trial (MANTRA – NCT04979442) in unresectable/metastatic DDLPS that recruited 175 patients.[160] This was following a successful phase 1 dose-escalation trial where a partial response (PR) was achieved in 4% of patients and stable disease in 64.2%.[161,162] Alrizomadlin (APG-115) in combination with pembrolizumab (a Programmed cell death protein - PD-1 - inhibitor) showed a PR in ~6% of LPS patient in interim results of a phase-2 trial (NCT03611868).[163] ALRN-6924, a dual MDM2/MDMX inhibitor, showed promise in a phase-1 trial (NCT02264613) showing a PR of 20% in LPS patients, and all showed stable disease.[164] Another dual MDM2/MDMX inhibitor siremadlin (HDM201) achieved PR in ~8% of TP53 wild-type LPS patients and 75% achieved stable disease in a phase 1 trial.[165] In some patients where there is TP53 dysfunction by damaging mutations MDM2/MDMX inhibitors may potentially not be effective.

Brigimadlin (BI 907828) is another MDM2 inhibitor that is currently undergoing a phase 2/3 clinical trial (Brightline-1 - NCT05218499) comparing use of brigimadlin with doxorubicin as a in first-line treatment of advanced DDLPS.[166] A phase 1a/1b (NCT03449381) dose-escalation study identified acceptable toxicity of brigimadlin and noted a 75% SD in DDLPS patients ranging from 1/5 to 22 months.[167]

Treatment with a CDK inhibitor, abemaciblib in a phase 2 clinical trial lead to an increase in progression-free survival (PFS) to – 76% at 12 weeks).[168] Following this success, abemaciblib is now undergoing a phase 3 trial (SARC041 - NCT04967521) in advanced DDLPS.[169] It was found that abemaciiblib induces tumour cell senescence, which corresponded to increased immune infiltration, but also may lead to downstream progression.[170] Palbociclib, another CDK4 inhibitor, showed a complete response in 2% of DD/WDLPS patients, and 12 week PFS was ~57%.[171] A phase II trial of ribociclib in combination with everolimus (NCT) an mTOR inhibitor, showed promise in DDLPS which attained a median PFS of 15.4 weeks, and 33% of patients showed SD or PR/CR at ≥ 16 weeks.[172] There is a phase 2 trial (NCT05694871) that is undergoing current recruitment for palbociclib combined with cemiplimab, a PD1 inhibitor, versus palbociclib alone that is currently recruiting for advanced DDLPS.

Table 1.1: Summary of small molecule inhibitors presently in active clinical trial for DDLPS.

Compound name (alias)	In combination with	Target (+ target of combination)	Recruiting trial phase	NCT number
Alrizomadlin (APG-115)	Pembrolizumab	MDM2 (+ PD1)	Phase 3	NCT03611868
Brigimadlin (BI 907828)	-	MDM2	Phase 2/3	NCT05218499
Palbociclib	Cemipolimab	CDK4 (+ PD1)	Phase 2	NCT05694871
Ribociclib	Everolimus	CDK4/6 (+ mTOR)	Phase 2	NCT03114527
Abemaciblib	-	CDK4/6	Phase 3	NCT04967521
Erdafitnib	-	Pan-FGFR	Phase 2	NCT03210714

Immunotherapies targeting PD-1 (programmed cell death protein 1) have also been assessed in DDLPS. Pembrolizumab showed a 10% overall response rate in the LPS cohort of 4 patients in the SARC028 phase-II trial LPS expansion cohort and achieved less success in a single-arm trial combining pembrolizumab with axitinib, a VEGF-inhibitor, where no response was observed in DDLPS patients.[173,174] Nivolumab alone or in combination with ipilimumab, a Cytotoxic T-lymphocyte-associated protein – 4 (CTLA-4) inhibitor showed no responses in DDLPS/WDLPS patients enlisted in the Alliance A091401 trial.[174] However, in a retrospective review of STS patients treated with nivolumab plus ipilimumab, a partial response was achieved in 17% DDLPS patients highlighting potential use as a combined therapy.[175] Despite advances in targeted and immunotherapy for DDLPS, there remains significant unmet clinical need for patients DDLPS.

It is also worth noting the MULTISARC (NCT03784014) phase 2/3 randomised clinical trial to evaluate the feasibility of NGS in advanced/metastatic STS which includes DDLPS.[176,177] It also seeks to determine whether NGS-guided therapy can improve patient survival outcomes. To do this it is applying RNA-seq and whole-exome sequencing to pair genomic alterations with candidate drugs in a tailored approach. These drugs include palbociclib, which given previous indications will likely be the tailored drug to DDLPS patients.

1.3 An introduction to networks:

Networks are methods of representing relationships in complex data as a connected graph.[178] Today, networks have found a common use in a variety of applications, perhaps most intuitively in modelling social interactions through online social media platforms.[179,180] Due to the vast rise of omic data within the field of molecular sciences since the late 90's, networks have found a place in being an intuitive method for interpreting such linear and non-linear relationships.[181-184] Many studies have utilised networks on cancer omics (molecular layers – e.g. transcriptomics) data derived from experimental assays or predictive modelling to identify prognostic biomarkers, cancer subtypes, cancer gene identification, and drug repurposing and discovery.[185-194] There are now a wide-range of databases and repositories which host pan-cancer molecular data, allowing for network-analysis to be applied to secondary data analysis (**table 1.3**).

Table 1.2: A summary of commonly utilised data repositories.

Repository name	Data types	Includes cancer	Includes LPS
The Cancer Genome Atlas (TCGA)	WES, RNA-seq, miRNA-seq, DNA methylation, protein expression, copy number arrays, clinical data	Yes	Yes
Gene Expression Omnibus (GEO)	Gene expression array and RNA-seq , DNA methylation	Yes	Yes
Array Express (AE)	RNA-seq, DNA methylation		
Clinical Proteomic Tumor Analysis (CPTAC)	Proteomics data on select TCGA projects	yes	No
International Cancer Genomics Consortium (ICGC)	WGS mutation data	Yes	Yes
DNA databank of Japan (DDBJ)	RNA-seq, gene expression array, WES, WGS, mutation calling,	Yes	Yes
Cancer Cell Line Encyopledia (CCLE)	gene expression arrays, copy number arrays, miRNA expression, protein expression, DNA methylation	Yes – Cell lines	Yes
Cancer Dependency MAP project (DepMap)	CCLE project with CRIPSR gene knockout cell viability data	Yes – Cell lines	Yes
Connectivity MAP	Drug-gene perturbation data (gene expression array, proteomics)	Yes – Cell lines	No

STRING	PPI	No	No
BioGrid	PPI	No	No
IntAct	PPI	No	No
JASPAR	TF-target	No	No
TRRUST	TF-target	No	No
miRTarBase	miRNA-target	No	No
Cansar.ai	Drug-target, drug combination data, ligandability	yes	No
The Drug Gene Interaction Database (DGIdb)	Drug-target	No	No
Drugbank	Drug-target	No	No
Human Metabolic Atlas (HMA)	Metabolic data	No	No
KEGG	Metabolic data	No	No
MsigDB	Gene function annotations	No	No

1.3.1 Fundamental network concepts

1.3.1.1 Nodes, edges and connectivity

Network graphs provide a simple yet powerful tool for understanding the relationships between components of a complex system. They are composed of nodes and their connections called edges (**Figure 1.3A**). In networks there are two main edge types, undirected and directed.[195] Undirected edges represent a simple connection between pairs of nodes and convey no information regarding directional flow. A directed edge implies information flow that can be organised hierarchically (e.g., a transcription factor (TF) that activates a gene). For both directed and undirected networks, edges can be assigned numeric values or "weights" (e.g. based on pairwise correlations among genes). Weighted networks depict a strength of connection according to the weight value.[196] There can also be more than one edges between the same nodes either representing inverse directionality (e.g., a reversible interaction) or depict different edge types which is sometimes referred to as multiplexing or stacking.[197] Similarly, networks can have two, or more, node types which are bipartite graphs (e.g., a network containing drugs and proteins).

Networks can be expressed mathematically as an adjacency matrix where all rows and columns represent the set of nodes belonging to that network (**Figure 1.3B).**[198,199] In unweighted networks, which is depicted in ***Figure 1.3B***, the cells can have a value of 1 or 0 according to whether an edge exists or not between two nodes, respectively.[199] In undirected networks the values on both sides of the main diagonal are symmetrical (e.g., a gene-gene correlation)

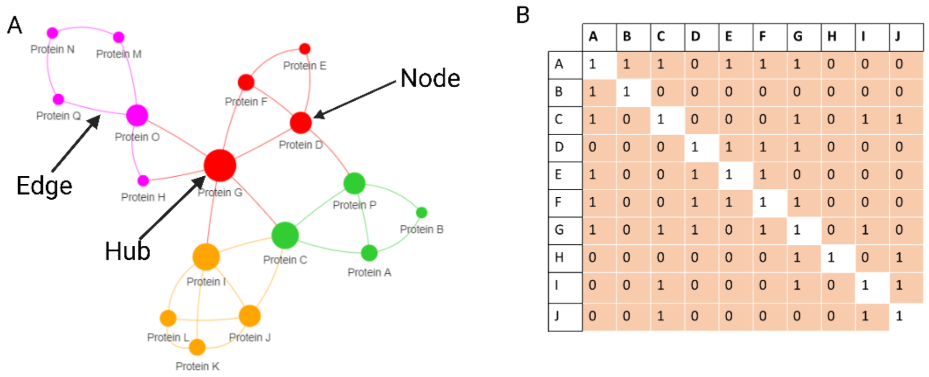
(**Figure 1.3B**) which is not the case for directed graphs where edge information travels unidirectionally across a single edge (e.g., signal transduction). A universal measure for the number of connections a node has, is called the degree, which is the number of edges a given node has.

The data represented within the adjacency matrix represents the networks organisation and architecture which corresponds to the workings of the systems being modelled. In the case of biological networks, patterns observed across the adjacency are of biological significance.[200-203] The direct relationship between network structure (topology) and biological function requires, for larger networks, mathematical modelling of the data. Notable properties of a biological network to model includes the degree distribution, centrality, and the presence of communities (often referred to as clusters, modules or cliques).

The degree distribution is the binning of nodes according to their degree (connectivity k) and can describe the probability that a node has a given k. The degree distribution can be described by many distributions.[204,205] For example, a network may have a homogenous distribution in its degree connectivity showing Poisson, binomial or exponential distributions where there is a more uniform distribution.[183] Typically network biology works under the principles that a network shows a scale-free distribution that can be described by a power law k^{-a}.[204] This is where degree distribution is right-skewed showing many nodes with low degrees and few nodes with high degree. From this another crucial feature of biological networks is identified – the hub, which are nodes with the highest degree and are binned on the tail end of the distribution.[206]

Hubs play a crucial role to network topology where deleting hubs can erase vast information from the network.[207] As biological networks are a reconstruction of patterns observed in biological data, then hubs are functionally important.[207] This notion is demonstrated in protein-protein interaction networks (PPINs) where deleting a hub protein is likely to be more lethal to an organisms function versus a non-hub – called centrality-lethality rule.[208,209] This rule has been observed in a range of eukaryotic organisms, and it can be inferred that due to the many interactions of the hub, they are potentially master regulators in signalling pathways.[207,208]

Centrality is a numerical measure of the importance of a node.[210,211] The degree centrality is the most common which is equal to the number of connections a node has. Other widely used node importance metrics include betweenness centrality, which measures how often a node lies on the shortest paths between other nodes (the flow of information through the network); closeness centrality, which measures how close a node is to every other node in the network (how fast a node can pass information to other nodes); and eigenvector centrality, which measures the degree of a node as well as the degree of its neighbours (the influence of the node on the network).[212-214]



	A	B	C	D	E	F	G	H	I	J
A	1	1	1	0	1	1	1	0	0	0
B	1	1	0	0	0	0	0	0	0	0
C	1	0	1	0	0	0	1	0	1	1
D	0	0	0	1	1	1	1	0	0	0
E	1	0	0	1	1	1	0	0	0	0
F	1	0	0	1	1	1	1	0	0	0
G	1	0	1	1	0	1	1	0	1	0
H	0	0	0	0	0	0	1	1	0	1
I	0	0	1	0	0	0	1	0	1	1
J	0	0	1	0	0	0	0	0	1	1

Figure 1.3: A: A mock protein-protein interaction network showing proteins (nodes) and their interactions (edges). Colours depict clusters (or modules) displays groups of proteins with high interconnectedness. Node size is proportional to the node degree (connectivity) **B:** An example of an adjacency matrix encoding edge information.

Biological data can be projected as a network using multiple tools available in many programming languages, including, igraph (R and, Python), visNetwork (R) and NetworkX (Python). These packages also have built in graph-based analyses tools to assess network-based metrics (e.g. degree, centrality etc). For those who are unfamiliar with these programming languages there are also user-friendly tools available, such as CytoScape, which includes multiple plug-ins for various graph-based applications.[215,216] Two commonly used networks in the identification of hub genes in cancer are gene co-expression networks (GCNs) and protein-protein interaction networks (PPINs). Gene regulatory networks are also common but are detailed in the ***Appendix A.1***.

1.3.2 Protein-protein interaction networks

PPINs model the repertoire of physical protein interactions that occur between proteins, where in the study of disease they typically serve to outline the underlying molecular mechanisms.[217,218]PPINs are usually unweighted and undirected graphs (no directional description to the edge).[219,220] Proteins will cluster into groups of interactivity, which can represent either modules of shared function, or physical protein complexes.[221] PPIs are derived from a range of biophysical, biochemical, genetic and computational methods, where examples of each are summarised in **table 1.4**.[222-226] There are a selection of online databases that provide PPI information, with some databases containing tools for network projection, the most common being STRING (**table 1.4**).[227]

Table 1.3: Examples of commonly described experimental techniques used to detect protein-protein interactions.

Method type	Method name	Approach	Throughput	PPIs
Biophysical	Nuclear magnetic resonance (NMR) spectroscopy	*In vitro*	Low	Direct
Biochemical	Co-immunoprecipitation	*In vitro*	Low	Direct and Indirect
Biochemical	Tandem affinity purification mass spectrometry (AP-MS – shotgun proteomics)	*In vitro*	High	Indirect
Genetic	Yeast-2-Hybrid (Y2H) system	*In vivo*	High	Direct
Genetic	Protein microarray	*In vitro*	High	Direct

1.3.3 Gene co-expression networks

GCNs are derived from a gene-gene correlation tests across a gene-sample matrix based on gene expression data derived from transcriptomic profiling (e.g. RNA-seq and microarray).[228] Pairwise correlation metrics can be used (e.g., Pearsons correlation coefficient) as well as other measures of correlations such as mutual information (MI) (e.g. in the initial co-expression analysis step in ARACNE – Algorithm for the Reconstruction of Accurate Cellular Networks) or partial correlation (e.g. PCIT – Partial Correlation and Information Theory).[229,230]

An unweighted GCN is derived from building an adjacency matrix based on gene-gene correlations that passes a hard-threshold defined by an arbitrary value (e.g. r > 0.7), statistical significance (e.g. p <0.05, t-distribution test), premutation testing, or spectral Laplacian properties.[231,232] Hard-thresholding in this manner may not retain biological insight as many genes may have a low correlation value, but still be important to the information flow through the network (i.e., it's overall topology).

For weighted networks soft thresholding, sometimes referred to as a smoothing parameter, is used to provide a continuous adjacency where the edge value represents the similarity or strength of connection.[233,234] This approach retains the continuous nature of pairwise correlations and ensures that the network generally exhibits scale-free topology through the retention of all gene connections.[196,228,234]

There are several tools available for the construction and analysis of weighted GCNs (WGCNs) that includes: petal, CEMItool, CoExp, GWENA, wTO, GeCoNet-Tool and weighted gene co-expression network analysis (WGCNA).[233-237] Despite differences in methodology the general framework is similar and includes all or in-part the following steps: 1) Calculating the gene adjacency, 2) transform degree distribution to approximate scale-free topology, 3) define network topology and 4) Identify modules of genes. Typically, these methods adhere to the guilt-by-association approach of networks where connected genes or genes within the same module are assumed to have similar or related biological functions and significance.[234]

The commonest method (with over 6165 search results in PubMed Central as of 2024) is WGCNA and some of the previously mentioned tools directly leverage WGCNA in their pipelines including CoExp and GWENA.[233,234,236] WGCNA is widely adopted across cancer studies and has been successful in identifying transcriptomic associations pertinent to disease conditions.[186,238-249] WGCNA is favoured for several reasons which includes the suite of tools available, the availability of in-depth guidance, stability among parameter changes, and favourable performance. Furthermore, WGCNA contains options for validating co-expression patterns and for data dimensionality reduction to simplify networks and observed patterns in gene connectivity.[234,250-252] An overview of the WGCNA methodology is described in **section 2.14.** Briefly, the crucial steps of the pipeline are:[234,253]

1. **Data preparation**: Raw gene expression data is pre-processed to remove noise and normalise the values.
2. **Sample clustering**: This step involves grouping samples based on their similarity in gene expression profiles. It helps in identifying outliers that do not conform to the general pattern.
3. **Network construction:** The construction of the GCN commences with the calculation of pairwise correlations between all genes across the chosen samples. The correlation matrix is then transformed into an adjacency matrix using a power function. This is followed by calculation of the Topological Overlap Measure (TOM).
4. **Module detection:** This step involves hierarchical clustering of the TOM, followed by dynamic tree cutting to define clusters of highly connected genes.[254]
5. **Module eigengene calculation:** The first principal component (PC1) of each individual module is calculated. This is referred to as the Module Eigengene (ME) and represents the gene expression profile of a module.[251]

6. **Relate modules to external traits:** This step involves correlating the MEs with external metadata to identify modules that are significantly associated with the trait of interest.

The application of WGCNA to cancer studies is broad and this includes the identification of co-expression modules relating to conditions, identification of biomarkers and potential drug targets.[186,238-249] There are several advantages to WGCNA over more conventional statistical approaches such as differential expression. It incorporates a systems approach which considers network topology which allows a holistic perspective of gene importance as shared units rather than singular genes passing a significance threshold. Then, co-expression can detail disease specific variation in a gene pairs expression value where differential expression cannot.[255]

Another advantageous feature of WGCNA is the incorporation of a data reduction step of ME calculation.[228,234,251,256] The expression profiles for each gene can then be correlated to the MEs to define a module membership metric (kME) (another form of connectiivty) and infer relationships with external data such as clinical variables (e.g., disease conditions), results from other bioinformatic pipelines, or other experimental techniques.[238,239,242,243,246] This "fuzzy" measure allows genes to show membership to multiple modules, highlighting the relationships beyond single module assignments.[234] MEs can also be projected as an eigen gene network (EGN) to proivde a higher order representation of the data, facilitating the visualisation of intramodualr relationships.[251,257] Additionally, feasture selection based on the MEs can improve the identificaiton of significant associations by mitigating the problems assocaited with high dimensional data. MiBiOmics , a web-based integrative pipeline, has taken advantage of this approach to quantify the realtive contribution of a given sample to module-trait associations.[258]

The use of WGCNA paired with differential expression analysis has been shown to improve inference, and it is generally not recommended to build a network following differential expression.[259] The ability of WGCNA partitioning to improve inference in subsequent functional annotations (e.g. via gene set enrichment analysis – GSEA) has been noted.[259] In addition, WGCNA partitioning was also found to improve performance of subsequent graph-based analysis.[260] Therefore use of WGCNA has the potential to identify hub genes that maybe key regulators in DDLPS biology for prioritisation of downstream drug-target screening. Furthermore, deriving gene-sets using this unsupervised approach is likely to identify novel interactions between genes, and improve functional annotations by approaches such as gene set enrichment analysis (GSEA).[259] Additionally, by using WGCNA to split up the GCN into gene sets improves the performance of sub network inference methods.[260]

Furthermore, the WGCNA R package contains a comprehensive method to validate modular expression patterns.[252] This validation process termed 'module preservation' contains multiple distinct tests which are summarised into a singular statistic which allows for easy interpretation. Lastly, WGCNA is more likely to find novel associations between genes as the partitioning of the GCN into modules is performed in an unsupervised manner and hence is a good fit for identifying novel therapeutic targets with no use of *prior* knowledge.

1.3.4 Extracting information from networks

There are several methods within the literature that aim to identify clusters of genes that maybe associated with disease mechanisms. A common method is to search for PPIs from a database using a list of differentially expressed genes related to a biological pathway of interest.[261,262] This allows for a quick identification of hub genes that maybe important. However, such methods are too reliant on the results from differential expression analysis, which can differ between methodologies and datasets used.[263] More comprehensive methods use a series of mathematical or topological methods to identify sub-networks which can then be investigated in subsequent analysis.[184,264,265]

Clinical or phenotypic traits are the most common types of data integrated with WGCNA modules.[253,266-268] Although it is possible to use results from other analyses including differentially expressed genes.[259] Once a module strongly associated with a disease trait is found (e.g., with poor outcome) then highly connected hub-genes can be identified within the module for further investigation., For instance, searching for known drug-target interactions for the hub-genes.[269] In addition, genes identified by genome-wide association studies can be assessed for their linkage with modules to detect all of the assumed components related to a disease.[270] WGCNA was developed to analyse gene expression profiles across patients and facilitate integration via the use of ME values. A subsequent study has utilised an approach to identify concordant hubs in WGCNA modules and PPINs, identified from PPI data based on differentially expressed gene patterns.[239]

Inspecting networks as a single network (monoplex graph) is an informative system-based approach. However, monoplex networks may oversimplify processes, and when viewed as individual layers can possibly lead to inferences that maybe misleading or incomplete.[197,271,272] Multiplex (or multi-layered) networks are a method of stacking monoplex networks that represent the same or different biological information type. Multiplex networks may have shared (homogenous) or different nodes (heterogenous) and different types of edges between nodes across different layers (or slices).[197,271,273] Additionally multiplexed networks can be categorized accordingly to node, edge and properties.[274]

Multiplexing allows the assessment of both inter and intra layer connections between nodes, which can be exploited to discern transcriptional and post-transcriptional layers of regulation. For example, the associations between miRNAs and protein expression[275] ,associations between gene co-expression, co-methylation and protein interaction[276], and similarities between patient groups based on gene expression, methylation patterns and miRNA profiles.[277] Multi-view networks (both multiplex and heterogenous networks) facilitate multi-omic data integration and you will often see multi-view analysis methodologies cited as network-based multi-omic data integration techniques.[278,279] These approaches could be incorporated within the methodologies of studies discussed here, in particular those which utilise data susceptible to noise (e.g., gene expression).[273]

Network diffusion models are a subset of network propagation and fusion methods.[278,280,281] These methods simulate a particle that diffuses through a network(s) and derives a measure of proximity, from which guilt-by-association can be applied. The most utilised network diffusion method is random walk with repeat (RWR).[192,282,283] A random walk simulates a particle that randomly moves from node to node. RWR is a random where after each step there is a probability (typically P = 0.7) of returning to the original seed node, specific by the user. The probability that the particle is at any given node at any given time can be used as a measure of proximity to the seed.[282] The effectiveness of this approach for integrating gene expression and copy number data for pathway analysis has been demonstrated.[284] Methods such as DART (denoising algorithm based on relevance network topology) and DRW (directed random walk; a random walk-based method) can only perform pathway inference on a single genomic profile (i.e., the integration of existing pathway information with gene expression data).[285,286]

Integrated DRW (iDRW) and random walk with restart on multiplex heterogenous networks (RWRMH) are two methods which seek to address this by incorporating multi-layered graphs within their methodology .[282,287,288] RWRMH performs a local neighbourhood-based search utilising a specific seed node. In addition to exploring areas within a monoplex layer, it can also walk to other layers via shared nodes. In RWRMH a parameter controlling the probability of the particle being at a given node in each layer can be set by the user. The main disadvantage to RWRMH is that it limits the integration of up to 5 network layers and it is limited to local searches with 199 nodes in addition to the seed. In addition, when shared nodes are specified it requires an input list that specifies node-node relations. A similar method, iDRW utilises a RWR to merge monoplex layers containing omics information with pathway-based gene-gene graphs to infer pathway activity profiles that are then associated with clinical outcome.[288] iDRW prioritises pathways that are correlated with poor outcome. A disadvantage over RWRMH is that merged pathway-level metrics are less interpretable at the gene-level compared to RWRMH. In addition, iDRW utilises directed graphs from prior knowledge sources (e.g. KEGG) and hence includes associated biases.

1.3.5 Network analysis in LPS

1.3.5.1 Hub genes

Network analysis and the identification of hub-genes is a well-recognised systems approach for identifying putative key molecules of cancer biology that could serve as therapeutic targets or biomarkers.[262,289,290] The application of network analysis to identify hubs is limited for DDLPS with only a few studies at the time of writing investigating LPS (***Table 1.4***).[239,249,262,289,291,292] The methods vary from building PPINs using online PPI databases (e.g., STRING) and differentially expressed genes as input, to intersecting multiple bioinformatic/statistical analysis such as WGCNA.

Table 1.4: Network analysis studies in LPS that have identified hub genes.

Ref.	LPS subtype	Method	Hub screening method	Hubs of interest	Clinical associations	Module preservation	Year
262	DDLPS	DEG PPIN	Degree centrality	APP, MDM2, CDK4	Not investigated	No	2017
289	LPS (MLPS – discovery and LPS validation)	DEG PPIN	Degree centrality	NIP7, RPL10L, MCM2	DRFS	No	2018
291	DDLPS	WGCNA, DEG, LASSO, survival analysis	kME and GS, LASSO results	UBE2C, ADIPOQ, PRC1	DRFS	No	2023
239	LPS (as part of a STS study)	WGCNA, PPIN (STRING)	Degree centrality	BUB1, RRM2, CENPF, KIF20A	OS, DRFS	Yes	2019
249	Retroperitoneal LPS	DEG, GSEA, WGCNA, LASSO	WGCNA, and LASSO filtering	PLCG1	OS	No	2023
292	Retroperitoneal LPS	WGCNA, Survival associations	Betweenness centrality	NINJ1	DRFS, OS	No	2024

LPS – Liposarcoma, DDLPS – dedifferentiated liposarcoma, MLPS – myxoid liposarcoma, WGCNA – weighted gene co-expression network analysis, PPIN – Protein-protein interaction networks, LASSO – Least Absolute Shrinkage and Selector Operator regression, GSEA – Gene Set Enrichment Analysis, DEG – Differentially expressed genes, OS – overall survival, DRFS – Distant recurrence free survival.

Some studies in LPS have identified hubs that predict survival trends serving as candidate biomarkers. For instance, Low gene expression of *Ninjurin-1* **(***NINJ1)* was identified to be predictive of a poorer DRFS and OS in retroperitoneal LPS (including DDLPS) compared to high expression.[292] NINJ1 is an emerging factor in mediating inflammatory activation following cellular lysis following cellular death processes and has been associated with a chronic inflammation.[293] NINJ1 has not been targeted in sarcoma, but has been the subject of potential therapy routes in inflammatory diseases, where inhibition notably reduced pro-inflammatory cytokine molecules.[293,294]

BUB1 mitotic checkpoint serine/threonine kinase B (*BUB1B*)*, Ribonucleotide reductase regulatory subunit M2* (*RRM2*)*, Centromere protein F* (*CENPF*) *and Kinesin family member 20A* (*KIF20A*) hub genes were found to predict poor OS and DFS in STS which were primarily mixed LPS subtypes.*[239]* These hubs all have notable functions in the cell cycle where shared function is not surprising given these were extracted from the same module. Whilst highlighted as biomarkers, these may have therapeutic potential in LPS, although no such studies have been conducted. *BUB1* has been found to be upregulated in several STS including high-grade MLPS and PLPS.[295] Furthermore, *RRM2* has been found to be upregulated in RLPS tissues, and knockdowns can inhibit G1/S cell cycle transition.[296] *CENPF* is upregulated in liposarcoma tissues being notably higher in DDLPS and has been associated with poor OS and DRFS.[297] Downregulation of *KIF20A* suppressed tumour growth and inhibited proliferation in STS cell lines and murine xenografts.[298]

Studies have built PPINs using differentially expressed genes in DDLPS which have identified hub genes that may represent putative targets or biomarkers.[262,289] One study identified Amyloid-beta precursor protein (APP) which has been found to promote cancer growth in Ewing's sarcoma, and has been implicated in other cancers including breast, prostate, pancreatic and melanoma.[299,300-303] The role of APP in DDLPS remains unclear. APP is targeted in the treatment of Alzheimer's, by inhibitory compounds for beta-secretase an enzyme that is important to APP processing.[304] Hence, APP is possibly an actionable cancer treatment and may warrant further investigation in DDLPS.

Another study identified hubs in MLPS including nucleolar pre-rRNA processing protein NIP7 (*NIP7*)*, Ribosomal protein L10 like (RPL10L),* and *Minichromosome maintenance complex component 2 (MCM2*) were identified as hub genes in MLPS that associated with distant recurrence free survival (DFRS).*[289]* *MCM2* has been found to show upregulation in both mRNA expression and protein expression in liposarcomas versus normal adipose tissue.[305]

Furthermore, the inhibition of MCM2 increased sensitivity of cancer cell lines to doxorubicin treatment.[305]

Another study leveraged WGNCA to identify modular hubs where *Adiponectin* (*ADIPOQ*)*, Ubiquitin-conjugating enzyme E2 C* (*UBE2C*)*, Polycomb repressive complex 1* (*PRC1*) were identified as potential biomarkers.[291] Another study used a similar approach[249] focusing on lipid metabolism associated genes in RPLS and identified *Elongation of very long chain fatty acids-like 2* (*ELOVL2*) and *Phospholipase C gamma 1* (*PLCG1*) as hubs.[249] They found that high expression of corelated to a lower predicted infiltration of immune cells and an immune-excluded phenotype, highlighting *ELOVL2* as a potential target for immunotherapeutic interventions.

To summarise, network-based analysis has been informative to studies in STS and in a more limited capacity to LPS. However, there is significant scope to expand the application of a network-based approach, particularly to DDLPS. With a wealth of data and comprehensive network-based methods available, it is feasible to apply a network-based approach to DDLPS to reveal genes that are key to the DDLPS disease mechanisms and potential candidates for therapeutic intervention.

1.3.5.2 Current limitations and future directions

Both studies by Yu et al[262] and Liu et al[289] would have likely benefitted from utilising the modular partitioning of WGCNA in deriving modules of interest associated with clinical variables or other omics data, which has proven to be successful in identifying key-genes and pathways in other cancers.[244,267,306]

The study by Zhu *et al*[239] emphasised the use of robust criteria for selecting validation datasets for module preservation analysis, and additionally utilising a prior-knowledge PPIN to aid in selecting for hub-genes associated with prognosis. Modules preservation is a technique that has not been used in any other LPS study despite it being a powerful and efficient metric in validating co-expression patterns between datasets.[252]

Currently studies have utilised WGCNA and PPIN based approaches to identify modules and hub-genes. Screening strategies usually involve selecting modules either associated with a clinical trait (e.g., survival) or centrality indices to screen for genes of interest. It would be beneficial to explore other sources of data for integration. Other studies integrate several computational experiments by assessing the gene overlap between significant or interesting results (e.g., genes that are screened out of a regression analysis or are overexpressed in cancer versus normal).[249,291] Such studies have been conducted in various cancers including breast[307-309], ovarian[248,310], colorectal[247], prostate[311], and cervical[312] among others. Studies conducted in gastric[248] and breast[310] cancers, adopted an approach utilising gene dependency data from the Cancer Dependency MAP (DepMap) project from the Broad Institute.[313]

The DepMap portal provides data from CRISPR-Cas9 knockout studies on various cancer cell lines, including three DDLPS cell-lines (LPS141, LPS510 and LPS853) derived from DDLPS patients.[150,314] From this data a gene effect score is inferred indicating whether a knockout has positive or negative effects on the viability of a given cancer cell line. As of the 2022 Quarter 3 (22Q3) DepMap data release, gene effect scores are inferred using the CHRONOS algorithm.[315] This algorithm works by modelling the observed depletions in single-guide RNA (sgRNA) abundance from CRISPR-Cas9 screens to determine the gene effect on cell fitness. Depletions in sgRNA indicates that there are fewer surviving cells carrying sgRNAs.[315,316]

The studies combining WGCNA and DepMap do not directly integrate the data with the GCN which could be achieved using the MEs.[248,310] Instead they use high/low dependency genes as a pre-filtering strategy which was a similar observation for LPS studies combining DEG analysis with WGCNA. As previously mentioned, this is typically not recommended as not only does its risk violating scale-free assumptions of WGCNA but can also decrease the accuracy of it.[259]

Furthermore, current studies have not used multiplex graphs to leverage multiple types of information or utilise methods to explore networks and extract information (e.g., RWR). Such an approach may be particularly useful for identifying relationships among modules, highlighting robust similarities (e.g., concordant co-expression edges) between datasets, and novel associations that may otherwise have been missed.[282,317] Finally, for the identification of putative targets, there is now a wealth of information on drug-target interactions that can be leveraged.[318-320] This has not as of yet been assessed for DDLPS, and allows the opportunity to identify hub genes where there is available chemical matter.

1.4 Research proposal

1.4.1 Research motivation

Current treatment options for DDLPS only reach 5yr overall survival targets in 50% of patients; frontline treatment in DDLPS cases fails to provide high rates of efficacy with poor survival and high rates of local/distal recurrence. Additionally, current treatment standards leave patients with long-lasting adverse side effects. Whilst current in-trial targeted treatment is promising, especially for MDM2 and CDK4 inhibitors, they are only successful in a proportion of patients due to disease heterogeneity. Hence, there is a need to find novel therapeutic strategies for DDLPS, which aim to increase patient survival whilst reducing the risk of adverse events. In addition, there is more work to be done in understanding DDLPS disease mechanisms. An integrative WGCNA based approach has not yet been applied to DDLPS with considerations for drug-target interaction data. This project aims to identify novel candidate therapeutic targets and to inform on DDLPS disease mechanisms.

1.4.2 Research Aims & Objectives

Hypothesis: An integrative systems biology approach using WGCNA can identify robust, disease-relevant hub genes in DDLPS that represent candidate druggable targets for therapeutic development.

Aim 1: **Data curation and construction of a robust GCN**

Objective 1.1: Use data repositories (NCBI, GEO and Array Express) to identify gene expression data on DDLPS samples with matched clinical data.

Objective 1.2: Perform clinical data analysis to identify key clinicopathological variables/features for WGCNA module integration.

Objective 1.3: Propose a gene expression dataset to be selected for network construction and validation.

Objective 1.4: Construct a robust GCN by exploring available WGCNA options and parameters and consider available filtering strategies to ensure network quality.

Aim 2: Identify robust modules of co-expressed genes associated with DDLPS biology

Objective 2.1: Identify the preservation levels of co-expression modules in independent validation gene expression datasets.

Objective 2.2: Derive a multi-parameter module score that integrates differential gene expression, gene essentiality data, and survival associations to prioritise disease relevant modules.

Objective 2.3: Utilise gene set enrichment analysis (GSEA) using publicly available gene set annotations to characterise the underlying biology of GCN modules.

Objective 2.4: Integrate available single-cell RNA sequencing data to confirm cell type associations within the GCN.

Aim 3: **Identify hub genes that can be targeted therapeutically through integrated gene network and druggability assessments**

Objective 3.1: Utilise graph-based centrality analysis (degree and eigen centrality) on the multiplexed module sub-networks to identify hubs.

Objective 3.2: Integrate protein-protein interactions (STRING database) to identify functional interactions between sub-graph genes and their interactors.

Objective 3.3: Integrate available drug-target information from available databases (therapeutic target database and chemical probes portal) to highlight putative druggable targets.

Chapter 2 Methodology

2.1 Ethical approval

This project has ethical approval from the University of Southampton Research Governance Office, ERGO ID: 64294 and ethical approval is scheduled to end in September 2025.

2.2 Data analysis interfaces

The R programming language (*version 4.2.2)* was utilised in Rstudio (*version 3.3)* on a Windows operating system.[321,322] Tasks requiring high-performance computation (e.g., read alignment and read quantification) and use of Unix (Linux) software (distribution version – 4.8.5-44) were performed on the Iridis 5 supercomputer at the University of Southampton. The Iridis 5 supercomputer was accessed using a Secure Shell (SSH) session via the MobaXterm programme (https://mobaxterm.mobatek.net). When working remotely, the Global Protect (https://docs.paloaltonetworks.com/globalprotect) Very Private Network software was used to access the University of Southampton intranet. Python environments were set using Anaconda3.[323] The working python environment was set to python version 3.11.4 and Ipython version 8.12.0 and the data analysis suites Spyder (available at https://www.spyder-ide.org/) (version 5.4.1) were used as interfaces for data analysis in Python.[324]

2.3 Data access and assurance

Data were accessed and obtained according to data usage and restrictions policies for given data and data repositories. Data from The Cancer Genome Atlas[325], accessed via the Genomic Data Commons (GDC) consortium[326], and Gene Expression Omnibus (GEO)[327] were open access. Data from the DNA Data Bank of Japan (DDBJ – available at https://www.ddbj.nig.ac.jp/) had restricted access and was obtained via application through the National Bioscience Database Center (NBDC), for which a data assurance policy was agreed and signed by myself (The DDBJ-NSDC account manager), Professor William Tapper (the coordinating supervisor for this project) and Professor Jon Strefford (the head of Cancer Sciences at the University Of Southampton). This data usage policy was renewed on the 29/09/2024 until the 29/11/2024. This data will be explored using bioinformatic approaches to identify novel candidate drug targets. The data will be stored in accordance with the data management plan, where public data will be securely stored on secure research drives that are maintained by the University of Southampton and accessible through authorised personnel. At the end of the project, all data, their location, analyses conducted with associated bioinformatic pipelines will be available on the research filestores accessible by project members in accordance with data usage policies.

2.4 Data download

TCGA raw read count data (processed by TCGA using STAR) was acquired using the TCGAbiolinks (version 2.28.4) in R studio. Gene expression data from Gene Expression Omnibus (GEO: https://www.ncbi.nlm.nih.gov/geo/) was acquired using the GEOquery (*version 2.64.2*) in Rstudio.[328-330] For the TCGA data, gene quantification and metadata was stored as a "rda" file using the SummarizedExperiment R package (version 1.30.2). The correctness and completeness of available metadata was tested against available metadata files from the published TCGA SARC research article.[64] Data access to utilize RNA-seq data deposited within the DNA Databank Japan (DDBJ) repository was requested and granted (01.10.2022). Data was downloaded over the Linux operating system using rsa key encrypted files to a secure restricted access and password protected file store using the rsync software (version 3.1.2). rsync was chosen as the preferred tool for safe file transfer through recommendation from the HPC Soton team due to the ability to download large data packages in steps, and for integrated Message-Digest 5 Hashing (MD5#)[331] integrity checks.

Clinical annotations for the NCC and IMS datasets were retrieved from the associated publication which also detailed additional clinical information for the TCGA SARC cohort in a format that easily allowed cross-study comparisons.[114] Additional clinical information that detailed the primary/recurrent status of the tumours and further treatment modality information was received upon contacting the corresponding data provider at the NCC (Dr Hitoshi Ichikawa, Department of Clinical Genomics, NCC).

Datasets identified within the GEO were retrieved using the GEO query R package (version 2.68.0). Datasets requested from the Samsung Medical Center (SMC) were downloaded using World Wide Web get (Wget, version 1.14) over a File Transfer Protocol (FTP) connection.[144] Datasets downloaded for this project are listed in *table 2.1*.

Table 2.1: Dataset manifest for downloaded gene expression data.

Dataset Accession	Dataset source	Associated publication
phs000178	**NCI TCGA**	**The Cancer Genome Atlas Research Network**[64]
JGAS000177	**DDBJ**	**Hirata et al.**[114]
JGAS000182	**DDBJ**	**Hirata et al.**[114]
GSE30929	**GEO**	**Gobble et al.**[332]
GSE159659	**GEO**	**Zuco et al.**[333]
GSE221494	**Institute Curie (GEO)**	**Gruel et al.**[143]

NCI TCGA – National Cancer Institute The Cancer Genome Atlas, GEO – Gene Expression Omnibus. The accession number provided for TCGA SARC is the dbGAP accession. The data from Gruel et al was downloaded peer-to-peer after data was shared from publication others prior to it being hosted on GEO.

2.5 Data integrity

Data integrity was assessed using the MD5# algorithm to ensure all files were content correct.[331] In GeoQuery and TCGAbiolinks MD5# checks are done automatically and when using Linux were performed manually for each directory using the md5sum command to generate the checksums upon retrieval and the '-c' option to specify for MD5# codes to be compared. If discrepancies between the original and generated MD5# codes were found, then the original data for the compromised file were redownloaded until no discrepancies were present.

2.6 Data storage

Public data was stored using the University of Southampton OneDrive. Private controlled and large data packages were stored using a research store in accordance with data assurance guidelines this research store is access restricted to only the relevant research students and academic staff.

2.7 FASTQ processing, read alignment and processing

FASTQ reads were downloaded using secure FTP interfaces utilising the wget and sftp commands on Linux OS. FASTQ paired end reads from RNA-seq experiments were assessed using the FASTQC package (version 0.11.9) on the Linux platform.[334] Where overrepresented adapter content was detected, the fastp (0.22.0) programme (using only the adapter trimming setting, all other parameters were set to off) was used to identify the adapters and remove them.[335] Pre and post QC alignments were then tested to ensure that QC improved alignment quality by inspecting the percentage of uniquely mapped reads using FASTQC html reports to view alignment log files that were produced by the STAR aligner.[336] Alignment and counting of FASTQ reads were performed using the STAR alignment package (version 2.7.10a) in Linux using the GRCh38 reference genome assembly.[336] BAM files were set to be sorted by coordinates and the option 'quantMode' was set to 'GeneCounts' which specifies STAR to perform gene quantification. The STAR package was chosen as it produces high-quality alignments without major parameter tuning, is a splice-aware alignment tool, and is fast when using large RAM pool allocations which were available via the HPC.[337]

2.8 Gene expression filtering and normalisation

For RNA-seq count data, lowly expressed genes which can generate false positives were removed using the threshold of 10 or more reads in at least 70% of samples in the edgeR R package (*version 3.38.4*).[338,339] To correct for potential differences in total read depth between samples, effective library sizes were calculated using the Trimmed Mean of M-component (TMM) method.[340] Scaling using TMM has been shown to have better performance for normalisation than other methods (e.g., RPKM and FPKM).[341] Then counts per million (cpm) were calculated for reads using the effective library sizes set by TMM. These normalised counts were then used in subsequent downstream bioinformatic analysis.

For microarray expression datasets, the probe intensity values were examined for prior background correction, which were found by inspecting the dataset annotations which are included in the data package retrieved using the getGEO() function from the GEOquery (version 2.70.0) R package. The GSE159659 microarray data was acquired and inspected. The values were verified to be log quantile normalised by assessing whether the 99th quantile is larger than 100 (values greater than 100 indicate the data has not been log transformed). This method of normalisation is common practice in microarray data analysis.[327] The Limma R package (*version 3.52.4*) was chosen for microarray data pre-processing due to its inclusion of background correction and normalisation methods.[342] In microarray data, there can be numerous probes per gene interactions. This redundancy helps ensure accurate and reliable measurement of gene expression but for downstream analysis it is often necessary to select a single probe per gene to avoid redundant information and to simplify the data. In cases where there were multiple interactions for probe-gene, duplicates were consolidated based on average signal intensity between the probes where the original non-averaged probes were removed from further analysis.

2.9 Random Seeds and reproducibility

For the entirety of this project, for stochastic functions requiring a random seed number, this was set as 333 (or 123 where stated). This was required for network projection functions as well as module preservation functions.

2.10 Data Manipulation

Data within R was manipulated (e.g., filtering, selecting, pulling, pivoting) using functions from the dplyr (*version* 1.1.4) R package. To subset strings in R, the 'stringii' function from the stringr (version 1.5.1) package was used. To modify strings, the 'gsub' function from base R (version 4.3.1) was used, and string pattern recognition was performed using either the 'regexpr' and 'grepl' from base R and BiocGenerics R (version 0.48.1) packages.[343] Data processed on Unix pipelines was performed using the SED (GNU version 4.2.2) and AWK (GNU 4.0.2) programmes.

2.11 General Data Visualisation

All (unless otherwise specified) generic plots were generated using the ggplot2 (*version 3.5.1*) R package. The ggplot theme was set as 'theme_bw' and the virdis, mako and turbo colour palettes were used to ensure colour-blind friendly visualisations. To label plot features the ggrepel (*version* 0.9.5) R package was used. Heatmaps were generated using the pretty heatmaps – pheatmap R package[344], hierarchical clustering on rows or columns (further specified in relevant chapter methodology sections) was performed using the default settings of a complete linkage algorithm on the Euclidean distance transformed values. The complete linkage prioritises smaller clusters that are compact and typically show better separation compared to other linkage methods including single and average that chain smaller clusters into larger clusters. Hence, the default setting of complete was not changed as finer cluster detail in visualisations was deemed beneficial for the reason of larger clusters potentially masking patterns of data that could be important and are overlooked.

2.12 Network manipulation and visualisation.

Networks were projected using visNetwork (version 2.1.2) and ggraph R package (version 2.2.1). For smaller networks requiring lower compute time networks were visualised as a spring-embedded projections with physics simulated using the barnesHut with the following settings: CentralGravity = 0.5, gravitationalConstant = -10000, springLength = 100, and springConstant = 0.05) using visnetwork. Where graphs had moderate complexity the 'VisIgraphLayout' function was used, and physics disabled to speed-up projections. Where networks were complex (i.e., multiple node/edge types) and dense the ggraph package was favoured due to quicker resolving times, and the ability to facet by edge and node properties. The tidygraph (version 1.3.1) and igraph (version 2.0.3) R packages were used prior to visnetwork for filtering, calculation, or selection of network properties to derive node and edge lists or modify their features.[345] Network features were selected arbitrarily based on the time to resolve an interactive network, the readability/interpretability of nodes and their labels and one that shows temporal stability when interacted with for the purpose of saving visualisations as a static image.

2.13 General Statistical methods

To assess whether the data followed a normal distribution, the Shapiro-Wilk test was applied using the R function 'shapiro.test' from the *Stats (version 4.3.1) R* package. Hypothesis testing between groups for normally distributed data was performed using the unpaired independent t test via the 't.test' function. Non-normal data was tested using the Wilcoxon rank sum test, this was conducted as opposed to normalising data as the Wilcoxon rank sum does not assume normality and is robust to skewed distributions. Categorical data was tested using a Fisher's exact test. Data summary statistics were generated using the 'summary' function to provide minimum, mean, median, maximum as well as lower (25th) and upper (75th) quantiles. Variance and standard deviation were calculated using the 'var' (or 'sd' where variance was not required) function, median absolute deviation calculated using the 'mad' function. Where required data summary metrics were generated using separate R stats functions 'mean', 'median', 'quantile' and if this was desired as a column or row-wise metric the functions were fed to into R apply functions. Applied statistical methods for clinical data inspection are described in detail in section 3.3.3-3.3.8.

2.14 Weighted Gene Co-expression Network Analysis (WGCNA)

WGCNA is a method for analysing and interpreting GCNs constructed from quantitative omics data, most commonly transcriptomic.[234] The approach seeks to identify modules of genes whose gene expression is correlated (co-expression) and sort them into modules of similar genes.[234,251,254] Guilt-by-association is adopter here as the approach assumes that genes co-expressed with genes that have known functions may also be involved in those functions.

There are several benefits to using WGCNA.[234,251,252,254,256,346] One of these include the generation of module eigengenes (ME) which are used to summarise the expression profiles of gene modules and to assess the relationships between modules (e.g. the correlation or adjacency between MEs).[251] They also act as an efficient way of integrating external omics data with the primary example being associations with clinical variables such as survival which aids in selecting modules that may be important to the disease. WGCNA can also perform a dynamic hybrid cutting algorithm when defining branch cut height in the hierarchical clustering used to partition the GCN into modules.[254] This algorithm defines preliminary clusters based on dendrogram branches that are distinct from surrounding clusters where the hub (the tip of the branch) is densely connected, excluding genes that are too from a cluster. It then uses the dissimilarity matrix to assign excluded genes to close clusters. The benefit of using a dynamic algorithm is that the cut height of the branches does not have to be pre-defined as is found in "static" branch pruning.

The first step in WGCNA is to calculate a pairwise gene correlation matrix, with values ranging from -1 to 1, using available correlation coefficients, commonly Pearsons but can also include other measures such as Spearman, Kendall, or the midweight bi-correlation, which are more robust to outliers than the mean-based Pearson's coefficient.[250] The correlation matrix is then transformed into an adjacency matrix through soft-thresholding by raising the correlation matrix to a power of β.

The next step in WGCNA is to construct a topological overlap measure (TOM), where pairwise similarity between genes is weighted by the connectivity of shared neighbours.[196,347] Finally, the TOM is converted into a dissimilarity measure (1-TOM) and modules of tightly co-expressed genes are defined by performing unsupervised hierarchical clustering and partitioning of the resulting dendrogram with a dynamic tree cutting algorithm.[196,254,347]

The WGCNA R package was used for the gene co-expression network (GCN) analysis of bulk RNA-sequencing data (TCGA SARC and NCC datasets). Further details are provided in the relevant chapter methodologies (see **section 4.3.2 & 4.3.6, section 5.3.3** and **section 6.3.2**) but in brief filtered and normalised gene expression counts (TCGA SARC and NCC – see **section 2.8**) were used as input to WCGNA. Quality control was conducted using sample dendrograms, principal component analysis, and sample network metrics. Gene adjacencies were calculated using the 'adjacency' function, specifying a "signed" network and a power of 12 (the justifications for choosing such options are explained in detail in **section 4.4.1**). To build a weighted (see **section 1.3.3**) GCN the adjacency was used as input to the 'TOMsimilarity' function specifying a "signed" network, and a TOMdenom of "min". The TOM dissimilarity (1 – TOM) was clustered using a hierarchical clustering algorithm with "complete" linkage, and modules were defined using a dynamic tree cutting algorithm with a sensitivity parameter ('DeepSplit') of four (justifications for these options are provided in **section 4.4.2**). Module eigengenes (ME) were then defined using the 'ModuleEigengenes' function and gene module membership (kME) (see **section 1.3.3**) were calculated by finding the correlation between gene expression values to the 'expression' values of the ME.

The TOM for the TCGA and NCC was calculated as previously described in **section 5.3.1** and **section 2.14.** This methodology was justified by the results obtained in **section 4.4.2** which showed that these parameters better recapitulated modular strcutures corresponding to biological patterns**.** In brief, the TOM was calculated using the 'TOMsimilarity' function using a signed adjacency from gene expression data. To account for differences between the NCC and TCGA data due to use of different mRNA enrichment pre-processing, the NCC gene expression matrix was filtered according to genes within the TCGA gene expression matrix.[64,114]

2.15 Eigengene Network (EGN)

Eigengene networks (EGN), a network of MEs (see **section 1.3.3**) was defined as the signed adjacency (0.5 * cor(I,j)^2) of the ME expression matrix. Diagonal values for the adjacency matrix were zeroed to prevent self-connections. A graph object for further analysis was then created using the WGCNA function 'ExportNetworkToVisAnt' where the adjacency threshold (a hard-edge threshold) was set as required (see **section 5.3.3.** and **section 6.3.2** – in brief thresholds were chosen to either cull or retain ME edges). EGN networks were then projected using methodology as set out in **section 2.12**.

2.16 Single-cell RNA-sequencing data description and processing

10X Genomics Chromium Single-cell data was kindly provided by Sarah Watson of the Institute Curie prior to publication of their findings which have since been published and data now hosted by GEO (GSE221494).[143] The data package provided included the raw RNA readcounts, processed reads, and cell annotations using markers as described in the published material.[143] The data has since been hosted by GEO (GSE221494). Several annotations were made available, but two were used, the ‘annotation.global’ which provided a broad overview of the cell types (lymphocytes, myeloid, tumour, endothelial, pericyte, normal adipose, and red blood cell) and the “annotation.detail” which included finer annotations (CD4+ T lymphocytes, NK CD56 bright cells, Exhausted CD8+ T lymphocytes, NKT cells, MDSC, CXCL8+ intermediate monocytes, cDC2, pDC, non-classical monocytes, CD4+ Treg lymphocytes, Mast cells, Pericytes, TAN, Tumor cells, NK CD56 dim cells, cDC1, classical monocytes, TAM3, Plasmacytes (Plasma cells), Cycling cells, TAM1, B lymphocytes, Endothelial cells, Normal adipocytes, Mesothelial cells). The raw data included 11 samples of primary untreated DDLPS patients; the raw data included 28029 cells.

The Seurat R package (version = 5.1.0) was used to process the data.[348] Single cell data was first filtered using available metadata to filter for samples from DDLPS patients and to remove red blood cells using the obtained annotations and metadata. A Seurat object was created using the ‘CreateSeuratObject’ function, specifying a gene to be expressed in a minimum of 3 cells, and a minimum of 200 expressed genes. For data normalisation, the data was split by patient and the ‘SCTransform’ function using the variance-stabilising transformation (sctransfomration) version 2 (V2) was used to retrieve the most variable genes.[349] The sctransform V2 was considered over other Seurat implemented transformations due to its favourable performance of other transformations in both the stabilising of variance and the speed of the algorithm.[349,350]

The features were then prepared for data integration by first using the ‘SelectIntegrationFeatures’ function considering all the 3000 most variable genes across patient data identified by sctransform V2. Next the data was prepared for patient integration using the ‘PrepSCTIntegration’ function to scale the 3000 genes. Anchors for integration were identified using the ‘FindIntegrationAnchors’ to identify pairs of cells that are similar in expression between patients for the aligning of data. For this step the normalisation method was set to the “SCT” SCTransform normalisation method as used previously. The data was subsequently integrated using the ‘IntegrateData’ function specifying the same SCTtransform normalisation method.

For dimensionality reduction of the single-cell data principal component analysis (PCA) was performed to identify principal components (PCs). These PCs were used as input for a graph-based clustering technique adopted by Seurat where for each cell the nearest neighbour using the ‘FindNeighbors’ function. This step identifies the similarity between two cells performing edge weight correction by looking at the topological overlap (the Jaccard similarity). Cells and their neighbours were then clustered uses the Louvain algorithm via the ‘FindClusters’ function to identify groups of cells similar in their gene expression (PCs). A resolution of between 0.4-1.2 is recommended by the authors of Seurat. Here a resolution of 0.8 to achieve a higher number of clusters. Finally, Uniform Manifold Approximation and Projection (UMAP) was used to reduce and visualise the data performed via the 'RunUMAP' function placing similar cells together in low-dimensional space, chosen to preserve local and global features of the data, as well as for fast computation time compared to other methods such a t-stochastic neighbour embedding (tSNE).[351] In addition, UMAP has shown the highest stability being able to separate cell types effectively. The minimum distance was set to 0.8 and the spread was set to 1.6 for easier cluster visualisation (less compact). The first 30 dimensions was used.

To identify cell gene expression markers, the transformed data (SCT) was first prepared using the 'PrepSCTFindMarker' function and subsequently markers were found using the 'FindAlMarkers' function. For this, only positive markers, those expressed in 50% of cells in each cluster and had a logFC of 0.5 versus other clusters to be considered a cell marker.

2.17 Integrating WGCNA and single cell data

To integrate the single-cell and WGCNA results an overrepresentation test (using the Fishers exact statistic) to identify overlap for cell cluster genes and WGCNA module genes conducted using the GeneOverlap R package (version 1.38.0). Two functions from this package were used. The first was 'newGeneOverlap' which constructs a 2x2 contingency table to test the overlap of the two gene sets. Here these two gene sets were, 1. module genes and 2. the inferred single cel markers. The second function 'testGeneOverlap' performs a hypergeometic test (Fisher's exact) to test whether the observed overlap is greater that what is expected by chance alone. If the intersection between gene sets was found to be less than two genes the significance was manually zeroed. This was to prevent significant results with sparse overlap. A -log10(p-value) of 1.3 was considered significant. Results from this analysis was visualised using the pheatmap R package (version 1.01.2).

2.18 Figure creation

Figures were created primarily using BioRender.com as credited in the appropriate figure legends. In addition, the Microsoft Office Suite was used to create a small proportion of figures. Any figure adapted or using data/information presented in published materials are given the correct citations in accordance with academic integrity.

Chapter 3 Dataset screening and clinicopathological metrics

3.1 Introduction

Unsupervised/semi-supervised gene networks, particularly gene co-expression networks (GCN – see chapter 1.2.3) modelled from gene expression data, have seen increased attention in the literature for offering novel insights into cancer, including DDLPS.[238,246,249,268,291,352,353] The successful analysis of GCNs requires several considerations on input data for reliable, robust and biologically meaningful inferences.[256,259] Sample number is likely the most crucial consideration for gene networks where generally it is recommended to have a cohort size of 30 or more samples.[259,354-356] For weighted gene co-expression network analysis, the authors recommend a minimum sample size of 15, although higher sample numbers are preferential for robust and reliable results.[259,356,357] Reduced sample numbers decreases the stability of correlation values used and can result in less reliable distributions of gene connectivity and can therefore pose challenges for identifying biologically meaningful modules and hub genes.[234,357-359] Over the last decade, there has been several initiatives that have increased the availability of DDLPS omics data, particularly transcriptomic.[64,114,332,333] This makes the application of WGCNA more feasible in the DDLPS space providing the datasets contain an appropriate number of samples.

Integrating sample or gene-level information in WGCNA is an important step for inferring biological/disease functions of gene co-expression modules and screening modules that may play a crucial role in disease biology.[256] One common strategy is to integrate clinical-trait data with modules.[234,238,239,241,243,246,251,256] Modules that either correlate with or are significantly associated with specific clinical traits can be selected. Therefore, the availability of sample phenotypic data is another important consideration when searching for data, especially those with annotations for clinicopathological variables that are known to be important to DDLPS. Metastatic and/or advanced disease are the primary clinicopathological factors discussed in current clinical management guidelines in the context of DDLPS.[29,61] Studies have identified several clinical variables as significant predictors of survival in DDLPS using multivariate Cox proportional hazard models.[93,107,360,361] These variables include patient age, patient sex, primary tumour site, tumour grade, stage, whether surgery was performed, metastatic disease, and the residual tumour left at resection margins. Integrating clinicopathological variables with WGCNA modules has yet to be performed at a comprehensive level in DDLPS.

WGCNA is discussed in-depth in ***chapter 1.3.3*** and has been selected for use in identifying disease-related modules and their hub genes for several reasons. Briefly, these include: The ability to derive disease-related gene sets based on gene expression data.[234] The use of module-eigengenes as a data-reduced summary of module-to-module and module-to-trait relationships.[251] The in-built tool for validating module co-expression patterns across data summarises network preservation statistics into a single and easily interpretable value.[252]

This chapter aims to first identify suitable datasets, following the data considerations discussed, for use in a WGCNA-based methodology to identify drug targets. Then, to identify any clinical features that can be used for integration with downstream WGCNA analysis.

3.2 Chapter aims and objectives:

Hypothesis: A sufficiently large discovery and validation dataset, enriched with comprehensive clinical information, will be well-suited for identifying novel drug targets using a Weighted Gene Co-expression Network Analysis (WGCNA) approach.

Chapter Aims and Objectives:

Aim 1: Identify and select an appropriate dataset for WGCNA and downstream validation of module co-expression patterns:

Objective 1.1: Evaluate available DDLPS datasets considering sample size and sample information.

Objective 1.2: Determine the most suitable datasets for downstream WGCNA analysis and module preservation.

Aim 2: Describe clinical characteristics of selected datasets.

Objective 2.1: Provide a comprehensive overview of the clinical data available in each dataset.

Objective 2.2: Highlight relevant patient demographics, tumour characteristics, treatment history and other clinical variables.

Aim 3: Identify clinical features for downstream module-based target exploration.

Objective 3.1: Identify clinical variables that significantly predict patient survival outcomes.

Objective 3.2: Highlight features that may inform downstream module screening for target exploration.

3.3 Methods

3.3.1 Identification of datasets

In addition to the TCGA SARC dataset, further gene expression datasets were identified by interrogating data repositories including, The Gene Expression Omnibus (GEO), Array Express (AE) and NCBI as well as the literature. The following combinations of search terms were "DDLPS", "dedifferentiated liposarcoma", "liposarcoma", with "gene expression", "transcriptomics", "microarray", or "RNA-seq".

RNA-seq datasets were prioritised over microarray. Firstly, to be concordant with TCGA SARC. Secondly, for the various technological advantages of RNA-seq versus microarray. RNA-seq provides whole-transcriptomics quantification of gene expression, has higher sensitivity allowing the capturing of low-abundance transcripts, can detect novel transcripts and alternative splicing events, has reduced platform-specific biases, and greater flexibility in data analysis compared with microarrays.[362,363]

Upon identifying datasets, they were further inspected and summarised according to their relative sample size, disease subtypes (for datasets encompassing multiple subtypes), availability of normal adipose controls, and any peer-reviewed publications associated with the data. Data were examined for the inclusion of clinicopathological annotations and additional matched metadata. Clinical annotations, particularly those related to outcome measures, were of particular interest for the purpose of highlighting gene co-expression module-trait associations that maybe pertinent to disease biology.

Subsequently, for the purpose of selecting datasets for use in WGCNA, datasets were screened to ensure an adequate sample size of DDLPS samples. Following recommendations from WGCNA authors, a minimum sample size of 20 was considered sufficient for downstream analysis. This sample size threshold was selected, over the minimum of 15, to ensure inferences made were reliable and robust. Fewer than 15 samples would produce too much noise, while larger sample sizes have been associated with networks exhibiting higher associations between connectivity and function.[364]

Datasets containing matched normal and/or other liposarcoma (or soft-tissue sarcoma) subtypes are essential for comparative gene expression studies. The results from differential gene expression analyses using these data will serve the downstream purpose of identifying gene co-expression modules that contain gene sets that are significantly enriched for biological functions upregulated in dedifferentiated liposarcoma (DDLPS) compared to either WDLPS or normal adipose tissue. Notably, datasets containing normal adipose tissue are particularly interesting. The upregulation in the transcription of a gene in disease versus normal tissue may represent potential therapeutic windows for clinical interventions.

Where possible data was derived from treatment naïve primary tissue that had not been subject to chemo-or-radio therapy before and after sample collection. For a project aiming to identify novel targets that are generalisable it is beneficial that these samples represent DDLPS biology, without any treatment induced alterations to transcriptional regulatory programmes. Recurrent disease samples were included in the analysis providing they had no pre-operative or post-operative CRT (by personal communication with the authors) and they were not identified as outliers based on their gene expression profile. It is not uncommon for DDLPS to be labelled as “primary” on first occurrence of a DDLPS tumour even when it is not the first occurrence of a liposarcoma.

If tissue was identified to be Secondary DDLPS by aetiological definition, these were accepted provided they pass subsequent QC testing, and the patient had not been subjected to post-operative C/RT. In instances where the tumour status was unknown, the data provider was contacted for more information. As an example of this process, the data obtained from DNA Databank Japan (DDBJ), which includes gene expression and whole exome sequencing (WES) data, had incomplete clinical information.[114] This data was derived from two separate institutions, the National Cancer Centre (NCC) – Japan (JGAS000182), and the Institute of Molecular Sciences (IMS) – Japan (JGAS000177). Both institutions were contacted and the NCC were able to clarify the tumour and treatment status of all samples. However, IMS have not yet been able to provide information on the systemic treatment status of the patients (pre-or-post operative chemotherapy) and the status of the tumour (Primary/Recurrent) after personal correspondence.

3.3.2 Data acquisition and integrity assurance

Data from the TCGA SARC was acquired using the TCGA biolinks R package (version 2.28.4) and the ‘GDCquery’, ‘GDCdownload’ and ‘GDCprepare’ functions.[328] Gene expression data from Gene Expression Omnibus (GEO: https://www.ncbi.nlm.nih.gov/geo/) was acquired using the GEOquery (*version 2.64.2*) using the ‘getGEO’ function.[329] These data packages included both gene expression data as well as available metadata pertinent to this chapter. Metadata (including clinical data) for the DNA Databank of Japan (DDBJ) National Cancer Center Japan (NCC) was acquired from a corresponding author from the associated publication directly (and through request via the NBDC for further information), and from supplementary files publicly available through the published article.[114] Data integrity was ensured using the MD5#

algorithm which is integrated into the 'GDCdownload' and 'GDCprepare'. Further general details on data acquisition and integrity are detailed in **section 2.3-2.5.**

3.3.3 Descriptive statistics and data completeness

Descriptive statistics for the patient, tumour and treatment characteristics were summarised using base R functions and Microsoft Excel where available. For quantitative variables the lower quartiles, mean, and standard deviation were calculated and tabulated. The finalfit R package (version 1.0.7) was used to produce a summary table stratified by vital status of the patients to assess variable split between dead or alive/censored at last point of follow-up. The proportion of missing observations was calculated for each available and any variable showing ≥20% missingness was removed from downstream analysis to preserve statistical power in survival models. Stratified summary tables were also used to assess case separation, which details how variables and their categories are distributed among the conditions between assessed. Where variables are perfectly separated between conditions, survival analysis is challenging as the coefficients will be inflated towards infinite values, which would translate to illogically high hazard ratios, and thus incorrect inference. Variables showing near or complete case separation were removed before univariate and multivariate analysis.

3.3.4 Survival analysis: Univariate and Multivariate Cox Proportional Hazards

Survival analysis was conducted using the following R packages: finalfit (version 1.0.7), survminer (version 0.4.9) and survival (version 3.5-7). The time to event was taken directly from available clinical annotations given in days to event (death, recurrence, progression, or disease-specific mortality). Time as a unit of years to event was calculated by dividing time in days by 365.25. If the event was not already binary encoded, "1 "was taken as the event occurring, whereas "0" was taken as the censor. Where required for Cox proportional hazards testing (**section 3.3.5**) and for plotting survival curves the 'Surv' function was passed to the 'coxph' function from the survival package.

3.3.5 Survival analysis: Variable Selection Strategy

To select variables for survival model construction, several criteria were used for each outcome measure available. Variables that passed missingness and case separation tests (**see section 3.3.3**) were considered for univariate testing. Mitotic rate and necrosis score were not included in multivariate models as they are used to calculate FNCLCC grade. To then identify which variables to include for the best fitted model to available data, a conditional backwards elimination approach was used. A significance cut-off of *p<0.20* for elimination was chosen as the stopping criterion, where variable elimination was questioned under the basis of clinical reasoning; Variables known to be important or typically included in LPS survival models were deprioritized for elimination. This was set to ensure that clinically important/interesting variables were not excluded from the model unless variables invalidate the model.[365] As the sample size of the cohorts is low (max = 50), co-linearity assessments were conducted at each step of the backwards elimination. To assess for co-linearity, variance inflation factor (VIF) was calculated, and if VIF >5 was found then correlation tests appropriate to the variable format were conducted. If variables were found to be co-linear, then the variables were either merged, if feasible, or the highest quality variable was selected based on data completeness, case separation, or continuous format. An additional stopping criterion was to reach the number of desired variables to achieve an event-to-variable ratio of ~10:1.[365]

3.3.6 Survival analysis: Variable Caterogisation

Variables were analysed in a continuous format, where possible, to preserve patterns within the data.[366,367] Categorization of continuous was not performed to avoid misleading or inappropriate interpretations from survival analysis.[368] In addition and for the exception of variables used in the grading of sarcoma tumours, ratified points at which measured clinicopathological variable can be categorized are not available in DDLPS disease.

3.3.7 Survival analysis: Model diagnostics and assumptions testing

In the first instance, to assess the assumptions of proportional hazards the Schoenfeld residuals were calculated and plotted against time (***Appendix A.2***). A chi squared test on the residuals was then performed between the observed and expected hazards using the 'coxzph' function from the survival R package.[369] Any variable found to be significant suggests non-proportional hazards and were then further assessed for parallelism of log-log survival curve plots as log-log survival curves should be approximately parallel and should not touch.[370] Variable categories where there was a single entry were set as *NA* and if this resulted in no possible contrast the variable was removed to protect against model overfitting. The variables retained for further analysis were used as an input to a univariate cox proportional hazards model.

Variables that were shown to breach the proportional hazards assumptions were not used as input into the multivariate model. Variable co-linearity was assessed using the variance inflation factor (VIF) in R studio with the rms package (version 6.7-1) using the 'vif' function.[371-373] Any variable with a VIF > 5 (strong co-linearity) was further assessed for correlation using Pearson correlation, Pearson chi-squared test and Kruskal-Wallis tests. Any variable found to show further evidence of co-linearity were not used as input into the multivariate Cox proportional hazards model, where the most complete or informative variable was kept.

3.3.8 Survival analysis: Survival plots

To generate Kaplan-Meier plots and log-log plots the ‘ggsurvplot’ function from the survminer R package was used, with the fit from the survival package as input. To generate log-log plots using the ggsurvplot function the ‘fun’ option was specified as “cloglog”. For the assessment of the Schoenfeld residuals the ggcoxzph function was used. To generate more appropriate scaling on the plot y-axis the ggcoxzph function was modified to remove the factor (d *) when scaling the y-axis to prevent incorrect scales which may lead to incorrect interpretation of the standard deviation of residuals from the 0 line. This is detailed in the R scripts provided.

3.4 Results

3.4.1 Identification and selection of datasets

In total seven datasets were found to contain data from dedifferentiated liposarcoma (DDLPS) patient samples for use in this project (***Table 3.1***). Datasets containing dedifferentiated liposarcoma patient samples are predominantly from studies on multiple soft-tissue and bone sarcoma histological subtypes (***Figure 3.1***). There are 554 samples across these seven datasets of which any liposarcoma and adipose normal tissue were selected for further screening, including 184 [33.2%] DDLPS samples, 72 [13.0%] WDLPS, 36 [6.5%] MLPS, 20 [3.6%] PLPS, 15 [2.7%] adipose tissue, 11 [2.0%] MLPS/RC, and other sarcoma [39.0%]. The majority of DDLPS (114 [61.9%]) gene expression data were generated using bulk RNA-seq which is beneficial for this study given the multiple advantages of RNA-sequencing technology over array-based technologies including the detection of RNA species and whole-transcriptome quantification.[374,375] Due to these benefits, RNA-seq experiments (TCGA SARC, NCC, IMS and SMC datasets) were prioritised for the construction and validation of the gene co-expression network.

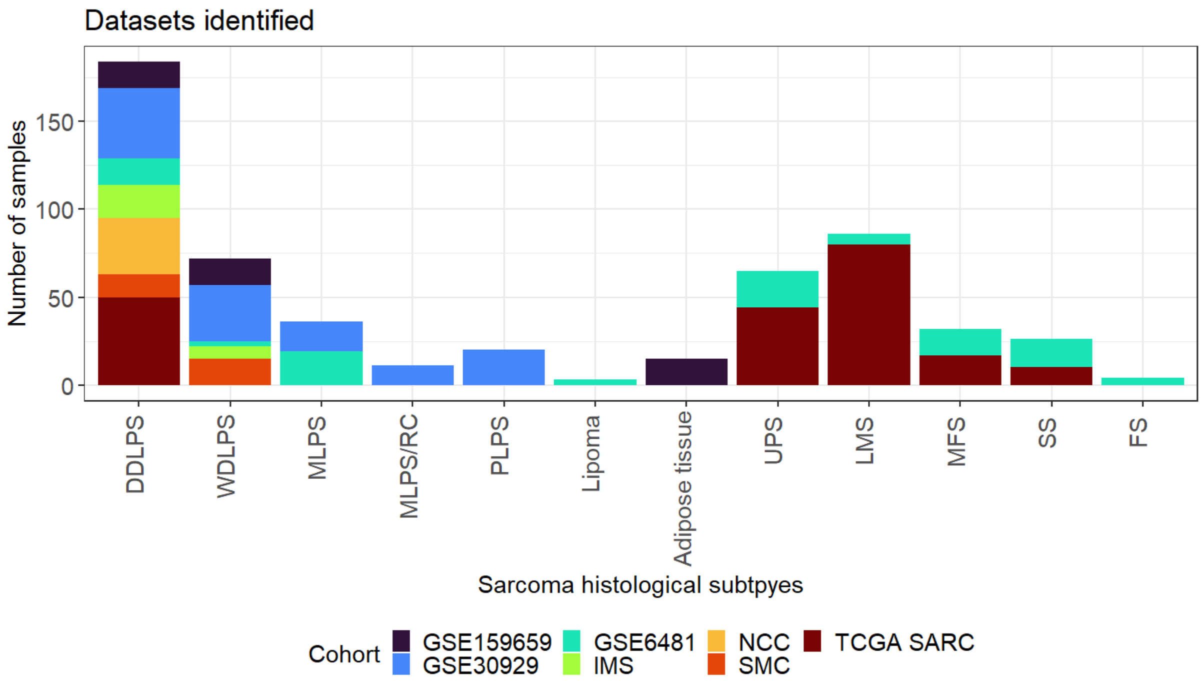


Figure 3.1. Sample counts for each sarcoma disease type by Cohort (dataset). GSE; GEO series data. TCGA; The Cancer Genome Atlas. IMS: Institute of Molecular Sciences. SMC: Samsung Medical Center. DDLPS: Dedifferentiated liposarcoma, WDLPS: Well-differentiated liposarcoma, MLPS: Myxoid liposarcoma, MLPC/RC: Round cell MLPS, PLPS: Pleomorphic liposarcoma, UPS: Undifferentiated pleomorphic sarcoma, LMS: Leiomyosarcoma, MFS: Myxofibrosarcoma, SS: Synovial sarcoma, FS: Fibrosarcoma.

Datasets were then screened for sample number meeting the recommended sample size of 20 as set out by the authors of WGCNA. The higher the number of samples used to construct a GCN results in more stable correlation values, a higher reproducibility of the network, and ability to capture biological annotations across gene-gene edges.[376,377] TCGA SARC (DDLPS, n = 50) and NCC (DDLPS, n = 32) were chosen for downstream analysis based on passing these criteria.

To increase the effective sample size, a horizontal integration strategy was considered, merging the available TCGA SARC and NCC data to gain a single dataset of 82 samples. However, TCGA SARC and NCC datasets use different RNA enrichment strategies for RNA-seq library preparation (table 4). TCGA SARC uses polyadenylated RNA selection (polyA+ selection), and NCC uses rRNA depletion (riboZero).[64,114] Whilst possible[378], it is not recommended to integrate RNA-seq data from these separate library types due to discrepancies in transcript abundance and the fraction of the transcriptome sequenced.[114,379,380] For this reason, and in favour of a validation strategy, the data were kept separate. TCGA SARC was selected for WGCNA due to larger sample size, more detailed clinical annotations, and that polyA+ selection retrieves a higher exonic coverage and shows higher accuracy in gene quantification.[379]

Table 3.1: Summary of DDLPS gene expression datasets discovered.

Source	Dataset ID	DDLPS (n)	WDLPS (n)	Adipose* (n)	Technology	Tissue/treatment status	Platform/library preparation	Passed screening
GEO	GSE30929	40	52	0	Array	Primary and treatment-naïve	GPL96	No
GEO	GSE159659	15	15	15	Array	Primary (Matched samples taken from the same patient)	GPL23159	No
GEO	GSE6481	15	0	0	Array	Unknown	GPL96	No
TCGA	SARC	50	0	0	RNA-seq,	Primary and treatment naïve	Illumina, PolyA+ selection	Yes
IMS	JGAS000177	19	7	0	RNA-seq	Primary and recurrent samples. Treatment status unknown	Illumina, PolyA+ selection	No
NCC	JGAS000182	32	0	0	RNA-seq	Primary and Recurrent. No systemic therapy.	Illumina, rRNA depletion (riboZero)	Yes
SMC	SMC	13	15	0	RNA-Seq	Primary and recurrent. No systemic therapy.	Illumina, PolyA+ selection	No

*Adipose is the normal tissue for comparison with DDLPS and WDLPS tumour samples. GEO: Gene Expression Omnibus. TCGA: The Cancer Genome Atlas. IMS: Institute of Molecular Sciences. NCC: National Cancer Center Japan. SMC: Samsung Medical Center. DDLPS: Dedifferentiated liposarcoma. WDLPS: Well-differentiated liposarcoma. “Passed screening” indicated which datasets were chosen for downstream analysis and which were excluded as they did not pass the screening criteria.

3.4.2 Patient, tumour, and outcome characteristics

Patient, tumour, and treatment characteristics were summarised for the TCGA SARC (DDLPS subset) and NCC cohorts (***Table 3.2***)**.**

3.4.2.1 Patient characteristics

The TCGA SARC was the only dataset to report patient age at diagnosis as a continuous variable, while the NCC summarised these into age categories at decade intervals. The age of the patients in the TCGA SARC (***Figure 3.2A***) was found to have a range of 51.8 years, with the youngest patient in this dataset being 34.9 years and the oldest being 86.8 years at diagnosis. The mean and median values for age in the TCGA SARC dataset were 63.5 and 62.0 respectively with the most frequent age category in both the TCGA SARC and NCC datasets being 60-69 years (***Figure 3.2B***).

The gender split was found to be 2.57:1 (male: female), with 59 male and 23 female patients (***Figure 3.2C***). For the TCGA SARC most patients, 49 were from white ethnic backgrounds where one had a Hispanic background, and 1 was unclassified (***Table 3.2***). The NCC data records all patients (n = 32) as being from Japanese patients (***Table 3.2***).

The majority of DDLPS tumours were in the retroperitoneum or abdomen (68 [83%]), with the remaining tumours locating to the extremity, shoulder, or girdle (11 [13%]) and the chest wall or back (4 [4%]) (***Figure 3.2D***). The age of the patients indicates that the majority of DDLPS samples within the TCGA SARC and NCC are from older demographics, whose sex was primarily male, with tumours occurring most commonly in the retroperitoneal or abdominal anatomical locations.

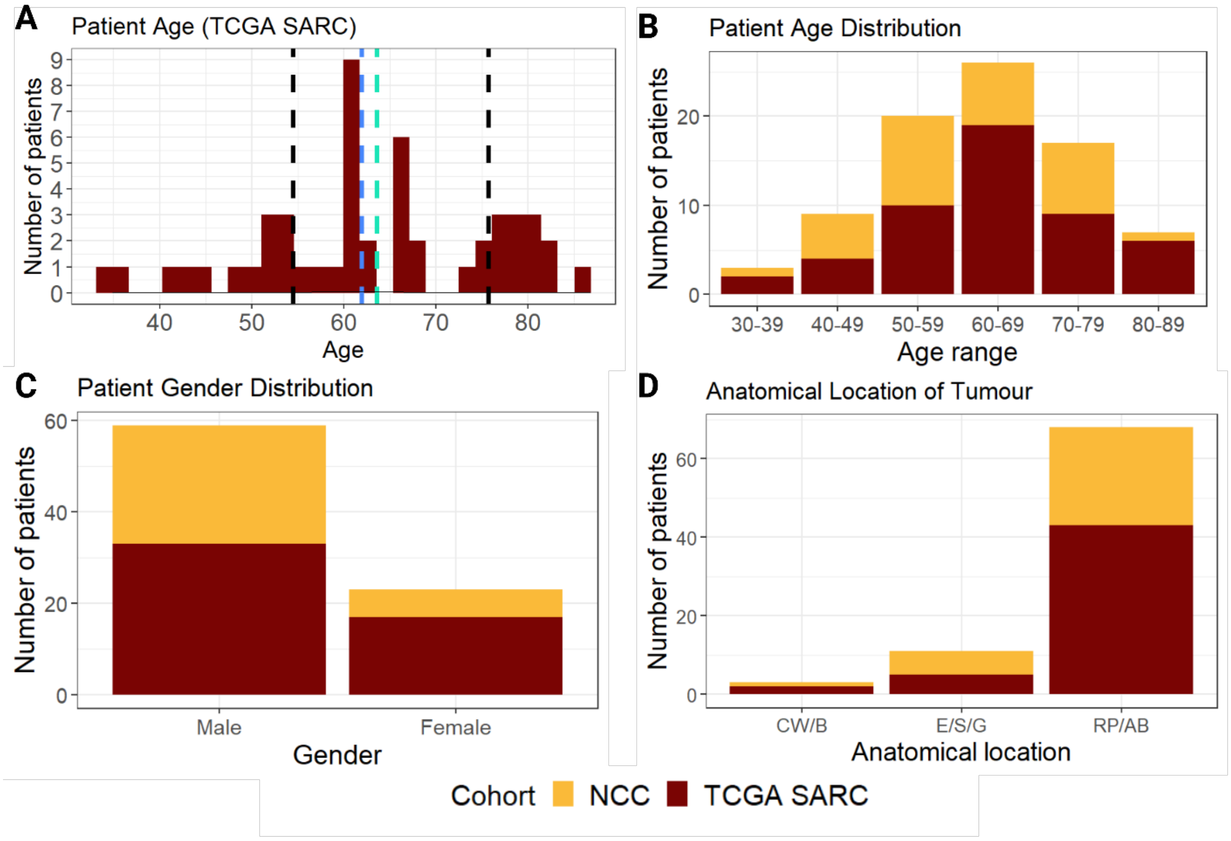


Figure 3.2. A: Quantitative patient age from the TCGA cohort. Black dotted vertical lines represent the lower(left) and upper (right) quartiles, cyan indicates the mean and turquoise indicates the median. **B:** Categorized patient age across TCGA and NCC cohorts. **C:** Patient gender split by cohort. **D:** Tumour anatomical location according to cohort.

3.4.2.2 Tumour characteristics

Most tumours showed microscopic tumour involvement (R1) after excision (43 [57%], followed by negative (R0) tumour involvement (30[39%]), and macroscopic tumour involvement (R2) being the least frequent (3 [4%]) (***Figure 3.3A***). Residual margins were either unclassified or not reported in 7 cases (labelled as RX) that were excluded from survival analysis. Tumour grading was available for the TCGA SARC dataset only, the majority of which were found to be grade 2 (37[74%]) followed by grade 3 (12 [24%]) and grade 1 (1 [2%]) (***Figure 3.3B***). Most TCGA SARC DDLPS tumours (49 [98%]) are high-grade (grade 2/3) as is typical for DDLPS.

The grading of sarcoma involves evaluating three tumour properties: the mitotic index, the percentage of necrosis, and the status of differentiation. These evaluations are used to assign a grade ranging from 1 to 3, with 3 being the highest grade.[67] The mitotic index is a measure of cellular proliferation as a percentage of cells in a microscopic survey across successive high power-fields (HPF). In sarcoma, the mitotic index is usually counted as the amount of mitosis across 10 HPFs although this is not specified in TCGA SARC metadata. In the TCGA SARC the mean mitotic index, was 7.16 (sd = 8.09), with a median of 5, and quartiles of 2 (lower) and 7 (upper) (***Figure 3.3C***). The mitotic index is used to derive a score: score 1 for 0-9 mitosis; score 2 for 10-19 mitosis; and score 3 for >19 mitosis. In the TCGA SARC 38 [76%] tumours have a mitotic rate score of 1, 7 [14%] have a score of 2, and the remaining 5 [10%] have a score of 3 (***Figure 3.3D***).

The percentage of necrosis is reported as an estimate of the percentage of the tumour that is necrotic. Tumours with no necrosis observed are assigned a necrosis score of 0, those with <50% are assigned a score of 1, and >50% a score of 2. In the TCGA SARC most tumours scored 1 (n = 28, [56%]), followed by a score of 0 (n = 21, [42%]) and by a score of 2 (n = 1, [2%]) (***Figure 3.3E***). The level of differentiation in the tumour was not available in the TCGA SARC metadata. Considering the sarcoma grading system these results indicate that DDLPS tumours in the TCGA SARC dataset are typically high-grade but generally have a low-proliferative index. This would suggest that the decision of assigning grades was driven by necrosis scoring and the status of differentiation. For a DDLPS tumour, where dedifferentiation is expected for diagnosis, this is not a surprising result.

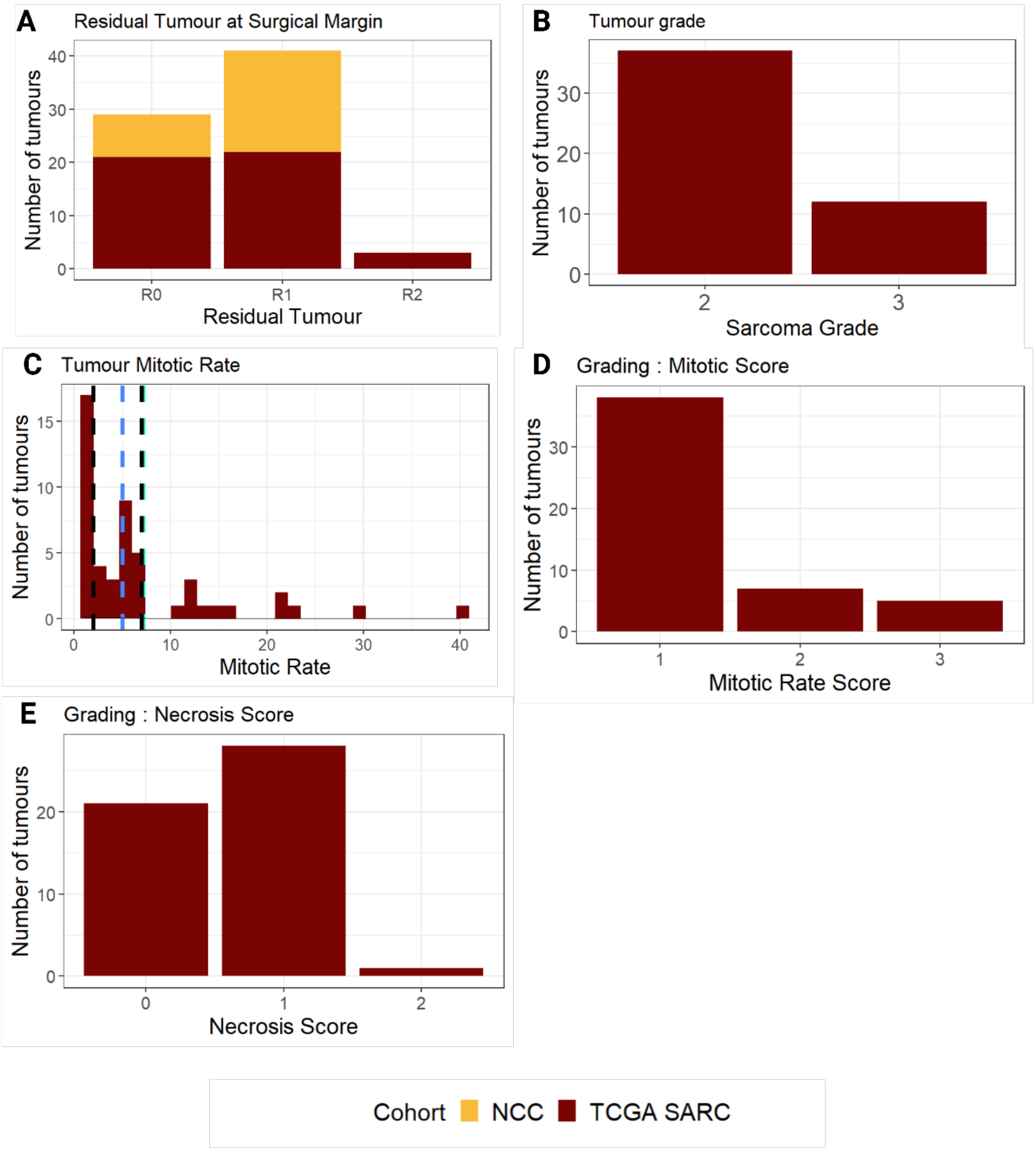


Figure 3.3: **A:** The reported involvement of tumour at resection margins in the TCGA SARC and NCC datasets. R0; Negative margins, R1; Microscopic tumour involvement, R2; Macroscopic tumour involvement. **B:** The reported FNCLCC (Sarcoma) grade of tumours in the TCGA SARC. **C:** The mitotic index (percentage of cells undergoing mitosis), and **D**; The miotic score of tumours within the TCGA SARC. **E:** The Necrosis score of tumours within the TCGA SARC cohort.

3.4.2.3 Treatment outcomes and survival

Surgical resection was conducted in 80 out of the 82 patients [98%] over both the TCGA SARC and the NCC (***Figure 3.4A***). Only two patients had not undergone surgical resection, suggesting that the tumour sample may have been obtained via a biopsy and that resection of the tumour was not appropriate (**Table 3.2**). Adjuvant radiotherapy was given to 8/81 [10%] patients (***Figure 3.4B***) indicating the possibility of locally advanced disease within those patients according to the treatment guidelines available for soft-tissue sarcoma.

Across both the TCGA SARC and NCC datasets four distinct survival outcomes were assessed (***Figure 3.4C, Table 3.2***). Disease-specific survival (DSS) and progression-free survival (PFS) were available for both datasets. Additionally, the TCGA SARC has data on overall survival (OS) and recurrence-free survival (RFS). The availability of PFS and DSS in both datasets enables validation of prognostic variables between datasets. In the TCGA SARC dataset, RFS was the most frequent outcome with 34 events (***Table 3.2; Figure 3.4C***). Conversely, in the NCC dataset, PFS had the highest frequency, with 22 events. Overall, PFS had the most frequent number of events across both datasets, with 51 events in 82 patients.

The estimated 5-year DSS was found to be 51% in the TCGA SARC which was slightly lower than 76% in the NCC. The 5-year PFS was found to be 31% in the TCGA SARC and 28% in the NCC rising to 29% within both datasets. Across both datasets the 5-year DSS was 63% and the 5-year PFS was 29%. The 5-year OS was 43%, decreasing to 23% for RFS in the TCGA SARC data. For the TCGA SARC both baseline survival probabilities using OS and RFS are within expectations for DDLPS.[18,22,361,381] The relationship between patient/tumour characteristics and survival are investigated in subsequent sections.

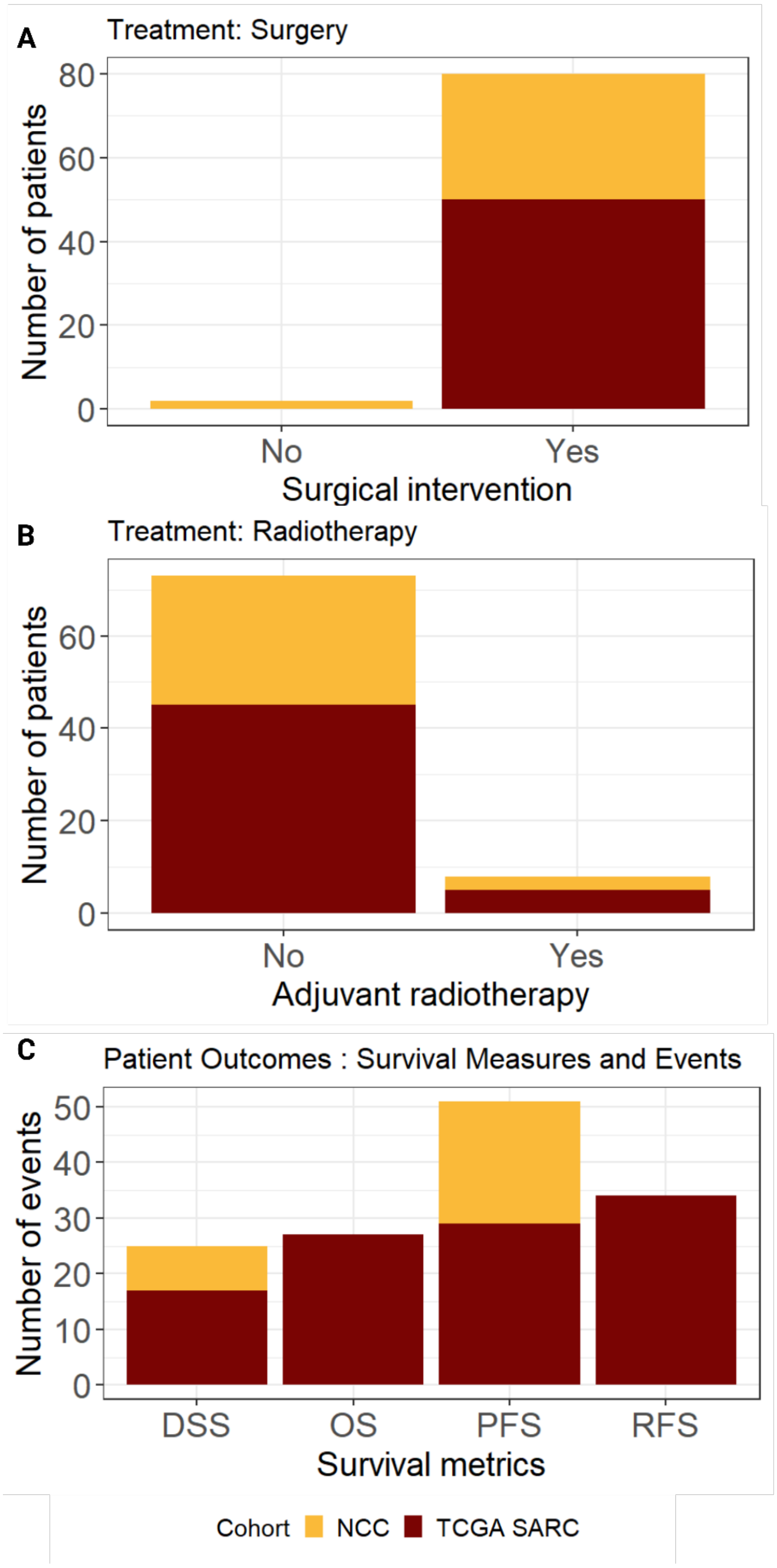


Figure 3.4: Treatment and outcomes. A: The number of patients that received adjuvant radiotherapy. **B**: Events per outcome measure by cohort.

Table 3.2: Patient, Tumour and Outcome characteristics.

	Dataset		
Variable	TCGA (*n = 50*)	NCC (*n = 32*)	Total
Patient characteristics			
Age	*n = 50*	-	-
lq | mean (sd) | uq	34.89 | 63.54 (12.76) | 75.68	-	-
Age (by range)	*n = 50*	*n = 32*	*n = 82*
30-39	2 [4]	1 [3]	3 [4]
40-49	4 [8]	5 [16]	9 [11]
50-59	10 [20]	10 [31]	20 [24]
60-69	19 [38]	7 [22]	26 [32]
70-79	9 [18]	8 [25]	17 [21]
80-89	6 [12]	1 [3]	7 [8]
Gender	*n = 50*	*n = 32*	*n = 82*
Male	33 [66]	26 [81]	59 [72]
Female	17 [34]	6 [19]	23 [28]
Ethnicity/Race/Demographic	*n = 49*	*n = 32*	*n = 81*
White	48 [98]	-	48 [59]
Hispanic	1 [2]	-	1 [2]
Asian (Japanese)	-	32 [100]	32 [39]
Location	*n = 50*	*n = 32*	*n = 82*
Chest wall or back	2 [4]	1 [3]	3 [4]
Extremity, shoulder or girdle	5 [10]	6 [19]	11 [13]
Retroperitoneum or abdomen	43 [86]	25 [78]	68 [83]
Tumour characteristics			
Resection margin	*n = 49*	*n = 27*	*n = 76*
R0	22 [45]	8 [30]	30 [39]
R1	24 [49]	19 [70]	43 [57]
R2	3 [6]	-	3 [4]
FNCLCC Grade	*n = 50*	-	-
1	1 [2]	-	-
2	37 [74]	-	-
3	12 [24]	-	-
Tumour size (cm)	*n = 49*	-	-
lq | mean (sd) | uq	13.0 | 19.2 (9.2) | 25. 0	-	-
Mitotic rate	*n = 50*	-	-
lq | mean (sd) | uq	2 | 7.16 (8.09) | 7	-	-
Mitotic scoring [%]	*n = 50*	-	-
1	38 [76]	-	-
2	7 [14]	-	-
3	5 [10]	-	-
Necrosis score [%]	*n = 50*	-	-
0 (no necrosis)	21 [42]	-	-
1 (Necrosis present <50%)	28 [56]	-	-
2 (Necrosis present >50%)	1 [2]	-	-
Treatment and Outcomes			
Distant recurrence [%]	*n = 49*	-	-
Yes	9 [18]	-	-
No	40 [82]	-	-
Surgical resection [%]	*n = 50*	*n = 32*	*n = 82*
Yes	50 [100]	30 [94]	80 [98]
No	0 [0]	2 [6]	2 [2]
Adjuvant RT [%]	*n = 50*	*n = 31*	*n = 81*
Yes	5 [10]	3 [10]	8 [10]
No	45 [90]	28 [90]	73 [90]
OS	*n = 50*	-	-
Events	23 [46]	-	-
RFS	*n = 50*	-	-
Events	34 [68]	-	-

Variable	TCGA (*n = 50*)	NCC (*n = 32*)	Total
DSS	*n = 47*	*n = 32*	*n = 79*
Events	17 [36]	8 [28]	25 [32]
PFS	*n = 50*	*n = 32*	*82*
Events	29[58]	22 [69]	51 [62]
1-year survival % (95% CI range)			
OS	84 (74-95	-	-
RFS	64(51-79)	-	-
DSS	85(75-96)	100 (100-100)	91 (85-98)
PFS	61(49-76)	84 (73-98)	70 (61-81)
3-year survival % (95% CI range)			
OS	59 (46-77)	-	-
RFS	43(31-61)	-	-
DSS	63(48-81)	94 (85-100)	77 (67-88)
PFS	43(31-61)	33 (20-55)	38 (28-51)
5-year survival % (95% CI range)			
OS	43 (28-65)	-	-
RFS	23 (12-43)	-	-
DSS	51(35-75)	76 (61-95)	63 (51-77)
PFS	31(18-52)	28 (15-51)	29 (19-43)
10-year survival % (95% CI range)			
OS	28 (13-59)	-	-
RFS	*NR*	-	-
DSS	44(27-72)	*NE*	58 (45-75)
PFS	*NR*	*NE*	*NE*

TCGA SARC: The Cancer Genome Atlas Sarcoma project. NCC: National Cancer Center Japan. Percentages for category levels were calculated columnwise and are reported as "[%]". For continuous variables the lower quartile (lq), mean, standard deviation (sd), and upper quartile (uq) were calculated and reported as "lq | mean (sd) | uq". Survival metrics are abbreviated as follows: OS: Overall survival, RFS: Recurrence-free survival, PFS: Progression-free survival, DSS: Disease-specific survival. *NE:* No further events observed for this timepoint. *NR:* No patients at risk for this timepoint.

3.4.3 Survival and clinical covariate analysis

3.4.3.1 TCGA SARC (DDLPS subset)

A total of fourteen clinical covariates and four measures of genomic complexity (taken from TCGA SARC metadata) were considered for survival analysis. Of these 18 variables, 17 variables passed quality control for missingness (<20%) with the site of distant recurrence being removed (***Table 3.3***). Variables were then tested for case separation for each survival outcomes before survival analysis (see subsequent sections). Missingness of 12% (6 of the 50 available) was identified in variables describing genomic complexity features which were tumour ploidy, mutation load, chromosome instability number (CIN) and genome doublings. The variable for new tumour event was found to be co-linear with local (chi squared = 20.33, *p* = *<0.001*) and distant recurrence (chi squared = 6.58, *p* = 0.01) variables and was removed. This left a total of 16 variables to be considered for univariate and multivariate analysis.

Table 3.3: TCGA SARC data missingness.

Variable	Total records	Percentage missingness	Missing QC result
Tumour weight (g)	50	0%	Pass
Age (years)	50	0%	Pass
Previous malignancy	50	0%	Pass
Sex	50	0%	Pass
Tumour size (cm)	49	2%	Pass
Residual tumour	49	2%	Pass
New tumour event	50	0%	Pass
Local recurrence	50	0%	Pass
Distant recurrence	50	0%	Pass
Site of distant recurrence	7 (of 9 maximum entries)	30%	Fail
Mitotic rate	50	0%	Pass
Necrosis score	50	0%	Pass
Multifocal disease	50	0%	Pass
FNCLCC Grade	49	2%	Pass
Tumour ploidy	44	12%	Pass
Genome doublings	44	12%	Pass
CIN	44	12%	Pass
Mutation load	44	12%	Pass

Data missingness was calculated as the percentage of missing records out of the maximum possible. Failing a variable was based on a missingness of ≥20%.

3.4.3.1.1 Overall Survival

Variables passing quality control for missingness were tested for case separation for overall survival (***Table 3.4***) and no variable was removed. Variables were then tested for their association with OS. To note on the difference in *p-values* between figures and text, the *p-values* reported in text are from the cox Proportional Hazards test, whereas the figures (e.g., ***Figure 3.5A*)** reports the log rank test statistic *p-value*. Univariate Cox analysis identified significant associations with the following variables: Distant recurrence (metastasis vs no metastasis, HR = 3.15 (1.28-7.76**), *p=0.012* ; *Figure 3.5A***), mitotic rate (HR = 1.06 (1.02-1.10), ***p=0.003***), FNCLCC grade (grade 2 vs grade 3, HR = 3.24 (1.35-7.75), *p* = **0.008; *Figure 3.5B***), age (in years) at diagnosis (HR = 1.04 (1.00-1.08), *p* = **0.035**), and genome doublings (HR = 2.87 (1.09-7.57), ***p* = 0.033; *Figure 3.5C***). Trend to significance was observed with residual tumour (HR = 2.24 (0.87-5.75), *p = 0.093*) and tumour ploidy score (HR = 1.50 (0.93-2.43), *p =0.099*) (***Table 3.5***).

Table 3.4: Summary of clinical variables within TCGA SARC stratified by vital status.

Variable	Category	Summary (sd)[%]		
		OS censor	OS event	Total
Tumour weight (g)	Mean (sd)	544.4 (339.1)	412.6 (353.7)	483.8 (348.7)
Age (years)	Mean (sd)	60.1 (12.9)	67.6 (11.6)	63.5 (12.8)
Previous malignancy	No	23 [85.2]	20 [87.0]	43 [86.0]
	Yes	4 [14.8]	3 [13.0]	7 [14.0]
Sex	Male	20 [74.1]	13 [56.5]	33 [66.0]
	Female	7 [25.9]	10 [43.5]	17 [34.0]
Tumour size (cm)	Mean (sd)	19.6 (8.3)	18.7 (10.3)	19.2 (9.2)
Residual tumour	R0	16 [59.3]	6 [27.3]	22 [44.9]
	R1/R2	11 [40.7]	16 [72.7]	27 [55.1]
Local recurrence	No	17 [62.9]	10 [43.5]	27 [54.0]
	Yes	10 [37.1]	13 [56.5]	23 [46.0]
Distant recurrence	Distant Metastasis	1 [3.7]	8 [34.8]	9 [18.0]
	New Primary Tumour	1 [3.7]		1 [2.0]
	No	25 [92.6]	15 [65.2]	40 [80.0]
Mitotic rate	Mean (sd)	4.8 (5.9)	9.9 (9.5)	7.2 (8.1)
Necrosis score	0	12 [44.4]	9 [39.1]	21 [42.0]
	1 or 2	15 [55.6]	14 [60.9]	29 [58.0]
Multifocal disease	No	17 [65.4]	13 [59.1]	30 [62.5]
	Yes	9 [34.6]	9 [40.9]	18 [37.5]
FNCLCC Grade	1	1 [3.7]		1 [2.0]
	2	22 [81.5]	15 [65.2]	37 [74.0]
	3	4 [14.8]	8 [34.8]	12 [24.0]
Tumour ploidy	Mean (sd)	2.7 (1.0)	2.7 (1.0)	2.7 (1.0)
Genome doublings	0	15 [65.2]	10 [47.6]	25 [56.8]
	1 or 2	8 [34.8]	11 [52.4]	19 [43.2]
CIN	Mean (sd)	420.6 (294.1)	465.4 (210.9)	441.2 (257.6)
Mutation load	Mean (sd)	62.1 (24.2)	65.2 (25.9)	63.5 (24.8)

Percentages were calculated columnwise and are reported as '[%]'. sd: standard deviation is reported as '(sd)'.
CIN: Chromosome instability number.

Table 3.5: Univariate Cox survival analysis on overall survival outcome measure.

Variable	**Categories**	**Number (sd) [%]**	**HR Univariate (CI, *p* value)**
Tumour Weight	Mean (sd)	483.8 (348.7)	1.00 (1.00-1.00, p=0.343)
Prior malignancy	No	43 [86.0]	-
	Yes	7 [14.0]	0.87 (0.26-2.93, p=0.817)
Sex	Male	33 [66.0]	-
	Female	17 [34.0]	1.68 (0.73-3.83, p=0.221)
Tumour size	Mean (sd)	19.2 (9.2)	1.00 (0.95-1.06, p=0.869)
Residual tumour	R0	22 (44.9)	-
	R1/R2	24 (49.0)	2.24 (0.87-5.75, p = 0.093)
Distant recurrence	No	41 [82.0]	-
	Yes	9 [18.0]	***3.15 (1.28-7.76, p=0.011)***
Mitotic rate	Mean (sd)	7.2 (8.1)	**1.06 (1.02-1.10, p=0.003)**
Necrosis score	0	21 [42.0]	-
	1	29 [58.0]	1.35 (0.58-3.13, p=0.483)
Multifocal disease	No	30 [62.5]	-
	Yes	18 [37.5]	1.73 (0.72-4.17, p=0.223)
FNCLCC grade	2	37 [75.5]	-
	3	12 [24.5]	**3.24 (1.35-7.75, p=0.008)**
CIN	Mean (sd)	441.2 (257.6)	1.00 (1.00-1.00, p=0.228)
Mutation load	Mean (sd)	63.5 (24.8)	1.01 (0.99-1.03, p=0.248)
Age (years)	Mean (sd)	63.5 (12.8)	**1.04 (1.00-1.08, p=0.035)**
Genome doublings	0	25 [56.8]	-
	1/2	19 [43.2]	**2.87 (1.09-7.57, p=0.033)**
Local recurrence	No	27 [54.0]	-
	Yes	23 [46.0]	1.26 (0.55-2.88, p=0.585)
Tumour ploidy	Mean (sd)	2.7 (1.0)	1.50 (0.93-2.43, p=0.099)

Bold text indicates significance. **CI** – Lower and upper 95% confidence intervals. **HR** – Hazards ratio**.** ***P values*** are from Wald test statistics for each variable. Percentages are indicated in [%] and standard deviations are indicated in (sd). Percentages are calculated column-wise per variable. **CIN** – Chromosome instability number. **RT** – Radiotherapy.

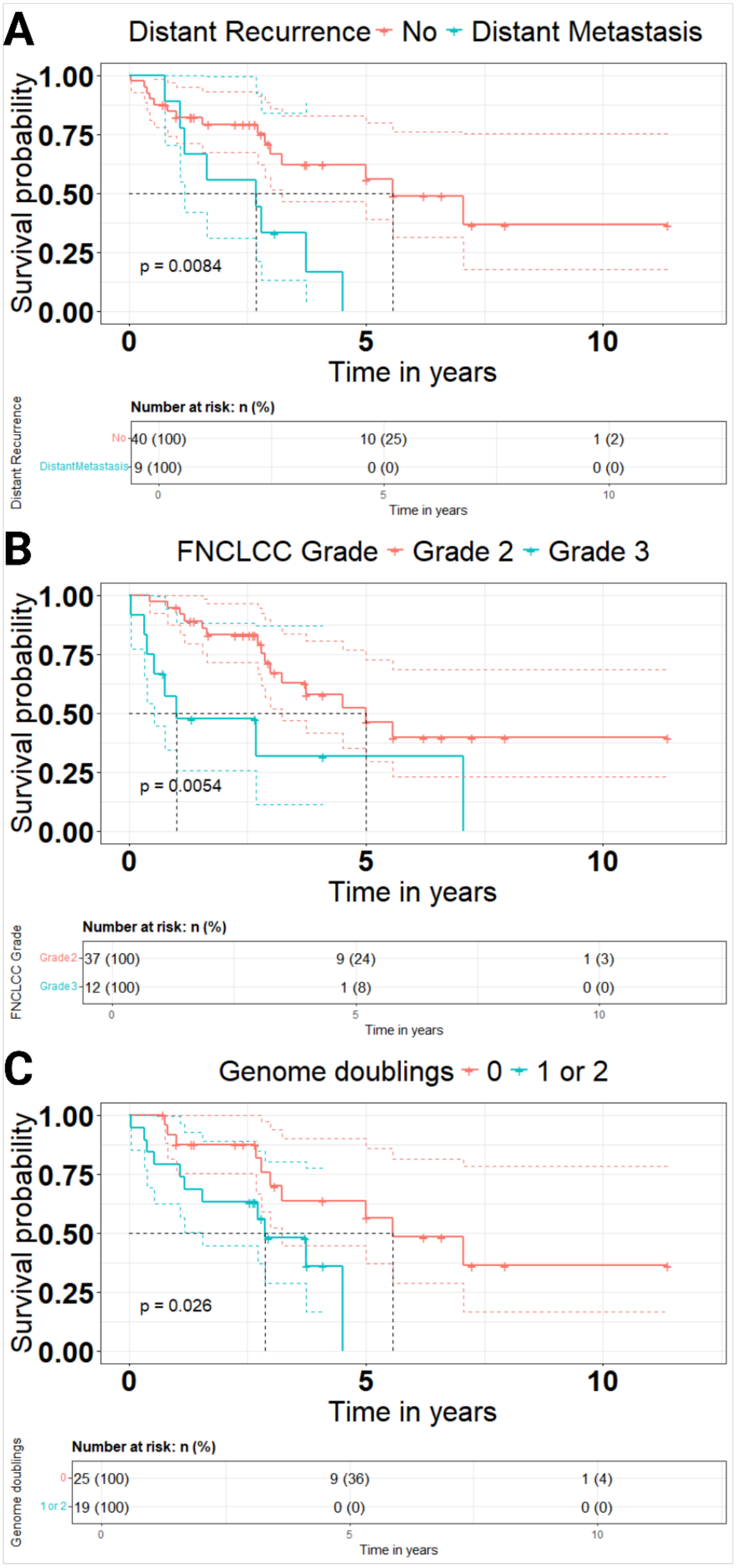


Figure 3.5: Overall Survival Kaplan-Meier plots for **A:** Residual tumour **B:** FNCLCC tumour grade and **C:** Genome Doublings. *P* value is from the logrank test. Fits are coloured according to variable factors. Dotted coloured lines represent the lower and upper 95% confidence intervals for respective plots. Grey dotted lines indicate the median survival times. Points plotted on survival curves represent patient censoring.

Multivariate analysis revealed FNCLCC grade (HR = 6.13 (2.06-18.23, ***p=0.001***) and residual tumour (HR = 4.31 (1.44-12.92, ***p=0.009***) to be significant predictors of overall survival (***Table 3.6***). Both variables predicted increased risk of patient death in higher graded tumours and incomplete resection cases. This model showed low correlation between variables and passed Schoenfeld residuals chi-squared test. The model concordance was found to be 0.726 and the likelihood ratio (LR) test on the global model showed significance (LR = 12.497, ***p* = *0.002***) on two degrees of freedom.

Table 3.6: Multivariate Cox PH results for OS in the TCGA SARC dataset.

Variables	**Categories**	**Number [%]**	**Multivariate Cox**	**Schoenfeld residuals test**	**VIF**
FNCLCC Grade	Grade 2	37 [75.5]	-	0.463, *p* = 0.104	-
	Grade 3	12 [24.5]	**6.13 (2.06-18.23, p=0.001)**		1.35
Residual tumour	R0	22 [44.9]	-	2.65, *p* = 0.496	1.35
	R1/R2	27 [55.1]	**4.31 (1.44-12.92, p=0.009)**		-

The number of patients included in the model was 48 with 22 events and 2 missing values. Concordance = 0.726 (standard error = 0.053), R^2 = 0.229, Likelihood ratio of test = 12.497 (degrees of freedom = 2, *p = 0.002)*. **Bold** text indicates a significant Wald test statistic (**p < 0.05***)*. An event to variable ratio = 11:1. Multivariate Cox results are presented as follows – HR (lower CI – upper CI, p-value). **HR** – Hazards ratio**. CI** – Lower and upper 95% confidence intervals. ***P values*** are from Wald test statistics for each variable. Percentages are indicated in [%] and are calculated columnwise per variable. Schoenfeld residuals testis reported as the Pearson Chi-square and the *p*-value. **VIF** – Variance inflation factor test for co-linearity.

3.4.3.1.2 Recurrence-free survival

For recurrence-free survival (RFS) co-variates describing recurrence events (local recurrence and distant recurrence) were removed as disease recurrence (recorded as relapse) is now the dependent variable. No variables were removed based on case separation between recurrence and non-recurrence (censor) (***Table 3.7***). A total of 14 variables were tested for association with RFS. Univariate analysis (***Table 3.8***) found that genome doublings were significant (0 vs 1 or 2 doublings, HR = 2.15 (1.00-4.58), ***p=0.049***; ***Figure 3.6***). Trend to significance was observed with residual tumour (R0 vs R1/R2, HR = 3.59 (0.68-8.82), *p* = 0.095) and tumour ploidy (HR = 1.40 (0.98-2.01), *p = 0.068*).

Table 3.7: Summary of variables within TCGA SARC stratified by recurrence.

Variable	Category	Summary (sd)[%]		
		Censor	Recurrence	Total
Tumour weight (g)	Mean (SD)	568.1 (377.6)	444.1 (332.6)	483.8 (348.7)
Age (years)	Mean (SD)	58.7 (11.8)	65.8 (12.7)	63.5 (12.8)
Previous malignancy	no	14 [87.5]	29 (85.3)	43 (86.0)
	Yes	2 [12.5]	5 (14.7)	7 (14.0)
Sex	Male	10 [62.5]	23 (67.6)	33 (66.0)
	Female	6 [37.5]	11 (32.4)	17 (34.0)
Tumour size (cm)	Mean (SD)	19.2 (8.4)	19.2 (9.6)	19.2 (9.2)
Residual tumour	R0	12 [75.0]	10 (30.3)	22 (44.9)
	R1/R2	4 [25.0]	23 [69.7]	27 [55.1]
Mitotic rate	Mean (SD)	5.4 (7.2)	8.0 (8.5)	7.2 (8.1)
Necrosis score	0	7 [43.8]	14 [41.2]	21 [42.0]
	1	8 [50.0]	20 [58.8]	28 [56.0]
	2	1 [6.2]		1 (2.0)
Multifocal disease	No	10 [66.7]	20 [60.6]	30 [62.5]
	Yes	5 [33.3]	13 [39.4]	18 [37.5]
FNCLCC grade	1	1 [6.25]		1 [2.0]
	2	12 [75.0]	25 [73.5]	37 [74.0]
	3	3 [19.75]	9 [26.5]	12 [24.0]
Tumour ploidy	Mean (SD)	2.4 (0.8)	2.8 (1.1)	2.7 (1.0)
Genome doublings	0	10 [76.9]	15 [48.4]	25 [56.8]
	1 or 2	3 [23.1]	16 [51.6]	19 [43.2]
CIN	Mean (SD)	387.9 (232.0)	466.3 (268.4)	441.2 (257.6)
Mutation load	Mean (SD)	61.4 (18.9)	64.5 (27.3)	63.5 (24.8)

Percentages were calculated columnwise and are reported as '[%]'. sd: standard deviation is reported as '(sd)'. **CIN**: Chromosome instability number.

Table 3.8: Univariate Cox proportional hazards results

Variables	Categories	Number (sd) [%]	Univariate Cox PH
Tumour weight	Mean (SD)	483.8 (348.7)	1.00 (1.00-1.00, p=0.317)
Previous malignancy	No	43 [86.0]	-
	yes	7 [14.0]	0.79 (0.30-2.06, p=0.626)
Sex	1 (Male)	33 [66.0]	-
	2 (Female)	17 [34.0]	0.84 (0.41-1.73, p=0.636)
Tumour size	Mean (SD)	19.2 (9.2)	1.01 (0.97-1.05, p=0.640)
Residual tumour	R0 (negative)	22 [44.9]	-
	R1/R2	27 [55.1]	1.75 (0.90-3.98, p=0.095)
Mitotic rate	Mean (SD)	7.2 (8.1)	1.02 (0.99-1.06, p=0.211)
Necrosis score	0	21 [42.0]	-
	1 or 2	29 [58.0]	1.18 (0.59-2.33, p=0.644)
Multifocal disease	No	30 [62.5]	-
	Yes	18 [37.5]	1.50 (0.74-3.06, p=0.259)
FNCLCC grade	Grade 2	37 [75.5]	-
	Grade 3	12 [24.5]	1.55 (0.71-3.35, p=0.269)
CIN	Mean (SD)	441.2 (257.6)	1.00 (1.00-1.00, p=0.343)
Mutation load	Mean (SD)	63.5 (24.8)	1.00 (0.99-1.02, p=0.601)
Age (years)	Mean (SD)	63.5 (12.8)	1.01 (0.99-1.04, p=0.305)
Genome doublings	0	25 [56.8]	-
	1/2	19 [43.2]	**2.15 (1.00-4.58, p = 0.049)**
Tumour ploidy	Mean (SD)	2.7 (1.0)	1.40 (0.98-2.01, p=0.068)

43 cases included in model with 30 events. Concordance = 0.637. ***NA*** = Not included in multivariate model. **Bold** text indicates significance in univariate and multivariate testing. **CI** – Lower and upper 95% confidence intervals. **HR** – Hazards ratio**.** ***P values*** are from Wald test statistics for each variable. Percentages are indicated in [%] and standard deviations are indicated in (sd). Percentages are calculated column-wise per variable. A total of 43 patients were included in the model and 7 were removed for missingness. CI – Lower and upper 95% confidence intervals. HR – Hazards ratio. *P value* are from Wald test statistics for each variable.

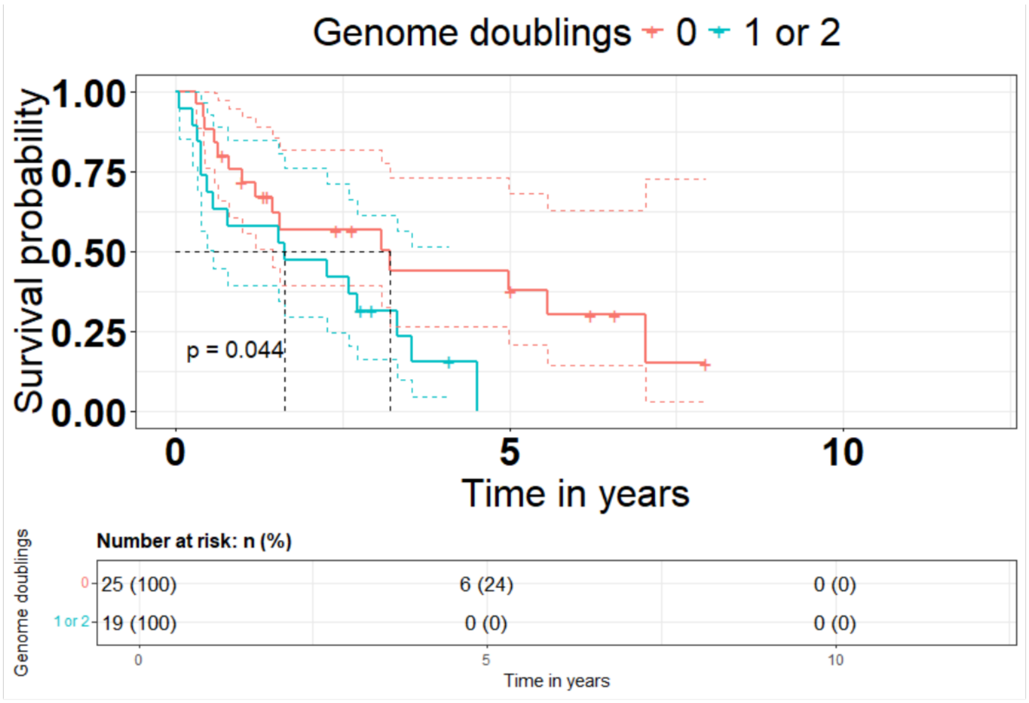


Figure 3.6: RFS Kaplan-Meier plot for genome doublings. *P* value is from the logrank test. Fits are coloured according to variable factors. Dotted coloured lines represent the lower and upper 95% confidence intervals for respective plots. Grey dotted lines indicate the median survival times. Points plotted on survival curves represent patient censoring.

Multivariate analysis both residual tumour (HR = 2.48 (1.09-5.62) ***p=0.030***) and FNCLCC grade (HR = 2.94 (1.07-8.06) ***p=0.036***) were significant for RFS (***Table 3.9***). Patient sex trended towards significance (0.55 (0.23-1.30), *p = 0.171*). Results indicate that higher graded tumours along with incomplete resection of tumours significantly increased risk of disease recurrence (including death).

Table 3.9: Multivariate Cox PH results for RFS in the TCGA SARC dataset.

Variables	Categories	Number [%]	Multivariate Cox PH	Schoenfeld test	VIF
Residual tumour	R0	22 [44.9]	-	0.010, p = 0.92	-
	R1/R2	27 [55.1]	**2.48 (1.09-5.62, p=0.030)**		1.20
FNCLCC Grade	Grade 2	37 [75.5]	-	0.041, P = 0.84	-
	Grade 3	12 [24.5]	**2.94 (1.07-8.06, p=0.036)**		1.55
Sex	Male	33 [66.0]	**-**	0.172, p = 0.85	-
	Female	17 [34.0]	0.55 (0.23-1.30, p=0.171)		1.33

The number of patients included in the model was 48 with 33 events and 2 missing values. Concordance = 0.613 (standard error = 0.059), R^2 = 0.139, Likelihood ratio of test = 7.165 (degrees of freedom = 3, *p = 0.067)*. **Bold** text indicates a significant Wald test statistic (**p < 0.05***)*. An event to variable ratio = 11:1. Multivariate Cox results are presented as follows – HR (lower CI – upper CI, p-value). **HR** – Hazards ratio**.** **CI** – Lower and upper 95% confidence intervals. ***P values*** are from Wald test statistics for each variable. Percentages are indicated in [%] and are calculated columnwise per variable. Schoenfeld residuals testis reported as the Pearson Chi-square and *p-value*. **VIF** – Variance inflation factor test for co-linearity.

3.4.3.1.3 Disease-specific survival

No variable was removed due to case separation (***Table 3.10***). In total 15 variables were tested for their association to disease specific survival (DSS). Univariate analysis revealed a single variable, chromosome instability number (CIN, HR = 1.00 (0.99-1.00), ***p=0.041***) as significant (***Table 3.11***). A strong trend to significance was observed for multifocal disease (HR = 2.49 (0.92-6.73), *p=0.071*).

Table 3.10: Summary of variables within TCGA SARC stratified by DSS event.

Variables	Category	Summary (sd) [%]		
		DSS censor	DSS event	Total
Initial weight	Mean (SD)	493.7 (330.8)	459.4 (373.4)	481.3 (343.1)
Previous malignancy	no	26 [86.7]	14 [82.4]	40 [85.1]
	yes	4 [13.3]	3 [17.6]	7 [14.9]
Sex	1	22 [73.3]	10 [58.8]	32 [68.1]
	2	8 [26.7]	7 [41.2]	15 [31.9]
Age (years)	Mean (SD)	66.1 [11.1]	62.0 (14.2)	64.6 (12.3)
Tumour size (cm)	Mean (SD)	18.3 (8.5)	20.5 (10.7)	19.1 (9.3)
Residual tumour	R0	13 [44.8]	8 [47.1]	21 [45.7]
	R1/R2	16 [55.2]	9 [52.9]	25 [47.8]
Local recurrence	No	17 [56.7]	8 [47.1]	25 [53.2]
	Yes	13 [43.3]	9 [52.9]	22 [46.8]
Distant recurrence	No	21 [72.4]	16 [94.1]	37 [80.4]
	Distant Metastasis	8 [27.6]	1 [5.9]	9 [19.6]
Mitotic rate	Mean (SD)	7.5 (7.3)	7.3 (10.0)	7.4 (8.3)
Necrosis score	0	11 [36.7]	9 [52.9]	20 [42.6]
	1	19 [73.3]	8 [47.1]	27 [57.4]
Multifocal disease	No	22 [75.9]	7 [43.8[	29 [64.4]
	Yes	7 [24.1]	9 [56.2]	16 [35.6]
FNCLCC grade	2	21 [72.4]	13 [76.5]	34 [73.9]
	3	8 [27.6]	4 [23.5]	12 [26.1]
Ploidy	Mean (SD)	2.6 (0.7)	2.9 (1.3)	2.7 (1.0)
Genome doublings	0	15 [55.6]	10 [62.5]	25 [58.1]
	1 or 2	12 [44.4]	6 [37.5]	18 [41.9]
CIN	Mean (SD)	508.4 (294.1)	343.2 (151.8)	448.6 (262.6)
Mutation load	Mean (SD)	63.2 (23.3)	65.4 (25.5)	64.0 (23.8)

Percentages were calculated columnwise and are reported as '[%]'. sd: standard deviation is reported as '(sd)'. **CIN**: Chromosome instability number.

Table 3.11: Univariate Cox PH results for DSS in the TCGA SARC dataset.

Variable	Categories	Number (sd) [%]	HR : Univariate (CI, p vlue)
Tumour weight	Mean (SD)	483.8 (348.7)	1.00 (1.00-1.00, p=0.771)
Previous malignancy	no	43 [86.0]	-
	yes	7 [14.0]	1.40 (0.40-4.95, p=0.601)
Sex	Male	33 [66.0]	-
	Female	17 [34.0]	1.06 (0.40-2.79, p=0.912)
Tumour size	Mean (SD)	19.2 (9.2)	1.02 (0.97-1.08, p=0.452)
Residual tumour	R0	22 [44.9]	-
	R1/R2	27 [49.0]	0.94 (0.35-2.49, p=0.898)
Distant Recurrence	No	40 [81.6]	-
	Distant Metastasis	9 [18.4]	0.23 (0.03-1.74, p=0.155)
Mitotic rate	Mean (SD)	7.2 (8.1)	0.99 (0.92-1.05, p=0.687)
Necrosis Score	0	21 [42.0]	-
	1 or 2	29 [58.0]	0.54 (0.21-1.40, p=0.205)
Multifocal disease	No	30 [62.5]	-
	Yes	18 [37.5]	2.49 (0.92-6.73, p=0.071)
FNCLCC grade	2	37 [75.5]	-
	3	12 [24.5]	0.70 (0.23-2.16, p=0.533)
CIN	Mean (SD)	441.2 (257.6)	**1.00 (0.99-1.00, p=0.041)**
Mutation load (burden)	Mean (SD)	63.5 (24.8)	1.00 (0.98-1.02, p=0.807)
Age (years)	Mean (SD)	63.5 (12.8)	0.98 (0.95-1.03, p=0.458)
Genome doublings	0	25 [56.8]	-
	1/2	19 [43.2]	0.77 (0.28-2.13, p=0.611)
Local recurrence	No	27 [54.0]	-
	Yes	23 [46.0]	1.49 (0.57-3.88, p=0.414)
Tumour ploidy	Mean (SD)	2.7 (1.0)	1.27 (0.79-2.06, p=0.322)

Bold text indicates significance at multivariate level. **CI** – Lower and upper 95% confidence intervals. **HR** – Hazards ratio**. CIN** – Chromosome instability number. ***P values*** are from Wald test statistics for each variable. Percentages are indicated in [%] and standard deviations are indicated in (sd). Percentages are calculated column-wise per variable

For DSS, variables included in the OS/RFS multivariate model (residual tumour and FNCLCC grade) were used to identify concordance between death and disease-specific death (***Table 3.12***). Both residual tumour (HR = 0.87 (0.32-2.35), *p=0.787*) and FNCLCC grade (HR = 0.80 (0.26-2.52), *p=0.706*) were not significant. Re-evaluating variables for DSS revealed that multifocal disease was found to be significant (HR = 3.04 (1.10-8.38) ***p=0.032***) when considering patient age. Tumours that were multifocal predicted an increased risk of disease specific death.

Table 3.12: Multivariate Cox PH results for DSS in the TCGA SARC dataset.

Variables	Categories	Number [%] Mean (sd)	Multivariate Cox	Schoenfeld residuals test	VIF
OS concordance					
FNCLCC grade	2	37 [75.5]	-	0.123, p = 0.73	-
	3	12 [24.5]	0.80 (0.26-2.52, p=0.706)		1.03
Residual tumour	R0	22 [44.9]	-	1.997, p = 0.16	-
	R1/R2	27 [55.1]	0.87 (0.32-2.35, p=0.787)		1.03
DSS model					
Multifocal disease	No	30 [62.5]	**-**	0.668, p = 0.41	-
	Yes	18 [37.5]	**3.04 (1.10-8.38, p=0.032)**		1.05
Age (years)	Mean (sd)	63.5 (12.8)	0.97 (0.93-1.01, p=0.119)	0.008, p = 0.93	1.05

The number of patients included in the **DSS model** was 45 with 16 events and 5 missing values. Concordance = 0.672 (standard error = 0.074), R^2 = 0.118, Likelihood ratio of test = 5.674 (degrees of freedom = 2, *p = 0.059).* **Bold** text indicates a significant Wald test statistic (**p < 0.05***)*. An event to variable ratio = 8:1. Multivariate Cox results are presented as follows – HR (lower CI – upper CI, p-value). **HR** – Hazards ratio**. CI** – Lower and upper 95% confidence intervals. ***P values*** are from Wald test statistics for each variable. Percentages are indicated in [%] and are calculated columnwise per variable. Schoenfeld residuals testis reported as the Pearson Chi-square and *p-value*. **VIF** – Variance inflation factor test for co-linearity.

3.4.3.1.4 Progression-free survival

Distant and local recurrence variables were removed from analysis of progression-free survival (PFS) as progression event (including recurrence) is the dependency. Previous malignancy was removed based on near complete case separation between conditions (***Table 3.13***). Fourteen variables were tested for univariate association to PFS (***Table 3.14***). No variable was found to be significant for PFS. Trend to significant was found in CIN (HR = 1.00 (1.00-1.00), *p* = 0.062) and FNCLCC grade 1.97 (0.91-4.28), *p = 0.085*). Hence, univariate analysis for PFS was non-informative.

Table 3.13: Summary of variables within TCGA SARC stratified by progression event.

Variable		Summary (sd)[%]

		PFS Censor	PFS event	Total
Initial weight (g)	Mean (SD)	433.3 (297.5)	520.3 (382.4)	483.8 (348.7)
Previous malignancy	no	20 [95.2]	23 [79.3]	43 [86.0]
	yes	1 [4.8]	6 [20.7]	7 [14.0]
Sex	1	16 [76.2]	17 [58.6]	33 [66.0]
	2	5 [23.8]	12 [41.4]	17 [34.0]
Age (years)	Mean (SD)	62.9 (11.4)	64.0 (13.8)	63.5 (12.8)
Tumour size (cm)	Mean (SD)	19.9 (6.3)	18.7 (10.9)	19.2 (9.2)
Residual tumour	R0	9 [42.9]	13 [46.4]	22 [44.9]
	R1/R2	12 [57.1]	15 [53.6]	27 [55.1]
Mitotic rate	Mean (SD)	5.5 (6.6)	8.4 (8.9)	7.2 (8.1)
Necrosis score	0	9 [42.9]	12 [41.4]	21 [42.0]
	1 or 2	12 [57.1]	17 [58.6]	29 [58.0]
Multifocal disease	No	15 [71.4]	15 [55.6]	30 [62.5]
	Yes	6 [28.6]	12 [44.4]	18 [37.5]
FNCLCC grade	2	19 [90.5]	18 [64.3]	37 [75.5]
	3	2 [9.5]	10 [35.7]	12 [24.5]
Tumour ploidy	Mean (SD)	2.8 (0.8)	2.7 (1.1)	2.7 (1.0)
Genome doublings	0	8 [47.1]	17 [63.0]	25 [56.8]
	1 or 2	9 [52.9]	10 [37.0]	19 [43.2]
CIN	Mean (SD)	531.3 (327.0)	376.0 (171.1)	441.2 (257.6)
Mutation load	Mean (SD)	60.8 (28.1)	65.5 (22.4)	63.5 (24.8)

Percentages were calculated columnwise and are reported as '[%]'. sd: standard deviation is reported as '(sd)'. CN: Copy number. CIN: Chromosome instability number.

Table 3.14: Univariate Cox PH results for progression free survival.

Variable	**Categories**	**Number (sd) [%]**	**HR Univariate (CI), p value)**

Tumour weight	Mean (sd)	483.8 (348.7)	1.00 (1.00-1.00, p=0.432)
Sex	Male	33 [66.0]	-
	Female	17 [34.0]	1.26 (0.60-2.65, p=0.541)
Tumour size	Mean (sd)	19.2 (9.2)	1.00 (0.96-1.05, p=0.876)
Residual tumour	R0	22 [44.9]	-
	R1/R2	27 [55.1]	0.85 (0.41-1.80, p=0.679)
Mitotic rate	Mean (sd)	7.2 (8.1)	1.03 (0.99-1.07, p=0.199)
Necrosis Score	0	21 [42.0]	-
	1 or 2	29 [58.0]	1.00 (0.48-2.11, p=0.990)
Multifocal disease	No	30 [62.5]	-
	Yes	18 [37.5]	1.63 (0.76-3.48, p=0.210)
FNCLCC grade	2	37 [75.5]	-
	3	12 [24.5]	1.97 (0.91-4.28, p=0.085)
CIN	Mean (sd)	441.2 (257.6)	1.00 (1.00-1.00, p=0.062)
Mutation load	Mean (sd)	63.5 (24.8)	1.01 (0.99-1.02, p=0.503)
Age (years)	Mean (sd)	63.5 (12.8)	1.01 (0.98-1.04, p=0.498)
Genome doublings	0	25 [56.8]	-
	1 or 2	19 [43.2]	0.75 (0.34-1.66, p=0.477)
Tumour ploidy	Mean (sd)	2.7 (1.0)	0.95 (0.61-1.49, p=0.831)

CI – Lower and upper 95% confidence intervals. **HR** – Hazards ratio**. *P values*** are from Wald test statistics for each variable. **CIN –** Chromosome instability number. Percentages are indicated in [%] and standard deviations are indicated in (sd). Percentages are calculated column-wise per variable.

Multivariate analysis using FNCLCC grade and residual tumours identified as being significant from the OS and RFS models were non-significant (***Table 3.15***). Reevaluating variables also revealed no significant variable. A trend towards significance was observed in FNCLCC grade (HR =1.89 (0.84-4.28, p=0.126) in the re-evaluated model.

Table 3.15: Multivariate Cox PH results for PFS in the TCGA SARC dataset.

Variables	Categories	Number [%] Mean (sd)	Multivariate Cox	Schoenfeld residuals test	VIF
OS/RFS concordance					
FNCLCC grade	2	37 [75.5]	-	0.10, p = 0.75	-
	3	12 [24.5]	1.89 (0.84-4.28, p=0.126)		1.03
Residual tumour	R0	22 [44.9]	-	1.36, p = 0.24	-
	R1/R2	27 [55.1]	0.94 (0.43-2.04, p=0.882)		1.03
PFS model					
FNCLCC grade	Grade 2	37 [75.5]	-	- 0.06, p = 0.80	-
	Grade 3	12 [24.5]	1.89 (0.84-4.28, p=0.126)		1.01
Multifocal disease	No	30 [62.50]	**-**	0.57, p = 0.45	-
	Yes	18 [37.5]	1.60 (0.74-3.47, p=0.236)	-	1.01

The number of patients included in the **PFS model** was 47 with 26 events and 3 missing values. Concordance = 0.607 (standard error = 0.056), R^2 = 0.083, Likelihood ratio of test = 4.088 (degrees of freedom = 2, *p = 0.130)*. **Bold** text indicates a significant Wald test statistic (**p < 0.05***)*. An event to variable ratio = 13:1. Multivariate Cox results are presented as follows – HR (lower CI – upper CI, p-value). **HR** – Hazards ratio**. CI** – Lower and upper 95% confidence intervals. ***P values*** are from Wald test statistics for each variable. Percentages are indicated in [%] and are calculated columnwise per variable. Schoenfeld residuals testis reported as the Pearson Chi-square and *p-value*. **VIF** – Variance inflation factor test for co-linearity.

3.4.3.1.5 Summary of TCGA SARC survival analysis

Across all outcomes six variables were found to be significant in univariate analysis and three in multivariate testing (***Table 3.16***). In multivariate analysis, FNCLCC grade and residual tumour were significant for both OS and RFS, although this significance could not be found in DSS or PFS. For univariate analysis, the direction of hazard ratio for variables is concordant across OS, RFS and PFS with similar trends in significance values. For example, FNCLCC grade predicts an increased risk of events across OS, RFS and PFS, and although not significant in PFS, there is a strong trend towards significance, showing some degree of concordance. This concordance is not observed for DSS. A similar trend is also observed at the multivariate level, although there is agreement in PFS and DSS models that multifocal disease increases risk of event, although this does not significantly predict PFS. Age was included in the RFS model in addition to FNCLCC grade and residual tumour as three variables were accepted due to there being 33 RFS events. This meant that an event to variable ratio of 10:1 could be retained for three variables (see **chapter 3.3.5**).

Table 3.16: Summary of Cox proportional hazards testing across patient outcomes.

Univariate Cox proportional hazards				
	Outcome measures (HR, *p value*)			
Variable	**OS**	**RFS**	**DSS**	**PFS**

Distant recurrence	3.15, *p* = ***0.012***	-	0.23, *p = 0.155*	-
Mitotic rate	1.06, *p* **= 0.003**	1.02, *p = 0.211*	0.99, *p = 0.687*	1.03, *p = 0.199*
FNCLCC grade	3.24, *p* = **0.008**	1.55, *p = 0.269*	0.70, *p = 0.533*	1.97, *p = 0.085*
Age	1.04, *p* = ***0.035***	1.01, *p = 0.305*	0.98, *p = 0.458*	1.01, *p = 0.498*
Genome doublings	2.87, *p* = ***0.033***	2.15, *p* = ***0.049***	0.77, *p = 0.611*	0.75, *p = 0.477*
CIN	1.00, *p = 0.228*	1.00, *p=0.343*	1.00, *p* = ***0.041***	1.00, *p = 0.062*
Multivariate Cox proportional hazards				
FNCLCC grade	**6.13, *p=0.001***	**2.94, *p=0.036***	-	1.89, *p=0.126*
Residual tumour	**4.31, *p=0.009***	**2.48, *p=0.030***	-	-
Multifocal disease	-	-	**3.04, p=0.032**	1.60, *p=0.236*

HR – Hazards ratio***. P values*** are from Wald test statistics for each variable. **OS:** Overall survival. **RFS:** Recurrence-free survival. **PFS:** Progression-free survival. **DSS:** Disease-specific survival. **CIN:** Chromosome instability number.

Inspecting the concordance index, the event-to-variable ratio, the model likelihood ratio (LR) and the LR p-value indicates that the OS model is the most robust, followed by the RFS model (***Table 3.17***). However, the RFS model did not reach significance in the global likelihood ratio statistic (7.165, *p* = 0.067) although was near to significance.

Table 3.17: Summary of Cox proportional hazards multivariate models by outcome metric.

Model	Variables (n)	Significant variables (n)	Events (n)	EV ratio	Concordance value	Likelihood ratio statistic	Likelihood ratio p-value
OS	2	2	22	11:1	0.726	12.497	***0.002***
RFS	3	2	33	11:1	0.613	7.165	*0.067*
DSS	2	1	16	8:1	0.672	5.674	*0.059*
PFS	2	0	26	13;1	0.607	4.088	*0.130*

OS: Overall survival. **RFS:** Recurrence-free survival. **PFS:** Progression-free survival. **DSS:** Disease-specific survival. **CIN:** Chromosome instability number. **EV ratio–** event-to-variable ratio. The likelihood ratio statistic is the global statistic on the model.

3.4.3.2 NCC Survival Analysis

Variables for the NCC were assessed for missingness (***Table 3.18***) and case separation between both progression and disease-specific death (***Table 3.19***). No data was found to have failed missingness checks. However, due to case separation for disease specific death, only age range and sex could be considered for survival analysis in NCC. Moreover, due to the number of events in NCC for DSS being only 8 meant that including two variables would give an event to variable ratio of 5, which is deemed not satisfactory here, hence, DSS could not be investigated. Due to case separation in PFS, primary tumour site could not be evaluated in univariate or multivariate analysis (***Table 3.19***). Due to low number of samples between categories of the adjuvant RT variable, the variable was not considered for selection in multivariate analysis but was tested in univariate analysis.

Table 3.18: Data missingness in the NCC dataset.

Variable	Number of entries	Percentage missing	QC result
Age range	32	0%	Pass
Sex	32	0%	Pass
Primary tumour site	31	3%	Pass
Surgery	32	0%	Pass
Residual tumour	27	15%	Pass
Adjuvant radiotherapy	31	3%	Pass
Recurrent or primary tumour	32	0%	Pass

Percentage missingness was calculated as the number of missing entries as a percentage of the number of maximum possible entries.

Univariate analysis revealed that residual tumour (HR = 5.66 (1.27-25.16), ***p=0.023***) was significant and age (70-89) trended towards significance (HR = 0.26 (0.06-1.06), *p=0.060*) (***Table 3.20***). Residual tumour (HR = 7.33 (1.60-33.52), ***p=0.010***) remained significant at the multivariate level (***Table 3.20***). Residual tumour significance is concordant with the significance found for OS and RFS in TCGA SARC.

Table 3.19: Case separation between conditions (progression and disease-specific death) for NCC data.

Variable	levels	DSS censor	DSS event	PFS censor	PFS event
Ag range	30-49	4 [16.7]	2 [25.0]	4 [40.0]	13 [59.1]
	50-69	12 [50.0]	5 [62.5]	6 [60.0]	3 [13.6]
	70-89	8 [33.3]	1 [12.5]	-	6 [27.3]
Sex	Male	19 [79.2]	7 (87.5)	9 [90.0]	17 [77.3]
	Female	5 [20.8]	1 (12.5)	1 [10.0]	5 [22.7]
Residual tumour	R0	8 [36.4]	-	6 [60.0]	2 [11.8]
	R1	14 [63.6]	5 [100.0]	4 [40.0]	15 [88.2]
Adjuvant RT	No	24 [100.0]	4 [57.1]	10 [100.0]	18 [85.7]
	Yes	-	3 [42.9]	-	3 [14.3]
Primary tumour site	Extremity, shoulder, or girdle	6 [26.1]	0 [0.0]	6 [60.0]	0 [0.0]
	Retroperitoneum or abdomen	17 [73.9]	8 [100.0]	4 [40.0]	21 [100.0]
Surgery	Yes	24 [100.0]	6 [75.0]	10 [100.0]	20 [90.9]
	No	-	2 [25.0]	-	2 [9.1]

Percentages were calculated columnwise and are reported as '[%]'. sd: standard deviation is reported as '(sd)'. **RT –** radiotherapy.

Table 3.20: Univariate and multivariate Cox PH results for PFS in the NCC dataset.

Variable	Category	Univariate	Multivariate	Schoenfeld residuals	VIF
Residual tumour	R0	-	-	-	-
	R1	**5.66 (1.27-25.16, p=0.023)**	**7.33 (1.60-33.52, p=0.010)**	0.004, *p* = 0.95	**1.02**
Age	30-49	-	-	-	-
	50-69	0.79 (0.30-2.10, p=0.632)	0.71 (0.23-2.18, p=0.545)	-	1.35
	70-89	0.26 (0.06-1.06, p=0.060)	0.20 (0.05-0.86, p=0.030)	2.040, *p* = 0.36	1.35
Sex	Male	-	-	-	-
	Female	1.74 (0.63-4.80, p=0.287)	-	-	-
Adjuvant RT	No	-	-	-	-
	Yes	1.47 (0.43-5.02, p=0.538)	-	-	-

Th number of patients included in the model was 27 with 17 events and 5 missing values. Concordance = 0.722 (standard error = 0.064), R^2 = 0.400, Likelihood ratio of test = 13.772 (degrees of freedom = 3, *p = 0.003)*. **Bold** text indicates a significant Wald test statistic (**p < 0.05***)*. An event to variable ratio = 8.5:1. Multivariate Cox results are presented as follows – HR (lower CI – upper CI, p-value). **HR** – Hazards ratio**. CI** – Lower and upper 95% confidence intervals. ***P values*** are from Wald test statistics for each variable. Percentages are indicated in [%] and are calculated columnwise per variable. Schoenfeld residuals testis reported as the Pearson Chi-square and *p-value* for multivariate analysis. **VIF** – Variance inflation factor test for co-linearity.

3.4.3.2.1 NCC data: Recurrent tumours

In the NCC, eight of the thirty-two samples were from recurrent tumours. This data was only available for the NCC data and not the TCGA SARC. A univariate and multivariate Cox proportional hazards model was constructed to identify the effect size of recurrent tumours on the PFS outcome. At the univariate level recurrent tumours showed increased risk of PFS event (HR = 2.498 (1.041-5.994), ***p* = 0.04**) but did not in multivariate analysis (***Table 3.21***). There were no significant differences between recurrent and primary tumours when corrected for multivariate analysis.

Table 3.21: Univariate and multivariate results for DSS with recurrent tumours

Variable	**All**	**PFS (HR, (CI), p)**	
		Univariate	**Multivariate**
Residual tumour	R0	-	-
	R1	**5.66 (1.27-25.16, p=0.023)**	**4.72 (1.01-22.15, p=0.049)**
Primary or Recurrent	Primary	-	-
	Recurrent	**2.50 (1.04-5.99, p=0.040)**	1.66 (0.61-4.55, p=0.324)

For PFS the number of patients included in the multivariate model was 27 with 17 PFS events and 5 missing values. Concordance = 0.656 (standard error = 0.062), R^2 = 0.272, Likelihood ratio of test = 8.563 (degrees of freedom = 4, *p = 0.014)*. The event-to-variable ratio was 8.5. **Bold** text indicates a significant wald test statistic (**p < 0.05***)*. **CI** – Lower and upper 95% confidence intervals. **HR** – Hazards ratio**. *P values*** are from Wald test statistics for each variable.

3.4.3.3 Comparing TCGA SARC and NCC datasets

3.4.3.3.1 Comparing PFS models

Concordant variables were taken from both the NCC and TCGA SARC DDLPS subset and used to generate separate PFS models (***Table 3.22***).[64,114] PFS was chosen for reasons outlined in **Chapter 3.4.3.2** (***Table 3.19***) despite PFS model failing to retrieve significance in TCGA SARC, DSS was not suitable for comparing between models due to complete case separation observed in the residual tumour variable. As compared to the NCC PFS model, no variable in the TCGA SARC data reached significance (reflecting results in **chapter 3.4.3.1.4; *Table 3.22***), although a concordant trend can be observed for both variables used. R1/R2 margins increases risk, although this factor is higher and significant in NCC data (HR = 7.33 (1.60-33.52), ***p* = 0.010).** However, a trend to significance is observed within the TCGA SARC data for patients with higher tumour cell content post-surgery (HR = 1.92 (0.86-4.29), *p = 0.113*). The difference in HRs is likely to be explained by the lower number of events available for the NCC data and may be an indication of data overfitting with proportionally larger confidence intervals as compared to the TCGA SARC data.

Table 3.22: Comparison of PFS models in NCC and TCGA SARC using concordant variables.

		NCC PFS model		TCGA SARC concordant PFS model	
Label	**Categories**	**Number**	**HR (multivariate)**	**Number**	**HR (multivariate)**
Residual tumour	R0	8 (29.6)	-	22 (44.9)	-
	R1/R2	19 (70.4)	7.33 (1.60-33.52, p=0.010)	27 (55.1)	1.92 (0.86-4.29, p=0.113)
Age (range)	30-49	6 (18.8)	-	6 (12.0)	-
	50-69	17 (53.1)	0.71 (0.23-2.18, p=0.545)	29 (58.0)	1.14 (0.33-4.00, p=0.835)
	70-89	9 (28.1)	0.20 (0.05-0.86, p=0.030)	15 (30.0)	0.95 (0.26-3.53, p=0.938)

Bold text indicates a significant Wald test statistic (**p < 0.05***)*. Multivariate Cox results are presented as follows – HR (lower CI – upper CI, p-value). **HR** – Hazards ratio**. CI** – Lower and upper 95% confidence intervals. ***P values*** are from Wald test statistics for each variable. Percentages are indicated in [%] and are calculated columnwise per variable. **NCC PFS model:** The number of patients included in the model was 27 with 17 events and 5 missing values. Concordance = 0.722 (standard error = 0.064), R^2 = 0.400, Likelihood ratio of test = 13.772 (degrees of freedom = 3, *p = 0.003)*. **TCGA SARC PFS model:** The number of patients included in the model was 49 with 29 events and 1 missing value. Concordance = 0.589 (standard error = 0.054), R^2 = 0.054, Likelihood ratio of test = 2.726 (degrees of freedom = 3, *p = 0.436)*. **NCC –** National Cancer Center Japan, **TCGA SARC –** The Cancer Genome Atlas Sarcoma Project.

To further compare the models, model performance indicators were compared (***Table 3.23***) showing that the PFS model performs better in the NCC dataset reaching a significant likelihood ratio (13.772, ***p* = 0.003**), although the event-to-variable ratio is lower in the NCC at 8.5:1 as compared to the TCGA SARC with 14.5:1. This, paired with the proportionally higher confidence intervals observed in NCC data (***Table 3.23***) indicates that the confidence of model validity is higher in TCGA SARC data. Schoenfeld residuals and VIF indicated no breaches.

Table 3.23: Cox PH model performance indicators for concordant models in NCC and TCGA SARC data.

Performance indicators	NCC	TCGA SARC
Concordance value	0.722	0.589
Likelihood ratio statistic	13.772 (*p* = 0.003)	2.726 (*p = 0.436*)
Number of events	17	29
EV ratio	8.5:1	14.5:1

EV – even-to-variable. **NCC –** National Cancer Center Japan, **TCGA SARC –** The Cancer Genome Atlas Sarcoma Project.

3.4.3.3.2 Combined data analysis

Concordant data in the TCGA SARC and NCC were then combined and analysed together.[114] (see section 3.3.2). Variables available in both datasets were assessed for missingness (***Table 3.24***). Primary or recurrent labels were not available for TCGA SARC, and the variable was removed for this reason. DSS data was not available for three TCGA SARC patients leaving 79 patients. Due to near complete case separation observed for both DSS and PFS in the primary tumour site variable, it was removed from subsequent survival analysis (***Table 3.25***).

Table 3.24: Variable missingness in TCGA SARC and NCC datasets

Label	Entries recorded	Percentage missingness	QC result
Age range	82	0%	Pass
Sex	82	0%	Pass
Site	82	0%	Pass
Surgery	82	0%	Pass
Residual tumour	76	7%	Pass
Adjuvant radiotherapy	81	1%	Pass
Recurrent or primary	32	61%	Fail
Percentages are calculated columnwise and are reported as "[%]" to 1.d.p.			

Univariate analysis (***Table 3.26***) revealed that the cohort (HR = 2.79 (1.14-6.84), **p=0.025; *Figure 3.7***) from which data was retrieved significantly predicted DSS outcome where tumours in the TCGA SARC had an increased risk of DSS event. Residual tumour showed a trend towards significance where incomplete resection showed increased risk for event (2.30 (0.85-6.26) p=0.102). For PFS the residual tumour (HR = 2.35 (1.19-4.63), ***p=0.014***; ***Figure 3.8***) and whether the patient received adjuvant radiotherapy (HR = 2.07 (0.92-4.61) p=0.077) trended towards significance.

Table 3.25: Case separation for DSS and PFS events in TCGA SARC and NCC datasets

label	levels	DSS censor	DSS event	PFS censor	PFS event
Age range	30-49	9 [16.7]	3 [12.0]	3 [9.7]	9 [17.6]
	50-69	30 [55.6]	15 [60.0]	16 [51.6]	30 [58.8]
	70-89	15 [27.8]	7 [28.0]	12 [38.7]	12 [23.5]
Sex	Male	40 [74.1]	17 [68.0]	21 [67.7]	38 [74.5]
	Female	14 [25.9]	8 [32.0]	10 [32.3]	13 [25.5]
Primary tumour site	Other	12 [22.2]		11 [35.5]	3 [5.9]
	Retroperitoneum or abdomen	42 [77.8]	25 [100.0]	20 [64.5]	48 [94.1]
Surgery	Yes	54 [100.0]	23 [92.0]	31 [100.0]	49 [96.1]
	No		2 [8.0]		2 [3.9]
Residual tumour	R0	24 [47.1]	5 [22.7]	19 [63.3]	11 [23.9]
	R1/R2	27 [52.9]	17 [77.3]	11 [36.7]	35 [76.1]
Adjuvant RT	No	51 [94.4]	19 [79.2]	30 [96.8]	43 [86.0]
	Yes	3 [5.6]	5 [20.8]	1 [3.2]	7 [14.0]
Cohort	NCC	24 [44.4]	8 [32.0]	10 [32.3]	22 [43.1]
	TCGA SARC	30 [55.6]	17 [68.0]	21 [67.7]	29 [56.9]

Percentages were calculated columnwise and are reported as '[%]'. **RT** – radiotherapy.

Table 3.26: Univariate analysis of PFS and DSS in the TCGA SARC and NCC datasets

Variables	Categories	Number [%]	DSS	PFS
Cohort	NCC	32 [39.0]	-	
	TCGA SARC	50 [61.0]	**2.79 (1.14-6.84, p=0.025)**	1.09 (0.63-1.90, p=0.758)
Residual tumour	R0	30 [39.5]	-	
	R1/R2	46 [60.5]	2.30 (0.85-6.26, p=0.102)	**2.35 (1.19-4.63, p=0.014)**
Adjuvant RT	No	73 [90.1]	-	
	Yes	8 [9.9]	2.24 (0.83-6.07, p=0.113)	2.07 (0.92-4.61, p=0.077)
Gender	Male	59 [72.0]	-	
	Female	23 [28.0]	1.76 (0.74-4.16, p = 0.201)	0.99 (0.53-1.86, p=0.975)
Age range	30-49	12 [14.6]	-	
	50-69	46 [56.1]	2.02 (0.46-8.89, p = 0.354)	0.85 (0.40-1.80, p=0.668)
	70-89	24 [29.3]	3.30 (0.66-16.47, p = 0.145)	0.54 (0.23-1.30, p=0.170)

Bold text indicates a significant Wald test statistic (**p < 0.05**). **CI** – Lower and upper 95% confidence intervals. **HR** – Hazards ratio**. RT –** radiotherapy**. *P values*** are from Wald test statistics for each variable. Percentages are indicated in [%] and are calculated columnwise per variable.

Multivariate analysis (***Table 3.27***) revealed that for DSS cohort (HR = 3.09 (1.19-8.01), ***p=0.020***) was significant where an increased risk was observed in the TCGA SARC. For PFS, residual tumour (HR = 2.29 (1.16-4.52)**,** ***p=0.017***) was found to be significant, where incomplete resection predicted a higher risk of disease progression. Patients that received adjuvant radiotherapy (HR = 2.10 (0.88-5.00), p=0.093) showed prediction of increased risk of disease progression, although the association was not significant. The DSS model Concordance was 0.664 and the model was significant (LR = 8.311, *p = 0.016*). The PFS model concordance was found to be 0.597, and the model was found to be significant (LR = 9.256, *p = 0.010*). In the Schoenfeld residuals test, cohort was found to border significance (chi-square = 3.773, *p* = 0.052) but was not found to violate proportional hazards in univariate testing.

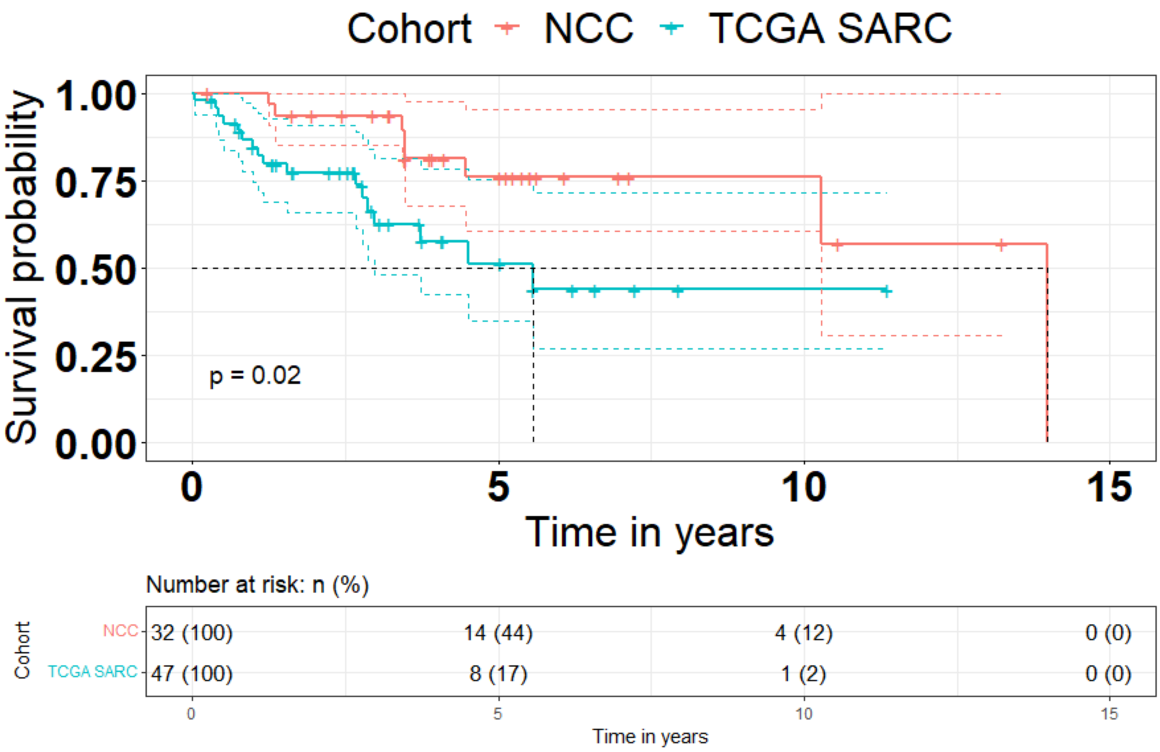


Figure 3.7: **Kaplan-Meier plot for DSS stratified by data cohort.** Fits are coloured according to variable factors. Dotted coloured lines represent the lower and upper 95% confidence intervals for respective plots. Grey dotted lines indicate the median survival times. Points plotted on survival curves represent patient censoring.

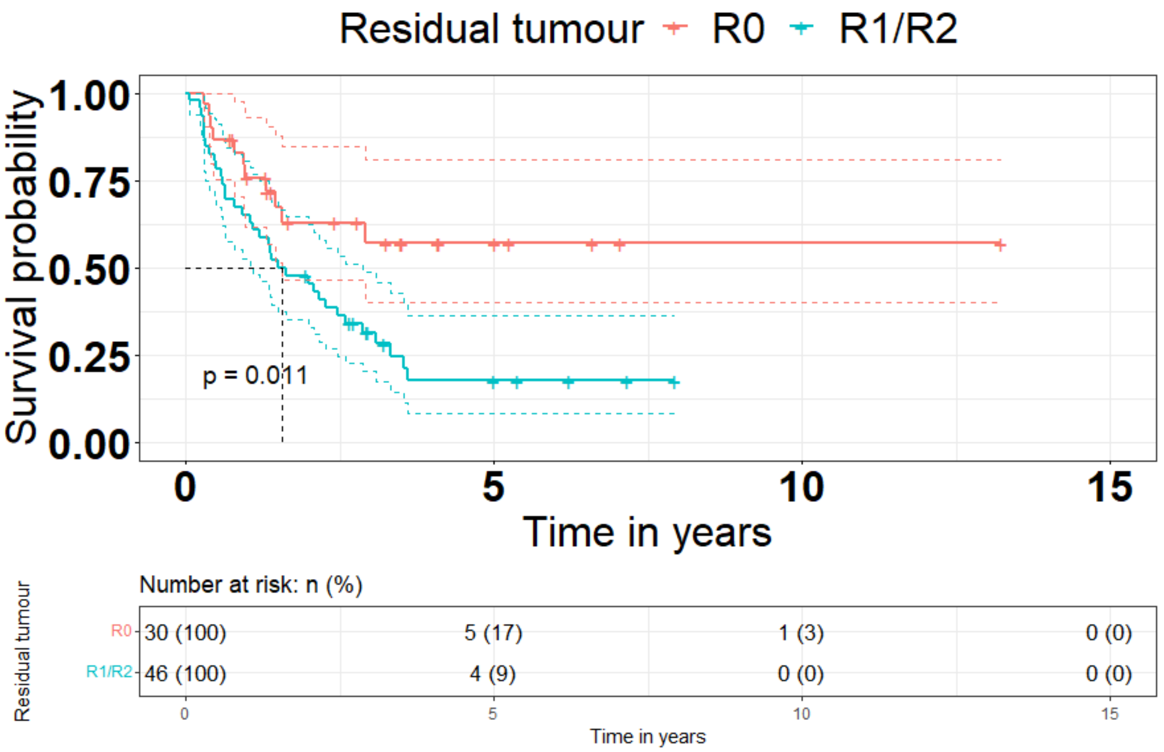


Figure 3.8: Kaplan-Meier plot for PFS stratified by residual tumour. Fits are coloured according to variable factors. Dotted coloured lines represent the lower and upper 95% confidence intervals for respective plots. Grey dotted lines indicate the median survival times. Points plotted on survival curves represent patient censoring.

Table 3.27: Multivariate analysis of PFS and DSS in the TCGA SARC and NCC datasets.

Variables	Categories	Number [%]	DSS	Schoenfeld test	VIF	PFS	Schoenfeld test	VIF
Cohort	NCC	32 (39.0)	-	-	-	-	-	-
	TCGA SARC	50 (61.0)	**3.09 (1.19-8.01, p=0.020)**	3.773 , *p* = 0.052	**1.02**	**-**	**-**	**-**
Residual tumour	R0	30 (39.5)	-	-	-	-	0.514, *p = 0.47*	1.00
	R1/R2	46 (60.5)	-	**-**	**-**	**2.29 (1.16-4.52, p=0.017)**	**-**	**-**
Adjuvant radiotherapy	No	73 (90.1)	-	-	-	-	-	-
	Yes	8 (9.9)	2.13 (0.78-5.77, p=0.139)	0.952, *p* = 0.329	1.02	2.10 (0.88-5.00, p=0.093)	0.145, *p* = 0.70	1.00

For the **DSS** model the number of patients included in the model was 78 with 24 events and 4 missing values. Concordance = 0.664 (standard error = 0.045), $R^2 = 0.101$, Likelihood ratio of test = 8.311 (degrees of freedom = 2, *p = 0.016)*. The event-to-variable ratio was 12:1 for DSS. For the **PFS** model the number of patients included was 76 with 46 events. Concordance = 0.597 (standard error = 0.039), $R^2 = 0.115$, Likelihood ratio of test = 9.256 (degrees of freedom = 2, *p = 0.010)*. The event-to-variable ratio was 23:1 for PFS. **Bold** text indicates a significant Wald test statistic (**p < 0.05***)*. **CI** – Lower and upper 95% confidence intervals. **HR** – Hazards ratio***. P values*** for HRs are from Wald test statistics for each variable in the model. Schoenfeld test is reported as the Pearson Chi-square and the associated p value (Chi-square, *p*). **VIF –** Variance inflation factor. Percentages are indicated in [%] and are calculated columnwise per variable.

3.5 Discussion

TCGA SARC and NCC have similar sample sizes (32 vs 50) and enough samples for WGCNA.[64,114] This presents an opportunity to identify preserved co-expression patterns among two datasets that pass the WGCNA recommended sample size.[252] The TCGA SARC is among the most comprehensive projects identified containing 50 DDLPS samples that underwent pathological review by an expert panel.[64] This review included an assessment of available pathology reports that included immunohistochemical (IHC) and molecular diagnostics (chromosome 12q15 copy number gain). Positive stains in IHC for MDM2/CDK4 and 12q extrachromosomal structures on cytogenetic reports are considered diagnostic indicators for WD/DDLPS.[29,60,61]

Other gene expression datasets found, whilst not chosen for WGCNA or validation, are suitable for other analysis. For example, GSE30929[332] and GSE159659[333] are microarray datasets with paired DDLPS-adipose or paired DDLPS-WDLPS samples, that can be used for differential gene expression between DDLPS and WDLPS and normal. Recently, GSE30929 and GSE159659 have been used in an integrated WGCNA approach in identifying gene expression biomarkers for DDLPS (compared to comparator tissue).[291] Differential gene expression analysis conducted using the GSE30929 cohort revealed significant upregulation of cell cycle genes in DDLPS compared to fatty acid metabolic processes and AMPK signalling processes in WDLPS. In their study, differentially expressed genes were cross-referenced with co-expressed genes that showed the strongest association with being a DDLPS sample, hence screening out genes that are differentially co-expressed and upregulated in DDLPS. It is expected that integrating results from differential gene expression analysis from GSE159659 or GSE30929 are likely to highlight modules pertaining to cell cycling.

Analysis of patient, tumour, and survival characteristics available within the TCGA SARC and NCC indicate that these datasets are representative of DDLPS as a disease, mirroring the conclusions made from NCC data generators.[114] These results also reflect what is observed in studies using larger datasets such as the demographic and clinical data available in the Survival Epidemiology, and End Results (SEER) database and the Canadian Institute for Clinical Evaluative Sciences (ICES) and other studies of similar scale.[18,22,361,381] The only minor differences is that in the TCGA SARC and NCC data, patient sex shows a slight predilection towards male patients, which is not reflected in these wider studies, although it is generally accepted now that DDLPS occurs at a higher rate in males than females.

FNCLCC grade and residual tumour were found to be significant predictors for OS and RFS. These variables are cited as being significant for LPS, and variables known to have clinicopathological importance were prioritised in the variable selection strategy (backward elimination).[84,93,107,360,361,382,383] Whilst this may have provided a bias towards known clinicopathological variables, it was desired to find a balance between a model that fits the data well but also one that is generalisable to the clinical understanding of DDLPS.[365] For OS and RFS multivariate models this was successful where there was concordance in the explanatory variables, and the results of both are in-line with clinical understanding.

No significant associations could be identified for PFS in the TCGA SARC, although some concordance in the trend of hazards ratios and p-values were noted. PFS is an important metric used in liposarcoma for clinical trial endpoints and may have been a measure to explore further downstream in target identified (e.g., to identify associations between gene expression and PFS). However, results here indicate this metric for the TCGA SARC should not be used for further analysis. Out of the four metrics analysed, performance indicators show that the OS Cox PH model should be brought forward for use in integration with WGCNA.

Cohort was found to be the only significant factor for DSS when assessing concordant data in TCGA SARC and NCC, both in univariate and multivariate analysis. Treatment guidelines for DDLPS are similar where surgery is the mainstay, with radio-or-chemo-therapy being considered for advanced or metastatic disease.[60] This difference is likely due to the TCGA SARC being a slightly larger sample size and sampling. Presence of residual tumour at resection margins were found to predict PFS in the NCC alone, and in the combined analysis of TCGA SARC and NCC data. However, when comparing models generated from concordant data in the TCGA SARC cohort separately, and in-line with the analysis of TCGA SARC data, no variable was shown to significantly predict PFS. Cohort was not a significant factor for PFS, and 5-years survival probabilities were found to be similar, where PFS events were more equal occurring between the cohorts. In addition, important clinical variables such as FNCLCC grade, distant recurrence, and multifocality where not listed in the NCC and hence could not be assessed. Having these concordant variables would have allowed for model validation between TCGA SARC and NCC, which will be a limitation moving forward.

Recurrent tumours in the NCC were not found to predict PFS or DSS when corrected for multivariate but was significant at the univariate level. Primary and recurrent tumour expression should be compared prior to use as a validation cohort for co-expression patterns.

The largest limitation in these analyses is the sample size. Not only does the small sample size limit statistical power and the generalisation of results to other studies, it also limits the margins by which robust exploratory data analysis can be conducted. For multivariate survival analyses it is suggested to keep a high event-to-variable ratio.[365] With restricted sample size (and hence small event numbers) it meant that multivariate models had to be kept small meaning important variables may have been excluded due to statistical power concerns. Despite this limitation however, results from survival analysis presented clinical features that are concordant with the literature. Whilst generally, missingness was not an issue in TCGA SARC, except for a few variables such as site of distant recurrence, and the measures of genomic complexity, with a higher sample number it may have been possible to impute values for data completeness.

Another limitation was for the distant recurrence variable. Distant metastasis is a known factor for DDLPS survival, and it was hoped that this would be a feature that could be investigated in WGCNA analysis. The clinical annotations for distant recurrence available for the TCGA SARC were reported as not-applicable "*NA*", "New Primary Tumour" or "Distant Metastasis". It was not possible, within the remits of the project, to elucidate whether "*NA*" was attributed to no distant recurrence event or whether distant recurrence was inspected for those patients. For the analysis in this chapter "*NA"* was taken as no distant recurrence event. Hence, it is possible that more patients went on to metastasise, although, it should be noted that the rates of metastasis observed in the TCGA SARC cohort matches the ~25% generally accepted in the literature.[36,104,108]

Furthermore, a conditional backwards elimination approach was used for variable selection. This method carries the limitations of overfitting data and provides vulnerability to variable co-linearity.[365] To reduce this limitation, the p-value for the stopping criteria was set at 0.2, and variables known to be important where questioned for removal in instances where they violated the criteria. This was done to avoid removing important variables and making a model that only describes the TCGA SARC data. Having a model that was generalisable is beneficial to the project aims given that it is sought to identify clinical targets for DDLPS.

Finally, the data used here was from secondary sources, often being multi-institutional, from non-UK sources and conducted up to ~15 years prior to this project start date. This means that the clinical practice may have differed slightly between patients and institutions, and that acquiring specific knowledge on variables was more challenging and the scope to gain addition data was limited.

In summary, dataset searching has identified suitable datasets that can be used in a discovery-validation strategy for WGCNA, and clinical data analysis has identified significant variables in the TCGA SARC data that can be integrated with subsequent WGCNA analysis to identify modules pertinent to disease.

Chapter 4 Investigating the relationship between modularity and robustness in the DDLPS GCN

4.1 Introduction

As outlined in detail in **Section 1.3.3.1**, WGCNA is a systematic approach to describe the patterns of gene correlation across different samples. Briefly, the main steps of the pipeline are: Data preparation, sample clustering, network construction, module detection, ME calculation, and then relating modules to external traits (see **section 1.2.3** and **section 2.14**)[234,251,253,254]

Each of these steps is critical and requires careful consideration before progressing to the next phase. For instance, in step 2, failure to exclude outliers can introduce greater heterogeneity in sample gene expression profiles, which in turn may affect gene connectivity values by altering the correlation coefficients in step 3. Consequently, this impacts upon how accurately hierarchical clustering and the dynamic tree cutting algorithm assigns genes to modules in step 4.[254] Sample outlier detection commonly employs methods such as clustering algorithms (for instance, K-means or hierarchical clustering), and principal component analysis (PCA). Each of these methods brings its own strengths and weaknesses.[346,384-386] The use of topological measures from a signed weighted sample network to identify outlying samples or groups is more contextually robust.[346] Properties assessed include scaled connectivity, which measures the overall connection strength between a given sample and all other samples in the network, and the clustering coefficient, which quantifies the connection strength among a sample's neighbours.[346,387] These network metrics can provide important insights into the relationship between samples.[387]

Another key consideration for WGCNA is that during the network construction phase (step 3), the adjacency data should exhibit an approximate scale-free topology.[234] The standard approach to verify this involves comparing the degree distribution of the data to a power-law model. This comparison is typically done by calculating the fitting index value (R^2) from a linear fitting model on a frequency plot of gene connectivity, which has been transformed to a log scale. Scale-free networks are non-homogenous in their connections, where hub genes are densely connected and easily distinguished. However, these networks are relatively rare, and networks can also follow other degree distributions, including an exponential distribution.[388,389] These are typically attributed to homogenous graphs where all nodes have approximately uniform connections and hubs are less separable.[390-392] Networks following an exponential distribution have been particularly useful in describing food webs, which are networks of 'who eats whom' in an ecological community, but can also be applied to molecular networks, such as protein-protein and gene regulatory networks.[389,392,393] Although, the majority of molecular networks have been shown to approximate a scale-free topology, an exponential distribution remains an alternative possibility. Therefore, an important QC step is to test whether the data is best described by a power-law distribution, indicating clear, distinguishable hubs, or an exponential distribution where gene connectivity is more homogenous.[394,395]

After network construction, WGCNA uses the hclust function from the Stats R package[321] to perform agglomerative hierarchical clustering and an iterative, top-down approach to cut branches of the dendrogram to define gene clusters (step 4).[254] The choice of linkage method in hierarchical clustering (***Figure 4.1***) and sensitivity parameter of the dynamic tree cut function can significantly impact the size and number of resulting clusters. Therefore, careful consideration and validation of the results are necessary at this phase. To determine the optimal method, assessment of intramodular connectivity (IMC), a measure of the cumulative connection strength a given module gene has with all other module genes, and the relationship between IMC and gene significance (GS) metrics are useful in this regard.[234] Biologically meaningful gene connectivity distributions will exhibit an association with GS scores, indicating which modules encapsulate features of biological importance.[256,387] Here standard validation approaches (e.g., Dunn Index, cophenetic correlation) would not be optimal amid many clusters where there is high dimensionality, noise and for a number of clusters, low separability using distance or similarity metrics.[396] To minimise noise, the strategy in WGCNA is to use metrics (e.g., IMC) which summarise the module as a whole and focus on modules with biological relevancy (high correlation between IMC and GS).[234] However, this strategy necessitates the proposal of suitable GS measures.

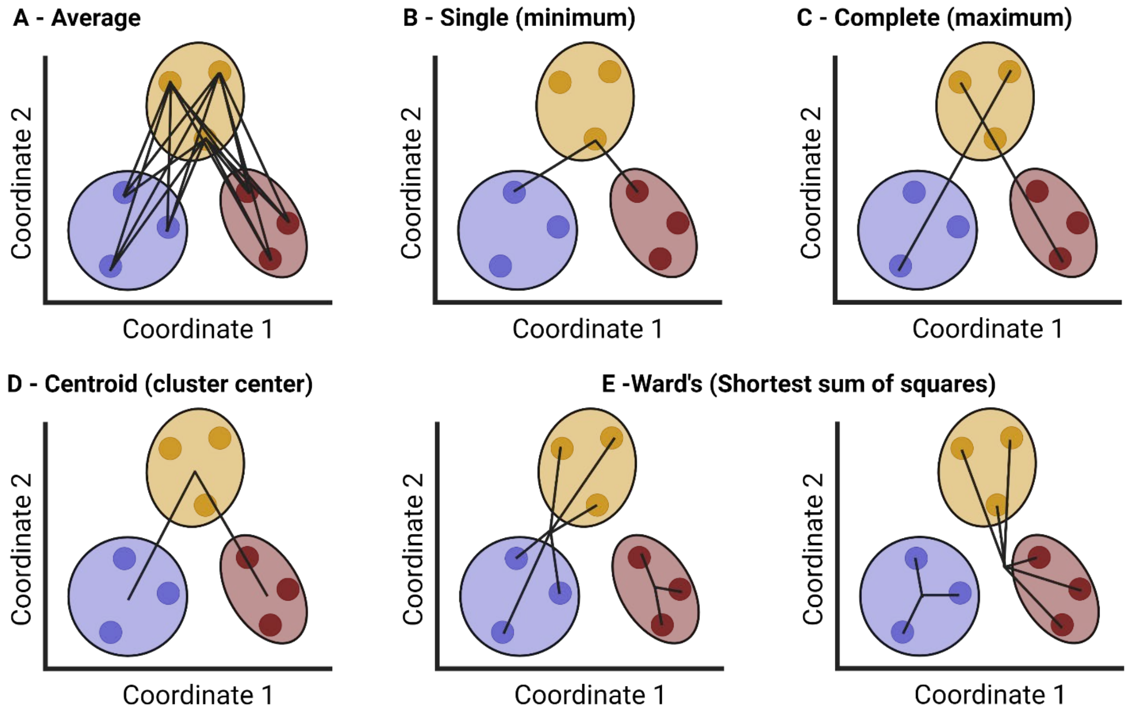


Figure 4.1: Demonstration of different linkage methods. **A** Average linkage takes the mean of all distances between two clusters. **B** Single linkage takes the smallest (minimum) distance between two clusters. **C** Complete linkage takes the largest (maximum) distance between two clusters. **D** Centroid linkages takes the center (mean distance within clusters) of between clusters. **E** Ward's linkage method calculates the error sum of squares of each cluster and chooses the cluster with the smallest value.

Results from the survival analysis (see **Section 3.4.3.1**) identified a multivariate model for overall survival within the TCGA SARC cohort specific to DDLPS. Using this model, the expression values for each gene will be used to construct gene-level multivariate models. The correlation of a gene's expression pattern with the deviance residuals from these models will be used as a GS measure, indicating an association with poor outcome (highlighting patterns where gene upregulation corresponds to increased risk). Another GS measure based on differential gene expression between DDLPS and WDLPS using the GSE159659 dataset, identified in **Section 3.4.1**. and discussed in **Section 1.1.5.** Differential expression results from this data have been used in previous studies implementing WGCNA[249,291] in a strategy to screen-out genes based on the intersect between co-expression modules with high DDLPS sample correlations and genes up and down regulated in DDLPS. These differential expression results were deemed suitable for highlighting modules with genes differentially expressed in DDLPS.

Two studies conducted in gastric[248] and breast[310] cancers, adopted an approach utilising data from the Cancer Dependency MAP (DepMap) project from the Broad Institute.[313] The DepMap portal provides inferred gene dependency and cancer vulnerability information from

CRISPR-Cas9 knockout studies on various cancer cell lines, including three DDLPS cell-lines (LPS141, LPS510 and LPS853) derived from DDLPS patients.[150,314] A limitation for the CHRONOS algorithm is that inferred dependencies may be incorrect as a gene with a complete and efficient knock-out for a weak scoring essential gene may score higher relative to a gene that should attain higher essentiality scores because the knockout was incomplete and showed poor efficiency.[315] CHRONOS also assumes a linear relationship between a given genes knockout and cell viability, assuming that cell growth rate is constant over time. Despite these limitations, this information can provide valuable insights into the dependencies of malignant cells and potentially highlight therapeutic targets. The thee proposed GS measures are summarised in ***Table 4.1***.

Table 4.1: Summary of proposed GS measures.

GS Measure	Description	Use	Source
Clinical GS	Correlation of a given genes expression values to the deviance residuals from a per-gene multivariate model.	Highlights regions of partition with strong associations to increasing or decreasing risk associated with upregulated expression.	Multivariate model identified in **Section 3.4.3** from survival analysis of TCGA SARC DDLPS clinical data.[64]
DEG GS	The logFC signed adjusted values from differential expression analysis of GSE159659.	Highlight regions of partitions that are positively or negatively associated with up and down regulated genes in DDLPS vs WDLPS.	Differential gene expression analysis of GSE159659 Microarray data.
Dependency GS	Gene effect score from the CHRONOS algorithm to infer gene essentiality in DDLPS cell lines.	To identify regions of partitions associated with cancer promoting and protective genes.	CRISPR-cas9 gene knockout screens conducted on DDLPS cell lines by the DepMap project.

Another step to assess the robustness of the GCN is through validation of modular co-expression patterns in a validation gene expression dataset using module preservation analysis.[252] Preservation is primarily represented through the composite Z-summary score which is a permuted and composite metric describing the significance of module preservation. This score combines seven correlation network statistics between the reference and test networks, the first four describe network density, then the remaining three describe network connectivity:

1. The mean correlation.

2. The mean adjacency.

3. The mean squared gene module membership values.

4. The mean correlation between genes and their module eigengene.

5. The correlation of intramodular connectivity's.

6. The correlation between gene module membership.

7. The correlation of modular correlation matrices between test and refence.

Density and connectivity measures are summarised by the median of respective metrics and are finally summarised by taking the mean of both density and connectivity measures to derive the summary statistic. To circumvent setting thresholds of significance for each individual metric, modulePreservation function randomly permutes module labels in the test network and recalculates preservation metrics for each permutation and is standardised into a Z statistic, presented as an asymptotic p-value where the null model is taken as no preservation following a normal distribution. In **Chapter 3.4.1** the NCC dataset was identified and proposed as a validation dataset to the TCGA SARC DDLPS data.

The overall goal of this results chapter is to delve into the intricacies of constructing a GCN using WGCNA. The process begins with a thorough inspection of the data at both the sample network and gene connectivity levels. This involves careful consideration of quality measures, such as outlier detection and ensuring scale-free topology. The chapter then explores various options for constructing the GCN, and after assessing relationships between IMC and GS measures as well as the validation of GCN modules, proposes the most robust GCN partition for downstream analysis and target screening.

4.2 Chapter aims and objectives.

Hypothesis: QC measures and GCN optimisation can collectively contribute to the definition of a robust DDLPS GCN.

Chapter Aims and Objectives:

Aim 1 – Identification and evaluation of sample outliers in RNA-seq datasets.

Objective 1.1: Identify outliers by assessing relationships between samples and characterising their properties.

Objective 1.2: Determine the optimal cutting point for outlier removal to ensure the robustness of subsequent analysis.

Objective 1.3: Evaluate the impact of outlier removal on the gene connectivity distribution within GCNs, thereby assessing the effectiveness of the outlier removal process.

Aim 2 – Determination of the most optimal WGCNA methodology for defining a GCN partition.

Objective 2.1: Construct a GCN using each available method to explore the impact of different methodologies on the resulting network structure.

Objective 2.2: Utilise GS measures to rank network partitions based on the strength of associations between gene connectivity and biological relevance, thereby identifying the most biologically meaningful network partition.

Objective 2.3: Conduct module preservation analysis using the NCC dataset to validate modular co-expression defined in the TCGA SARC DDLPS GCN.

4.3 Methods

4.3.1 RNA-seq FASTQ read processing, alignment, and quantification

Paired-end reads from RNA-seq experiments, in FASTQ format, were evaluated using the FASTQC package (version 0.11.9) on the Linux operating system.[334] The alignment and quantification of these FASTQ reads were performed using the STAR alignment package (version 2.7.10a) against the reference human genome, assembly version GRCh38.[336] The STAR package was selected for its ability to produce accurate alignments without extensive parameter tuning, its splice-aware alignment capabilities, and its efficiency when allocated large RAM pools.[337] The quality of alignment was assessed by examining the percentage of uniquely mapped reads. This was achieved by utilising the FASTQC html reports to scrutinise the alignment log files generated by the STAR aligner.[336] The resultant BAM files were configured to be sorted by coordinates, and the option 'quantMode' was set to 'GeneCounts', instructing STAR to count the number of reads per gene to quantity gene expression. The gene structure and positional annotations for the alignment were supplied through the GRCh38.108 gene transfer format (GTF), which was downloaded from Ensembl.[397]

4.3.2 WGCNA quality control

The WGCNA was implemented using the R package 'WGCNA' (version 1.72-1).[234] Before constructing the GCN, sample relationships were examined through the assessment of scaled connectivity and clustering coefficient from a weighted sample network. This was achieved using the 'adjacency' function in WGCNA, with a soft-thresholding power of 2. Network metrics for each sample were calculated using the 'fundamentalNetworkConcepts' function in WGCNA.[387,398]

To identify outliers, the above approach was combined with PCA and hierarchical clustering of principal components (PCs), using the FactoMineR (*version 2.1.1*) and factoextra (*version 1.0.7*) R packages. For outlier removal, the most appropriate threshold of median absolute deviation (MAD) around the median was identified by assessing alterations to gene connectivity distributions. The initial assessment involved evaluating the R^2 value of a linear fit to the log-log connectivity for thresholds of MAD around the median. To characterise the outliers, available clinical data was inspected, and variables were compared between groups using appropriate tests according to normal distribution and format (see statistical methods, **chapter 2.3**).

Gene connectivity was calculated using the 'softConnectivity' function in WGCNA, with the biweight midcorrelation ('bicor') set as the correlation function and a signed network specified. In WGCNA, correlation values are soft thresholded by raising the coefficients to a given power, which retains gene connectivity whilst suppressing weak connections towards zero, thus promoting scale-free topology.[196]Gene connectivity for candidate powers were tested based on the distribution of gene connectivity, max and min values of connectivity, and the number of unassigned genes in downstream clustering (using the "complete" linkage method, a sensitivity parameter of 3, and the TOM min).

The goodness-of-fit of gene connectivity distribution to a power law or exponential model was tested using the powerlaw Python package (version 1.5).[395] The 'fit.distribution_compare' function was used to compare the power law and exponential distribution. This function uses a likelihood ratio test to decide on the better fitting model. A significant p-value (p < 0.05/log10(p) > 1.3) from the likelihood ratio test indicates that one of the two models significantly describes the data. Testing was conducted for each soft-thresholding power (2 to 20 with intervals of 2) on the original non-cut data (DDLPS50), the chosen MAD threshold, the next most stringent threshold, and the threshold using the median. The NCC RNA-seq data was also assessed using the same methods as outlined above as detailed in the ***Appendix A.3***.

4.3.3 Gene-level multivariate Cox regression

To identify modules associated with clinical outcome, the multivariate overall survival model, as described in **Section 3.4.3,** was employed. It was found that the residual tumour and FNCLCC grade were significant variables for survival in this multivariate OS model. To further refine the model, gene expression values for a given gene corresponding to the sample were included as an additional variable for all genes in a stepwise manner, thereby creating a gene-wise multivariate model for each gene. Subsequently, the deviance residuals, which detail the expected risk of an event, were correlated with the respective gene expression values. This correlation resulted in the derivation of 'CorDeviance', designated as the "Clinical GS" metric, which was used to identify where strong patterns of upregulation infer increased or decreased risk of death.

4.3.4 Differential gene expression

For microarray expression datasets, the probe intensity values were examined for prior background correction, which were found by inspecting the dataset annotations which are included in the data package retrieved using the getGEO() function from the GEOquery (version 2.70.0) R package. The GSE159659 microarray data was acquired and inspected. The values were verified to be log quantile normalised by assessing whether the 99th quantile is larger than 100 (values greater than 100 indicate the data has not been log transformed). This method of normalisation is common practice in microarray data analysis.[327] The Limma package was chosen for microarray data pre-processing due to its inclusion of background correction and normalisation methods.[342] In microarray data, there can be numerous probes per gene interactions. This redundancy helps ensure accurate and reliable measurement of gene expression but for downstream analysis it is often necessary to select a single probe per gene to avoid redundant information and to simplify the data. In cases where there were multiple interactions for probe-gene, duplicates were consolidated based on average signal intensity.

The following Limma R package functions were used to perform the differential expression test. The initial fit was made using the linear model fitting function 'lmFit', then the contrasts for DDLPS vs WDLPS, WDLPS vs Adipose and Adipose vs DDLPS were fitted to the linear fit. The 'eBayes' function was used to apply an empirical Bayes statistic to the model with prior degrees of freedom set to 0.01. The 'TopTable' function was used to retrieve gene statistics for differential expression and the False Discovery Rate ("fdr" – Benjamini-Hochberg method) method was used to adjust p-values which was ranked by the B-statistic (log-odds). To derive the GS score, the FDR-adjusted p-value from differential expression with the contrasts of DDLPS vs WDLPS, WDLPS vs Adipose, and Adipose vs DDLPS was used. This p-value depicts the significance of genes across all contrasts. This was -log10 transformed for interpretability and signed by the logFC value from the DDLPS vs WDLPS contrast. This GS score was named the "DEG GS". The GSE159659 dataset was chosen over another microarray dataset, GSE30929, for deriving the GS score as it contained a higher overlap (84.2% in GSE159659 versus 60.6% in GSE30929) with the 16,032 genes in TCGA SARC DDLPS post-filtering and had matched normal adipose samples.

4.3.5 Gene dependency map

The cancer dependency MAP (DepMap) is an ongoing project by the Broad Institute that maintains a comprehensive database of cancer cell lines, including three DDLPS cell lines (LPS141, LPS853, LPS510). Each of these cell lines was queried for data from CRISPR gene knockout screens.[313,399] The CRISPR gene knockout effect score for each gene derived from the CHRONOS algorithm was obtained using the depmap R package (*version 1.10.0*)[248,310,315] and inverted, so that positive values indicated gene dependency. The mean gene effect value across all three cell lines was used as a GS score, which was named the "Dependency GS".

4.3.6 Partition quality assessment

To define the most robust GCN partition, various options for hierarchical clustering and dendrogram cutting within WGCNA were explored. The linkage methods considered included Ward.D, Ward.D2, single, complete, and average (***Figure 4.1***). These methods were selected due to their compatibility with the "hybrid" method for dynamic tree cutting in the 'CutTreeDynamic' function in the WGCNA R package.[400] A gene tree, or dendrogram, was constructed for each linkage method using the dissTOM as input to the 'hclust' function, specifying it as a distance object 'as.dist'. These gene trees where then used as input to the 'CutTreeDynamic' function, employing the "hybrid" method, and adjusting the deepSplit parameter from one to four. The deepsplit option controls the sensitivity of dendrogram cutting, where higher values generate more modules with fewer genes. There are also two main methods for calculating the TOM, where the topological overlap between two genes can be calculated as either the minimum sum between gene adjacencies (TOM min) or the mean of adjacencies (TOM mean). The various options provide 40 unique GCN partitions; here are labelled according to the options used (e.g., min_average_4, corresponds to a GCN partition derived using the TOM min, average-linkage hierarchical clustering, and a cutting sensitivity of four).

Each GCN partition may exhibit different IMC values, which represent the density of gene co-expression within a module. Higher values suggest a more densely co-expressed module. The patterns of pairwise relationships between kWithin and the GS measures will also differ among GCN partitions. A GCN partition that contains modules with a high correlation between the kWithin and the GS measures indicates potentially biologically significant modules. This not only reveals patterns of biological relevance within the network but will also assist in downstream module screening, as modules will be flagged as interesting based on high positive or negative values.

To assess the connectivity distribution in the network, the intramodular connectivity (kWithin) for each module was calculated using the 'intramodularConnectivity' function from the WGCNA R package. This function used the module labels from the tree cutting algorithm and adjacency matrix, with the scaleByMax option set as "TRUE". This option scales the modular connectivity values by the most connected gene in the module, allowing connectivity values to be normalised for module size and enabling comparison of partitions with variable-sized modules. The IMC values were summarised for each module by calculating the mean scaled connectivity values.

To identify the partition that best captured functional relevance in DDLPS, the kWithin values for each module were correlated with the three measures of GS (as described in **Section 4.2.3, 4.2.4** and **4.2.5**). The correlation between the kWithin and these three gene significance measures were then summarised using the maximum and minimum values, and the partition with the highest ranked metric was chosen.

4.3.7 Module Preservation

To validate modular co-expression patterns within the TCGA SARC DDLPS GCN, module preservation analysis was conducted using the independent NCC dataset (**See Section 3.4.1**).[252] For the preservation analysis conducted, the 'modulePreservation' function from the WGCNA R package was used with 500 permutations, a 'maxModuleSize' and a 'maxGoldModuleSize' of 248.[252] The networkType was set as "signed" and the correlation coefficient used was the 'bicor' for its efficient computation time.[250] The guidelines for interpreting preservation results as set by the authors of WGCNA module preservation analysis are used, which suggest any module with a 2 < Zsummary < 10 shows significant evidence of preservation, a Zsummary > 10 is considered highly significant, and any module with Zsummary <2 is considered to have no significant evidence of preservation.

4.4 Results

4.4.1 Optimisation of WGCNA

4.4.1.1 Outlier detection and removal

Principal component analysis of TCGA SARC DDLPS (n = 50) gene expression data was employed, in conjunction with hierarchical clustering, to identify sample outliers. This approach led to the identification of three primary clusters within the PCA space (***Figure 4.2A-B***). However, the first two PCs only accounted for ~26% of the data variance, suggesting additional sample heterogeneity. These distinct clusters were also discernible in the sample network metric plot (***Figure 4.2C***).

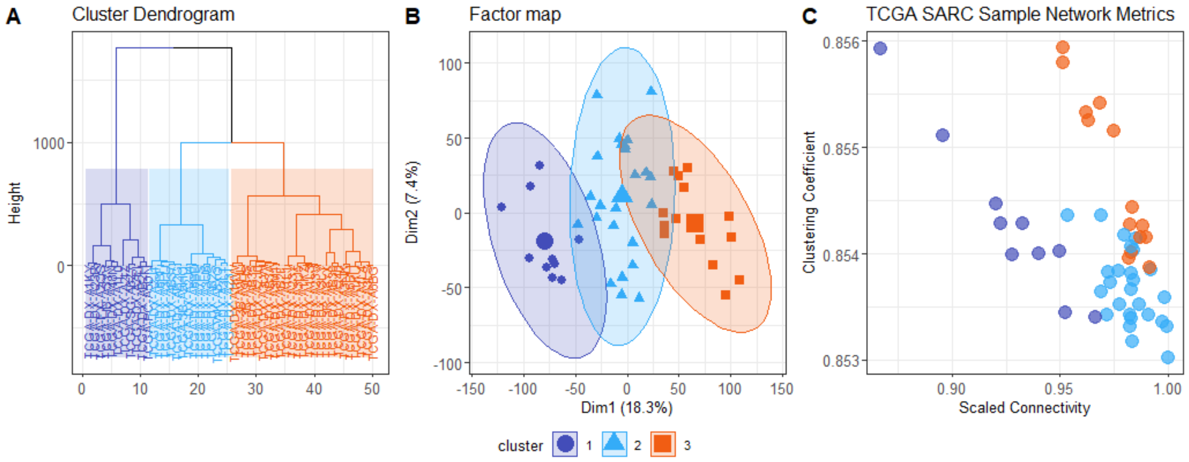


Figure 4.2: **Sample clustering for the TCGA data A**: Hierarchical clustering using the principal components from a PCA. **B**: Top two dimensions by variance explained with hierarchical clusters projected. **C:** Sample network metrics detailing relationship between samples, coloured by clusters from **A**.

To explore these further, the methylation, reverse phase protein array (RPPA) and copy number (CN) clustering analyses results available from the existing TCGA publication was cross referenced with the clusters from the HCPC (figure 4.2A ; ***Figure 4.3A-B***).[401] This revealed that CN and RPPA clusters showed no significant overlap (***Table 4.2***). However, the methylation clusters did (*chisq = 25.42, -log10pvalue = 4.14 ;* ***Table 4.2***). Further analysis of the standardised residuals from the chi squared test (chisq standardised residuals = 5, where anything >2 is significant) and percentages of the overlap (80%) revealed a significant overlap between HCPC cluster 1 and Methylation cluster M5 existed (***Figure 4.4***).

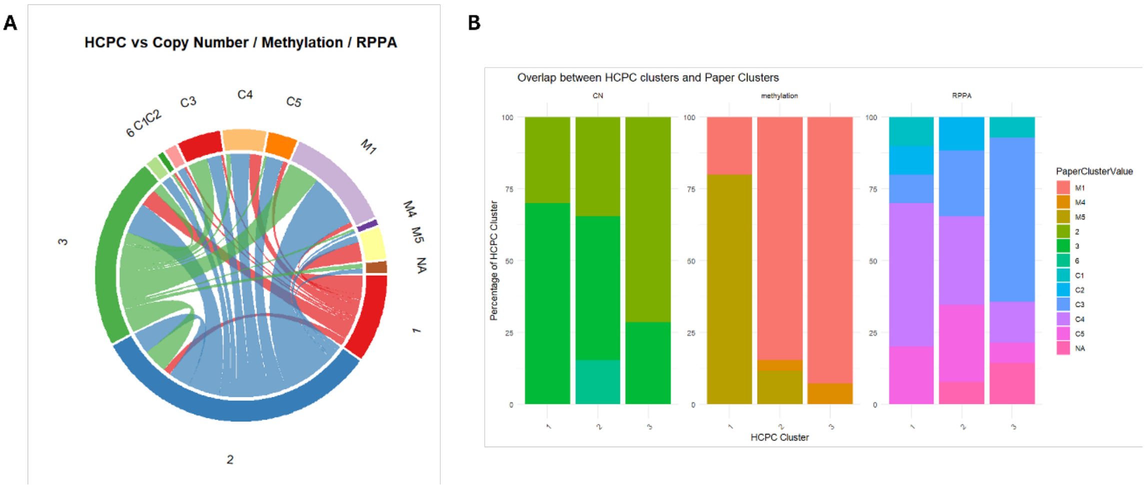


Figure 4.3: Overlap between methylation clusters and the clusters identified through hierarchical clustering of principal components (HCPC). **A** Chord diagram of cluster representation of HCPC clusters among the RPPA (reverse phase protein array), CN (copy number) and M (methylation) clusters. **B** Percentage representation of data as displayed in **A.**

Table 4.2: Chi squared statistics for HCPC clusters versus Methylation, CN and RRPA clusters from TCGA publication.[64]

Cluster type	Chi Squared statistic	DF	-log10(pvalue)	Significant
Copy Number	9.18 4	5	1.2465	No
Methylation	25.42	4	4.3830	Yes
RPPA	14.82	10	0.8571	No

DF – degrees of freedom. RPPA – Reverse Phase Protein Array.

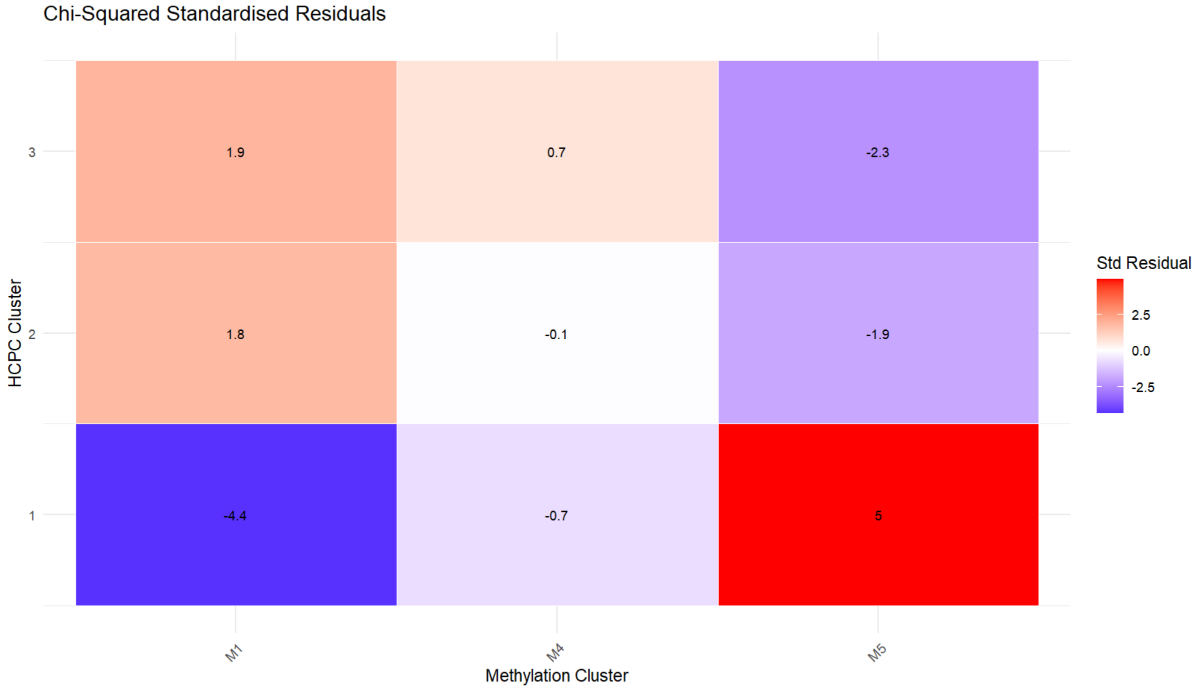


Figure 4.4: Chi squared standardised residuals for HCPC clusters versus methylation clusters. HCPC – Hierarchical clustering of principal components. M – Methylation clusters. Std residual– standardised residual from a chi-squared contingency table.

To further explore the relationship between these clusters and having identified a significant overlap between the cluster 1 and Methylation cluster 5 (M5) differential expression and gene set enrichment analysis was subsequently conducted. For this clusters 2 and 3 were merged to contrast with cluster 1. Genes upregulated in cluster 1 were found to be enriched for immune processes (***Figure 4.5***) including leukocyte mediated immunity (leukocyte mediated immunity, -log10padj = 126). This may suggest a relationship between DNA methylation and transcriptional programmes involving the immune system.

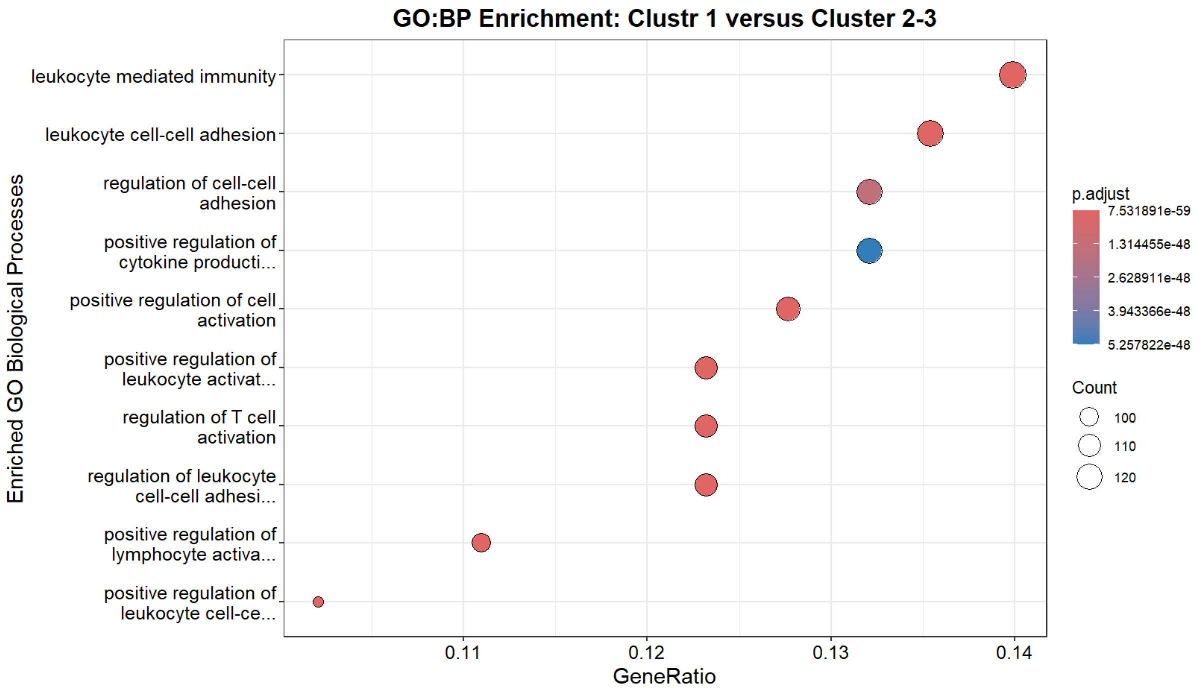


Figure 4.5: Gene Ontology (GO) Biological Process (BP) enrichment plot for genes differentially expressed between HCPC clusters 1 versus clusters 2-3. Dot colour indicate the Benjamini-Hochberg adjusted p value. Dot size indicates the number of genes intersecting with that term.

This analysis was repeated for the NCC gene expression data (n = 32). In this data PCA clustering (***Figure 4.6A-B***) identified three clusters. A lack of available adjunct data in the NCC metadata meant that cross referencing clusters with molecular annotations (e.g., protein expression or methylation, as was used in TCGA data above) was not possible.[114] Further analysing the NCC samples on a sample network (***Figure 4.6C***) reveals a tighter clustering of clusters as compared to the TCGA data where clustering coefficients and scaled connectivity show greater linearity (***Figure 4.2***). Linearity between these would suggest that sample clustering is the result of variations along a continuous scale (e.g., expected biological variation) rather than discreet groups due to biological or technical variance.[252,346]

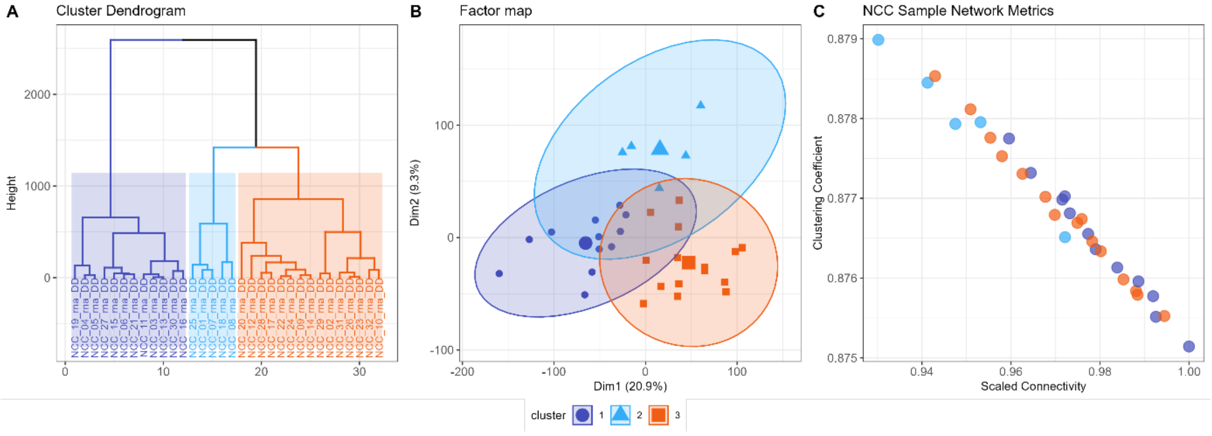


Figure 4.6: Sample clustering for the NCC data A: Hierarchical clustering using the principal components from a PCA. **B**: Top two dimensions by variance explained with hierarchical clusters projected. **C:** Sample network metrics detailing relationship between samples, coloured by clusters from **A**.

Differential expression and subsequent gene ontology enrichment analysis was conducted to assess differences in enriched functions between sample clusters (***Figure 4.7***). Cluster 1 vs Cluster 2 was enriched for immune processes. For example, leukocyte activation involved in immune response (-log10padj = 19.88) (***Figure 4.7A***). Cluster 2 versus cluster 1 DEGs are enriched for extracellular matrix organisation processes (e.g., extracellular matrix organization, -log10padj = 9.97) (***Figure 4.7B***). Cluster 2 versus cluster 3 is enriched for cellular development processes (e.g., renal system development, -log10padj = 6.81 (***Figure 4.7C***). Cluster 3 versus cluster 2 are enriched for immune cell processes (e.g., leukocyte migration, -log10padj = 17.58) (***Figure 4.7D***). Cluster 1 vs Cluster 3 are annotated for neuronal (e.g., neuromuscular processes, -log10padj = 2.68) and ion transport (e.g., regulation of monoatomic ion transmembrane transport, -log10padj = 3.20) (***Figure 4.7E***). DEGs in Cluster 3 versus Cluster 1 are enriched for extracellular matrix organisation (e.g., extracellular matrix organisation, -log10padj = 3.59) (***Figure 4.7F***).

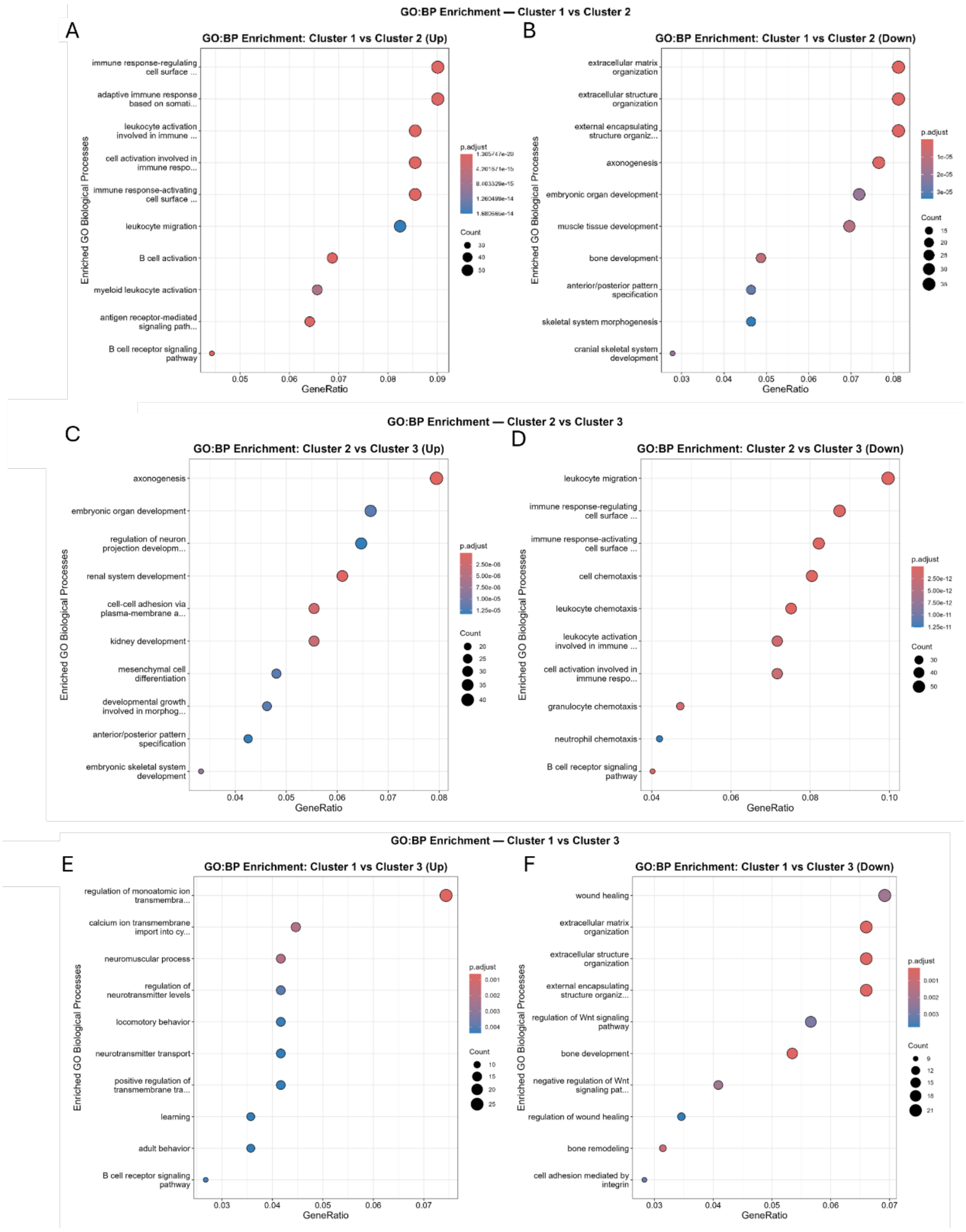


Figure 4.7: Gene Ontology (GO) Biological Process (BP) enrichment plot for genes differentially expressed (up and down) between HCPC clusters. Cluster 1 versus cluster 2 are denoted in **A** upregulated and **B** downregulated. Cluster 2 versus 3 is presented in **C** for upregulated genes **D** for downregulated. Cluster 1 versus cluster 3 for **E** upregulated genes and **F** for downregulated genes. Dot colour indicate the Benjamin-Hochberg adjusted p value. Dot size indicates the number of genes intersecting with that term.

The potential impact of outlier removal on scale-free topology was explored in the TCGA by assessing sample clustering coefficients and their scaled connectivity, using thresholds of MAD around the median. This was undertaken to identify a data subset exhibiting uniform gene expression between samples, thereby facilitating robust WGCNA.[346] Soft thresholding powers of 1 to 20 were tested, revealing that a threshold of 2xMAD around the median maximised the R^2 fitting index of a scale-free distribution for powers ≥ 12, as proposed by the authors of WGCNA for a signed GCN (***Figure 4.8***).[402] Using the 2xMAD threshold led to the identification of 14 outlier samples (***Figure 4.9A***), which were also evident when projecting a sample network (***Figure 4.9B***). These samples were subsequently removed.

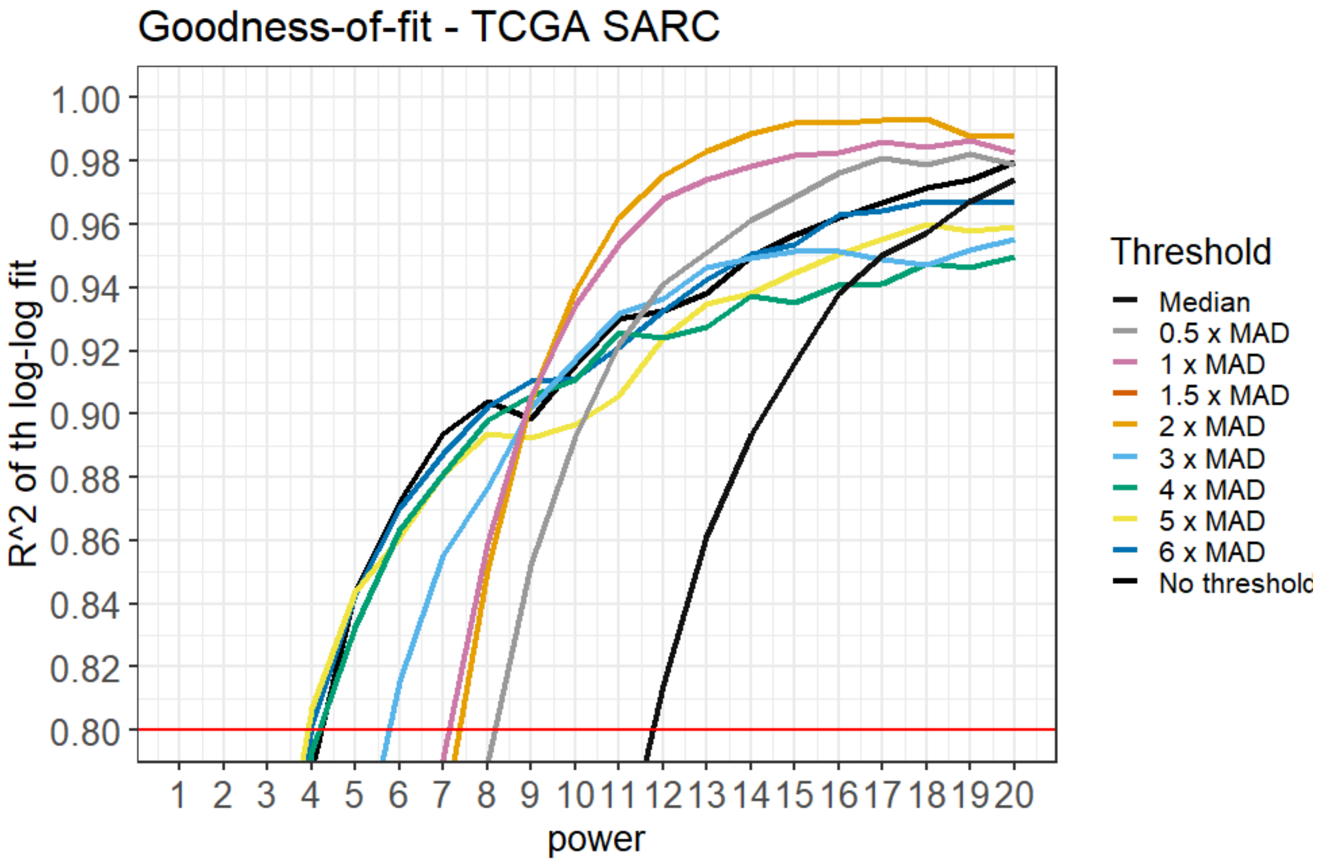


Figure 4.8: R^2 of the linear model fit for connectivity across values of β for the tested thresholds. The red dotted line indicates the value of R^2 that suggests an approximately scale-free network which is set out by the authors of WGCNA as being 0.8.

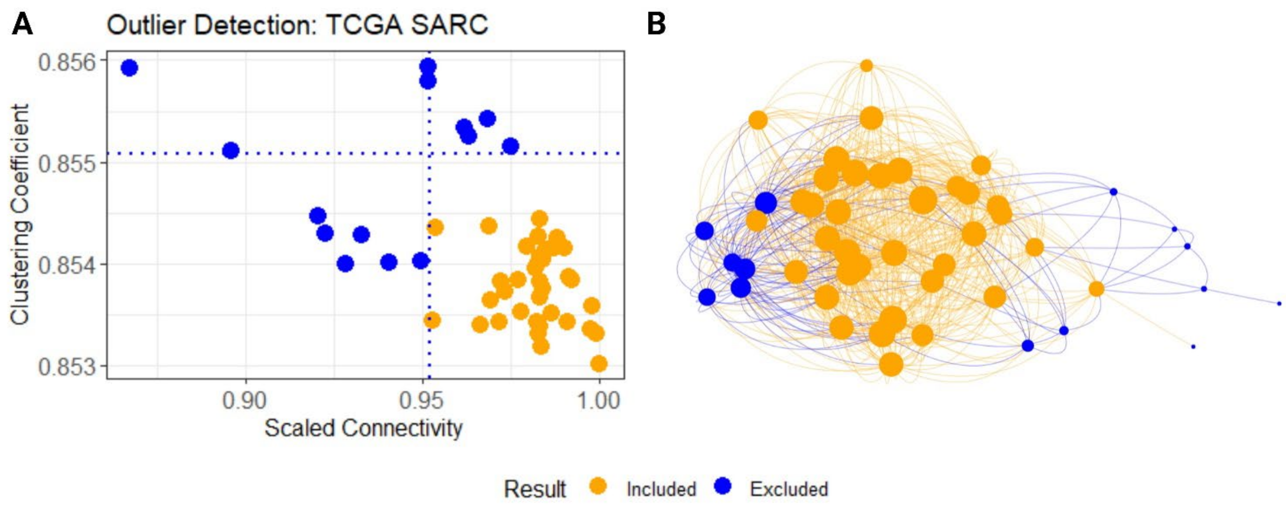


Figure 4.9: A Sample network metrics coloured by whether samples violate the 2*MAD threshold (blue line). **B** Corresponding sample network. Node size indicates node degree.

4.4.1.2 Characterising outliers

Clinical data was examined to characterise the outlying samples. As stated in **Section 3.5,** all samples underwent review by a pathology panel as per the requirements for the TCGA SARC study.[64] Our analyses (**Section 3.4.2**) indicated that these samples exhibit DDLPS disease characteristics that align with other studies in the literature, thereby confidently classifying these samples as DDLPS.

Statistical tests for each variable between included or excluded samples revealed significant differences in tumour mutational load (*p = 0.01;* ***Figure 4.10A***), and FNCLCC grade (*p = 0.01:* ***Figure 4.10B****)*. Disease-free status on follow-up (*p = 0.05:* ***Figure 4.10C***) is bordering on significance. Tumour ploidy (*p = 0.09*) and local recurrence (*p = 0.06*) exhibited trends towards significance (***Table 4.3***). Given that the grade was found to be an important predictor in the OS and RFS multivariate models, this may suggest that outlier samples have distinct survival probabilities.

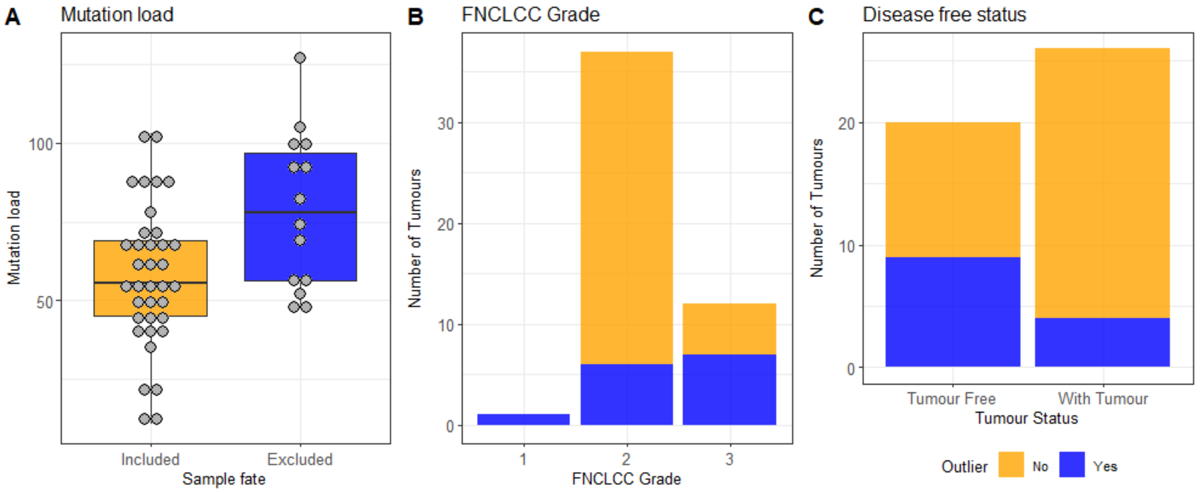


Figure 4.10: **A** Mutation load stratified by sample removal **B** FNCLCC Grade of patients by sample removal **C** Disease free status of patients on follow-up.

Table 4.3: Comparison of clinical characteristics between included and excluded TCGA samples.

Variable	Included (mean (sd) median)	Excluded (mean (sd) median)	p.value
Tumour weight (g)	504.44 (363.57) 445.0	430.71 (313.3) 390.0	0.62[a]
Tumour Size (cm)	18.45 (8.39) 19.5	21.16 (10.94) 18.0	0.41[b]
Avg12q copy number	22.50 (3.38) 25.0	21.07 (4.98) 22.5	0.43[a]
Mitotic Rate	6.72 (7.79) 5.0	8.29 (9.05) 2.5	0.89[a]
Ploidy	2.58 (2.06) 2.1	3.04 (1.07) 3.1	0.09[a]
Subclonal genome fraction	0.15 (0.17) 0.1	0.12 (0.09) 0.1	0.92[a]
CIN	434.6 (270.23) 377	458.36 (230.46) 424.5	0.65[a]
mutational load	57.63 (22.49) 55.5	78.64 (24.75) 78.0	**0.01[b]**
Age	62.85 (12.94) 61.9	65.32 (12.58) 63.9	0.54[b]
-	**Included (n = 36)**	**Excluded (n = 14)**	-
FNCLCC grade			**0.01[c]**
1	0	1	
2	31	6	
3	5	7	
Residual tumour			0.51[c]
R0	17	5	
R1	16	8	
R2	3	0	
Adjuvant Radiotherapy			0.13[c]
Yes	2	3	
No	34	11	
Previous malignancy			0.38[c]
Yes	4	3	
No	32	11	
Disease-free at follow up			0.05[c]
With tumour	22	4	
Tumour free	11	9	
New tumour event			0.53[c]
Yes	18	5	
No	18	9	
Local recurrence			0.06[c]
Yes	20	3	
No	16	11	
Distant recurrence			0.41[c]
Yes	8	1	
No	28	12	
Primary tumour site			0.37[c]
Retroperitoneal/Upper abdominal	32	12	
Lower extremity		1	
Upper extremity	1		
Superficial Trunk		1	
Chest	1		
Lower abdominal/Pelvic	2		
Gender			0.19[c]
Male	26	2	
Female	10	7	
Necrosis Score			0.36[c]
0	15	6	
1	21	7	
2	0	1	
Genome doublings			0.18[c]
0	20	5	
1 or 2	11	8	
Multifocal disease			0.74[c]
Yes	14	4	
No	21	9	

Superscript denotes statistical test [a]**: Wilcoxon rank-sum test.** [b]: **Unpaired independent T-test.** [c]**: Fisher's exact test. Bold *p.value*** indicates a significant (*p < 0.05*) result. **(sd): Standard deviation**

Survival probabilities were calculated for OS, RFS and DSS outcome measures, and subsequently compared between the included and outlier samples (***Table 4.4***). The outlier cohort exhibited a lower number of events, although survival probabilities were consistent for the first and third years but were notably higher for the fifth year. This is likely attributable to the small sample size in the outlier group.

Table 4.4: Survival probabilities for included and outlier samples for OS, RFS and DSS survival outcome measures.

Survival measures	**Included**		**Outlier**	
OS	**Probability**	**Number at risk | Number of events**	**Probability**	**Number at risk | Number of events**
1 year	0.83	30 | 6	0.86	10 | 2
3 year	0.58	14 | 7	0.64	4 | 2
5 year	0.37	7 | 4	0.64	3 | 0
RFS				
1 year	0.61	21 | 14	0.70	6 | 4
3 year	0.45	10 | 5	0.40	3 | 3
5 year	0.18	4 | 6	0.40	2 | 0
DSS				
1 year	0.81	24 | 6	0.92	12 | 1
3 year	0.58	10 | 5	0.72	6 | 2
5 year	0.49	5 | 1	0.57	3 | 1

OS – overall survival, **RFS –** recurrence free survival, **DSS** – disease specific survival.

At the univariate level, no significant difference in survival outcomes was observed between the groups (***Table 4.5; Figure 4.11)***. A new variable, termed "Outlier", was incorporated into the OS, RFS, and DSS multivariate models (see **section 3.4.2.1** and refer to ***Table 4.4***). These results suggested that for OS (HR = 0.18 [0.05-0.66], ***p=0.010***) and RFS (HR = 0.35 [0.13-0.95], ***p=0.039***), but not for DSS (HR = 1.26 [0.35-4.57], *p=0.723*), being classified as an outlier was a significant predictor of improved survival outcome. Each of the OS, RFS and DSS models were then re-evaluated for both the included and outlier groups. The OS and RFS models demonstrated that significant variables maintained their significance in the included group (***Table 4.6***), but not in the separate outlier group (***Table 4.7***).

Table 4.5: Multivariate models modified with whether samples were included or excluded.

OS model	**Variable categories**	**Variable levels**	**HR (univariable)**	**HR (multivariable)**
Residual Tumour	R0	22 (44.9)	-	-
	R1/R2	27 (55.1)	2.24 (0.87-5.75, p=0.093)	4.69 (1.60-13.76, p=0.005)
FNCLCC Grade	2	37 (75.5)	-	-
	3	12 (24.5)	3.24 (1.35-7.75, p=0.008)	15.38 (4.17-56.72, p<0.001)
QC Result	Included	36 (72.0)	-	-
	Outlier	14 (28.0)	0.69 (0.25-1.91, p=0.473)	0.18 (0.05-0.66, p=0.010)
RFS model				
Residual Tumour	R0	22 (44.9)	-	-
	R1/R2	27 (55.1)	1.89 (0.90-3.98, p=0.095)	2.84 (1.26-6.39, p=0.012)
FNCLCC Grade	2	37 (75.5)	-	-
	3	12 (24.5)	1.55 (0.71-3.35, p=0.269)	4.88 (1.71-13.89, p=0.003)
Gender	1	33 (66.0)	-	-
	2	17 (34.0)	0.84 (0.41-1.73, p=0.636)	0.58 (0.26-1.33, p=0.202)
QC Result	Included	36 (72.0)	-	
	Outlier	14 (28.0)	0.68 (0.29-1.56, p=0.358)	0.35 (0.13-0.95, p=0.039)
DSS model				
Multifocal Disease	NO	30 (62.5)	-	-
	YES	18 (37.5)	2.49 (0.92-6.73, p=0.071)	3.20 (1.10-9.25, p=0.032)
Age	Mean (SD)	63.5 (12.8)	0.98 (0.95-1.03, p=0.458)	0.96 (0.92-1.01, p=0.120)
QC Result	Included	36 (72.0)	-	-
	Outlier	14 (28.0)	0.63 (0.20-1.94, p=0.420)	1.26 (0.35-4.57, p=0.723)

OS – overall survival. **RFS** – Recurrence free survival. **DSS** – Disease specific survival

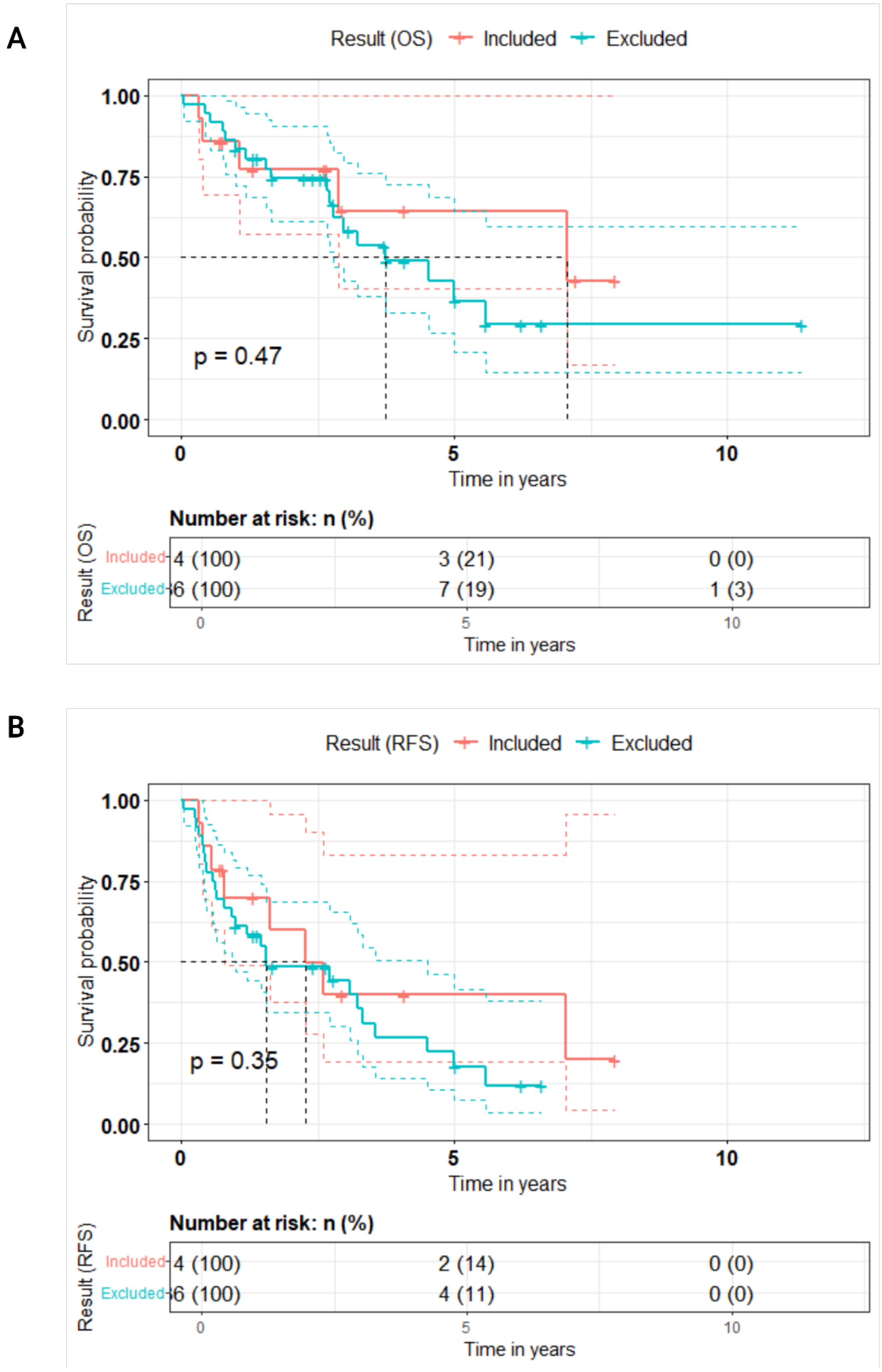


Figure 4.11: Kaplan-Meier plots for includes versus excluded samples for the **A** OS outcome measure and **B** the RFS outcome measure. Fits are coloured according to variable factors. Dotted coloured lines represent the lower and upper 95% confidence intervals for respective plots. Grey dotted lines indicate the median survival times. Points plotted on survival curves represent patient censoring.

Table 4.6: Multivariate models using the filtered DDLPS36 (MAD*2 cut) data.

OS model (included)	**Categories**	**Number**	**HR (univariable)**	**HR (multivariable)**
Residual Tumour	R0	17 (47.2)	-	-
	R1/R2	19 (52.8)	2.60 (0.93-7.32, p=0.070)	4.24 (1.29-13.93, p=0.017)
FNCLCC Grade	2	31 (86.1)	-	-
	3	5 (13.9)	13.74 (3.82-49.45, p<0.001)	25.53 (5.62-116.05, p<0.001)
RFS model (included)				
Residual Tumour	R0	17 (47.2)	-	-
	R1/R2	19 (52.8)	1.62 (0.72-3.67, p=0.244)	2.13 (0.90-5.05, p=0.085)
FNCLCC Grade	2	31 (86.1)	-	-
	3	5 (13.9)	4.60 (1.58-13.38, p=0.005)	6.04 (1.97-18.57, p=0.002)
Gender	1	26 (72.2)	-	-
	2	10 (27.8)	0.89 (0.37-2.16, p=0.804)	0.68 (0.27-1.69, p=0.402)
DSS model (included)				
Multifocal Disease	No	21 (60.0)	-	-
	Yes	14 (40.0)	2.55 (0.81-8.08, p=0.111)	2.86 (0.89-9.24, p=0.079)
Age	Mean (SD)	62.9 (12.9)	0.97 (0.92-1.02, p=0.196)	0.96 (0.91-1.01, p=0.141)

OS – overall survival. **RFS –** recurrence free survival. **DSS** – disease specific survival.

Table 4.7: Multivariate models using the outlier samples from the 2*MAD cut.

OS model	**Categories**	**Number**	**HR (univariable)**	**HR (multivariable)**
Residual Tumour	R0	5 (38.5)	-	-
	R1/R2	8 (61.5)	1.53 (0.16-14.80, p=0.712)	6.89 (0.28-169.05, p=0.237)
FNCLCC Grade	2	6 (46.2)	-	-
	3	7 (53.8)	2.11 (0.34-12.95, p=0.420)	6.61 (0.35-123.75, p=0.207)
RFS model				
Residual Tumour	R0	5 (38.5)	-	-
	R1/R2	8 (61.5)	3.73 (0.44-31.33, p=0.225)	6.17 (0.52-73.03, p=0.149)
FNCLCC Grade	2	6 (46.2)	-	-
	3	7 (53.8)	1.02 (0.25-4.10, p=0.981)	3.40 (0.24-48.39, p=0.366)
Gender	1	7 (50.0)	-	-
DSS model				
Multifocal Disease	NO	9 (69.2)	-	-
	YES	4 (30.8)	2.23 (0.29-16.85, p=0.439)	4.26 (0.08-227.88, p=0.475)
Age	Mean (SD)	65.3 (12.6)	1.04 (0.95-1.14, p=0.421)	0.96 (0.80-1.16, p=0.693)

OS – overall survival. **RFS –** recurrence free survival. **DSS** – disease specific survival.

Finally, given the results from **section 4.4.1.1,** the relationship between the cut point used and the methylation clusters form the TCGA SARC publication was assessed.[64] This identified that these clusters are independent (***chisq = 5.34, p-value = 0.068***)). However, in assessing the residuals it was identified that the M5 methylation cluster was significantly (chi square standardised residuals = 2.22; ***Figure 4.12***) represented in the included sample set (n = 36) but not the excluded (n = 14).

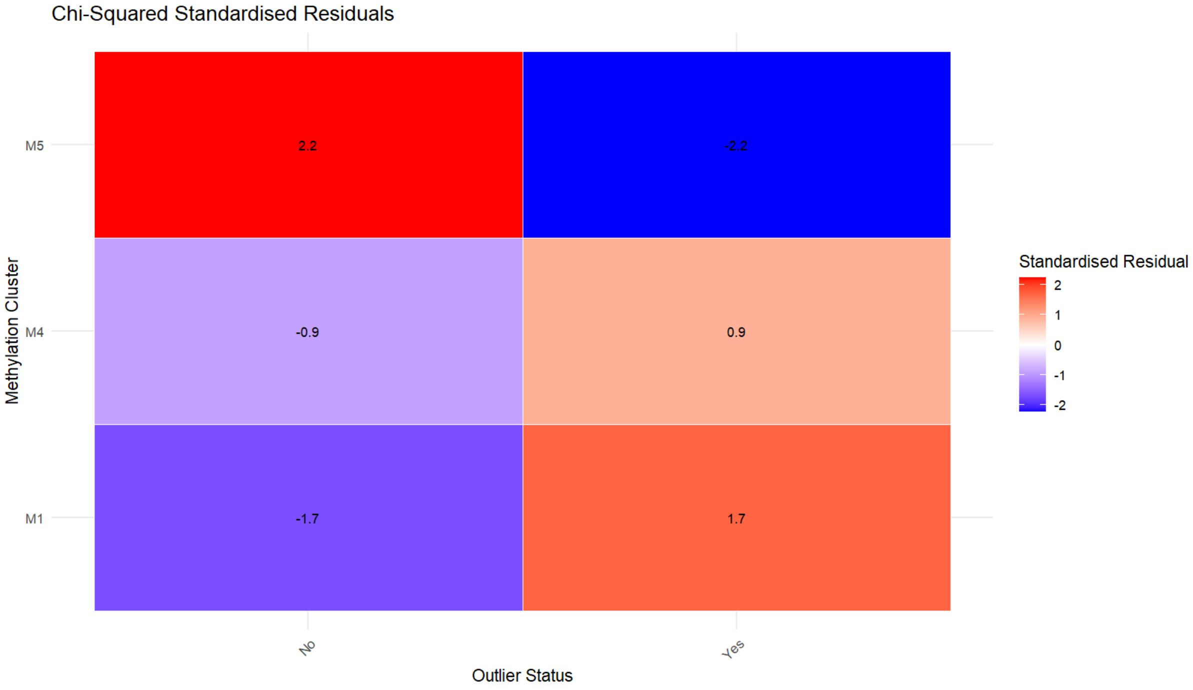


Figure 4.12: The standardised chi squared residuals for outlier status versus Methylation Cluster. Outlier status is the status of whether samples passed the sample network metrics quality control. Methylation cluster denotes the cluster that DDLPS patients were assigned in the TCGA SARC publication based on DNA methylation.[64]

4.4.1.3 Scale-free topology analysis

To determine whether the gene expression data exhibited scale-free topology, a likelihood ratio test was utilised to assess whether the data was best described by a power law or an alternative exponential model. The test was applied to four datasets: the 2xMAD threshold data (n = 36, DDLPS36), the original data (n = 50, DDLPS50), a 1.5xMAD threshold (n =34, DDLPS34), and using the median value (n = 16, DDLPS16).

The results indicated that the DDLPS36 data adhered more closely to a power law distribution for powers beta ≥ 10 (as shown in ***Figure 4.13A***). This contrasted with DDLPS50, where such a pattern was not observed for the powers tested (***Figure 4.13B***). Furthermore, the test statistics and significance values for a power law distribution were found to be higher in the DDLPS36 dataset compared to the original data (as depicted in ***Figure 4.13A-B***). This trend is also observed in the DDLPS34 data which provides similar results to the DDLPS36 data (***Figure 4.13C***). When using the median as the cutting threshold, an exponential distribution begins to become more apparent (***Figure 4.13D***) and is more comparable to the original data (***Figure 4.13A***). This suggests that both the DDLPS36 and DDLPS34 data subsets provides a more unform dataset for WGCNA analysis as compared to the original data and using the median.

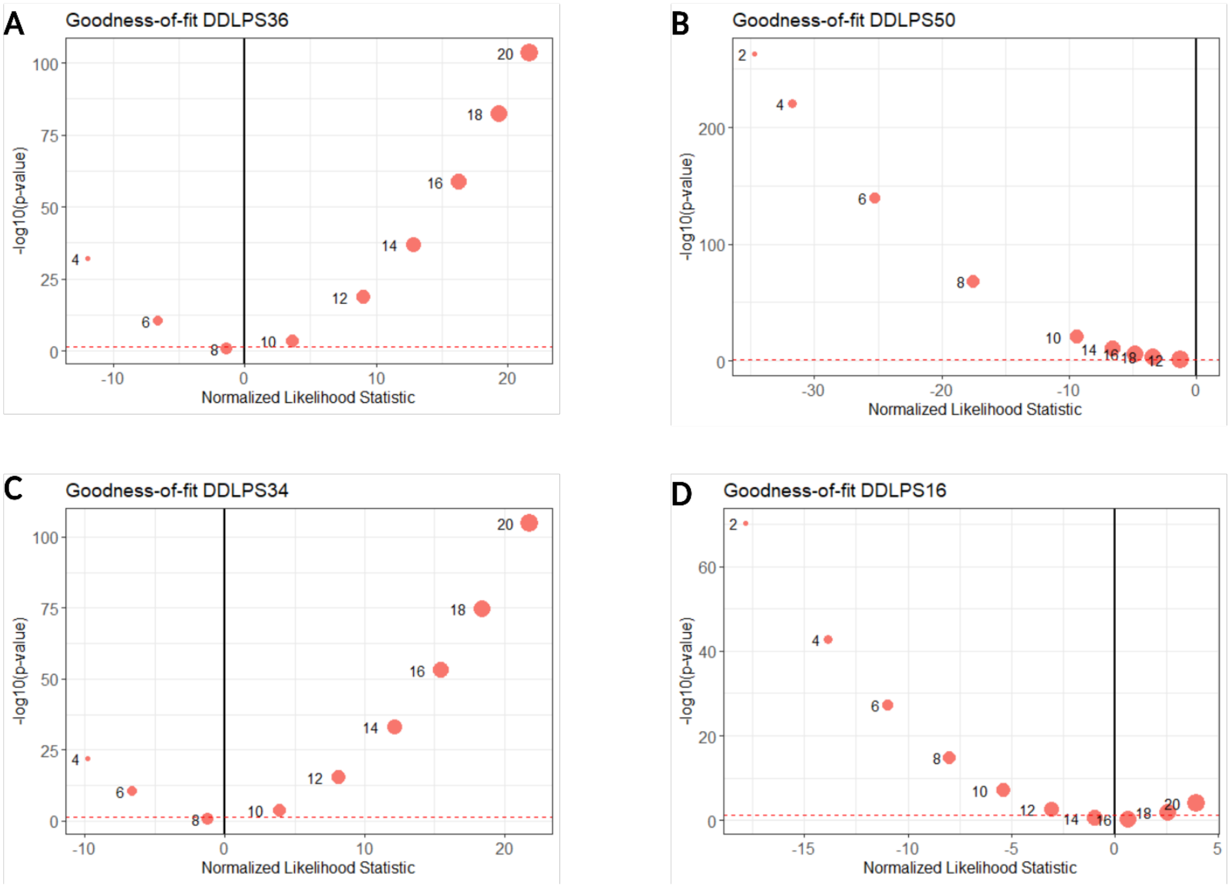


Figure 4.13: Goodness-of-fit of connectivity distribution for power law versus an alternative exponential distribution for: **A.** A 2*MAD threshold (DDLPS36) on sample network metrics. The result for a power of 2 could not be reported as it retrieved "NA" in statistical testing. **B.** Connectivity distribution with no sample network metric filtering (DDLPS50). **C.** A 1*MAD threshold (DDLPS34) and **D.** Using the Median scaled connectivity and clustering coefficient as the threshold (DDLPS16). Red dotted line indicates a significant result from a likelihood ratio test (-log10(p-value) ≥ 1.3) where any point above the line is significantly described by one of the distributions tested. The normalised likelihood ratio test statistic details which model is best fitted where any point right of the solid black line is best described by a power law distribution and any point left of the black line is best described with the alternative exponential distribution.

In this approach where sample number is an important consideration, it is desirable to maintain as many samples as possible whilst maximising scale-free topology. For this reason, DDLPS34 was not considered over the DDLPS36 subset. Then the DDLPS36 was compared against the best of 50 random permutations of outlier exclusions (using the 36:14 split of DDLPS36) (***Figure 4.14A-B***). This showed that for most powers, the gene connectivity of random partitions was highly significantly described by an exponential distribution versus a power law (***Figure 4.14B***). However, for powers of 14, 16 and 20, it is observed that the random shuffling is described better by a power law distribution (***Figure 4.14B***).

Samples flagged for exclusion based on outlying expression levels using a 2*MAD threshold on sample network metrics are associated with higher mutational load, tumour grade and have better survival outcomes (OS and RFS) when adjusting for tumour grade and residual margin, but not in univariate testing. Due to sample size, it is difficult to draw robust conclusions. The outlying samples did not represent a specific disease subtype, based on annotations available, and removal was considered to benefit the identification of drug targets using the GCN.

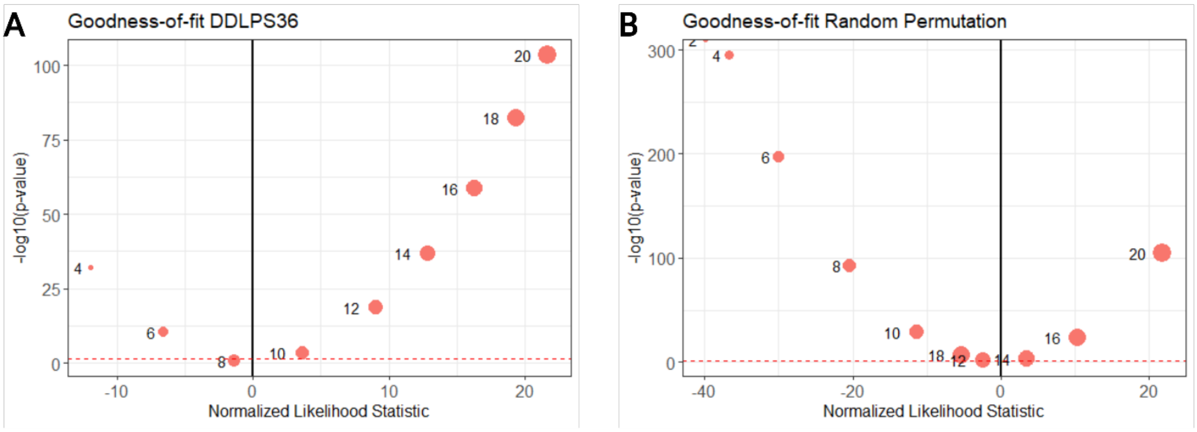


Figure 4.14: Goodness-of-fit of connectivity distribution for power law versus an alternative exponential distribution for: **A** The DDLPS36 data and **B** The maximum values from 50 permutations or random removals of 14 samples from the original cohort. Red dotted line indicates a significant result from a likelihood ratio test (-log10(p-value) ≥ 1.3) where any point above the line is significantly described by one of the distributions tested. The normalised likelihood ratio test statistic details which model is best fitted where any point right of the solid black line is best described by a power law distribution and any point left of the black line is best described with the alternative exponential distribution.

4.4.1.4 Selection of the soft-thresholding power

The selection of the most suitable soft-thresholding power is an important step in WGCNA. In this analysis, the commonly used power of 12 was compared against the power of 14, where the R^2 value inflects (***Figure 4.8*** – 2xMAD) as well as the power of 10 where a power law fit becomes significant (***Figure 4.14A***). To illustrate the effects of soft-thresholding and the necessity for higher powers in signed networks, a power of 6 was also evaluated. The gene connectivity distribution for these powers was subsequently tested (***Figure 4.15A-H; Table 4.8***).

For the power of 6, the connectivity distribution appeared more homogenous, resembling an exponential-like distribution (***Figure 4.15A***). In agreement with the goodness-of-fit testing (***Figure 4.15B***), the low R^2 value of 0.64 indicated that at this power, the data did not exhibit scale-free topology (***Figure 4.15B***). This is supported by the power of 6 giving a very high mean connectivity of 463 (***Table 4.15A***).

The power of 10 (***Figure 4.15C***) shows a better R^2 fit (***Figure 4.15D***) of 0.93 as compared to the power of 6, although contains genes that have a markedly higher connectivity with a mean value of 81 (***Figure 4.15C; Table 4.8***) as compared to power of 12 with a mean value of 39 (***Figure 4.15C; table 4.8).*** In comparing powers of 12 and 14, it was observed that the power of 14 skews gene connectivity towards a closer fit for a scale-free topology as compared to the power of 12 (***Figure 4.15C-E***). However, powers > 12 resulted in a substantial reduction in both mean (20) and maximum (179) connectivity in the network (***Figure 4.15E-F; Figure 4.15H-G; Figure 4.16A-C; Table 4.8***). This also led to a sizeable increase in the number of unassigned genes from downstream clustering (e.g., power of 14, 7027/16032 – ***Table 4.8***).

The power of 10 (0.93; ***Figure 4.15D***) had a reduced R^2 fitting index as compared to the power of 12 (0.97; ***Figure 4.15F***) and showed lower significance in the formal goodness-of-fit testing (***Figure 4.14B***). In WGCNA, "grey" modules, are typically used to label a module for genes that are "unassigned" or independent in their patterns of co-expression. The power of 10 provides a smaller "grey" module, containing just 59 genes (***Table 4.8***) as compared to the 2,012 for power of 12. However, the power of 12 shows a gene connectivity distribution that adheres better to the assumptions of a scale-free topology. The selection of the soft-thresholding power is a balance between achieving a scale-free topology and maintaining meaningful connectivity in the network. After careful consideration, a soft-thresholding power of 12 was chosen, as this value provided the best balance between these two objectives.

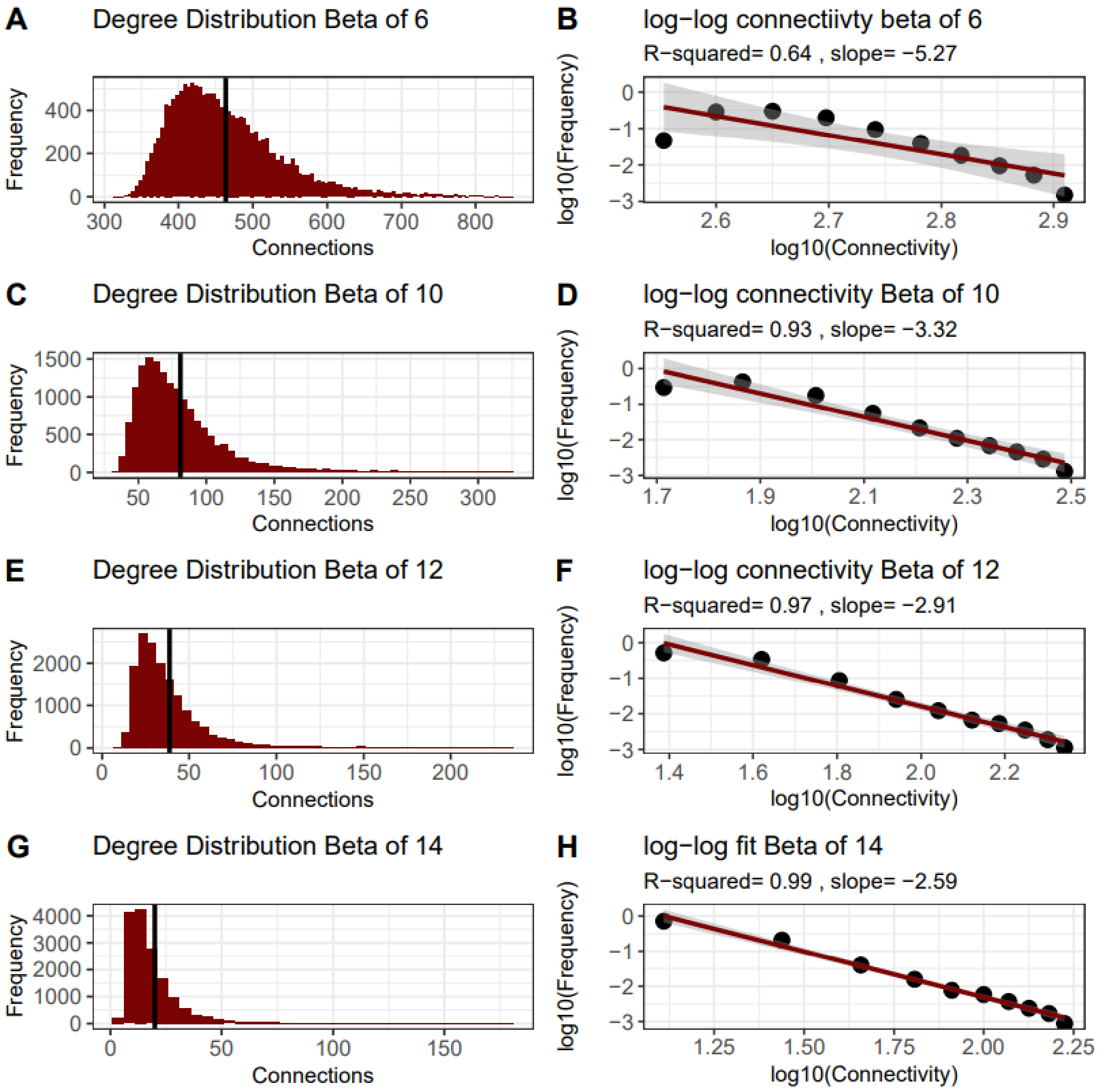


Figure 4.15: Assessment of degree distribution and fit to power law distribution for three powers of beta. **A** degree distribution and **B** fitting index for beta of 6. **C** degree distribution and **D** fitting index for beta of 10. **E** degree distribution and **F** fitting index for beta of 12, the recommended power for WGCNA. **G** degree distribution and **H** fitting index for beta of 15.

Table 4.8: Summary of gene connectivity values and unassigned genes from downstream clustering analysis

power	Max size	Mean size	Size of grey module	Max connectivity	Mean connectivity
6	202	59.38	NA	846.47	463.57
10	178	61.43	59	323.87	81.05
12	263	57.46	2012	233.95	38.80
14	215	50.88	7027	178.81	20.06

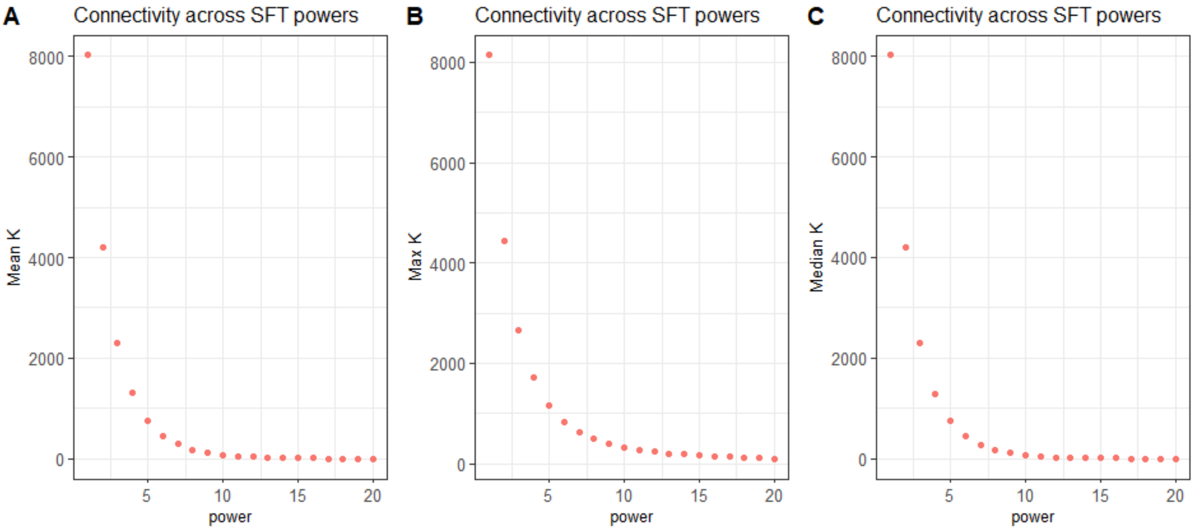


Figure 4.16: Summary of connectivity (k) across powers. **A** The mean k, **B** the max k and **C** the median k.

4.4.2 Evaluating GCN partitions

4.4.2.1 GCN partition ranking

Overall, there were 40 distinct GCN partitions assessed, each with a different number of modules and number of genes within those modules (***Table 4.9***). Maximum module sizes range from 931 to 12,553, minimum modules correspond to the baseline minimum value set for dynamic tree cutting of 30 ranging from 30 to 1,342. Mean module sizes range from 60 to 4,008 genes. In WGCNA, "grey" modules. The size of the grey modules ranges from 0 to 10,120 (***Table 4.9***). When considering max module sizes, the grey module was not counted, as it was observed that for some partitions the grey module was very large (e.g., containing >50% of expressed genes) (***Table 4.9***). Wards D method provided the largest modules overall with no "grey" module. Ward D2 and Single provided similar module sizes, followed by average, and then complete linkage producing the module with the smaller number of genes. Complete (paired with low sensitivity parameter values) and single linkage methods produced the largest "grey" modules. Wards methods produced no grey modules, with average producing the second smallest grey modules. Increased sensitivity parameter for the dynamic cut increased the number of modules and decreased the module size. A mean TOM threshold produced smaller modules compared to a minimum TOM threshold although the overall pattern was similar.

Table 4.9: Descriptive metrics for module number and sizes across GCN partitions.

Partition	Number of modules	Max module size	Min module size	Mean module size	Size of "grey" module
Min complete 4	266	1201	30	60.27068	1201
Min complete 3	245	2012	30	65.43673	2012
Mean complete 4	239	1960	30	67.0795	1960
Mean complete 3	213	2869	30	75.26761	2869
Min complete 2	138	5844	30	116.1739	5844
Mean complete 2	130	5453	30	123.3231	5453
Mean average 4	114	1445	33	140.6316	582
Min average 4	90	2098	31	178.1333	463
Mean complete 1	74	8556	32	216.6486	8556
Mean average 3	74	1434	37	216.6486	758
Mean ward.D2 4	72	931	32	222.6667	0
Min average 3	71	2130	36	225.8028	561
Min ward.D2 4	69	1162	42	232.3478	0
Min complete 1	68	10120	30	235.7647	10120
Mean ward.D2 3	39	2201	73	411.0769	0
Min ward.D2 3	33	1816	83	485.8182	0
Mean average 2	30	1991	58	534.4	1449
Min average 2	28	2513	39	572.5714	645
Min ward.D 4	19	3722	86	843.7895	0
Mean ward.D 4	19	4209	81	843.7895	0
Min ward.D2 2	18	3411	83	890.6667	0
Mean ward.D2 2	18	4662	88	890.6667	0
Min ward.D2 1	14	4043	83	1145.143	0
Mean ward.D 3	13	4958	90	1233.231	0
Mean ward.D2 1	13	6299	88	1233.231	0
Mean average 1	13	2187	203	1233.231	1769
Min ward.D 3	12	4908	86	1336	0
Min average 1	11	3296	167	1457.455	1953
Mean single 4	7	4891	827	2290.286	4891
Min ward.D 2	6	6876	156	2672	0
Min single 2	6	6313	955	2672	6313
Min single 3	6	6314	955	2672	6314
Min single 4	6	6312	954	2672	6312
Mean ward.D 2	6	6790	258	2672	0
Mean single 1	6	5498	1342	2672	5498
Mean single 2	6	5497	1342	2672	5497
Mean single 3	6	5499	1342	2672	5499
Min single 1	5	5469	959	3206.4	5469
Min ward.D 1	4	12195	260	4008	0
Mean ward.D 1	4	12553	258	4008	0

Partition nomenclature details the **TOM calculation** used as either **Mean** or **Min** , then the hierarchical clustering **linkage method** used being one of **Single, Complete, Average, Ward.D** and **Ward.D2,** then followed by the sensitivity parameter of the dynamic cutting algorithm from **1** to **4.** Module size calculations were made after the exclusion of the "grey" module which depicts unassigned (or independent) genes.

Partitions showed a range of IMC values across module size (***Figure 4.17***). Those derived using the mean TOMs yielded the highest overall IMC value, likely attributed to the average module size (***Figure 4.11***). GCN partitions derived from complete linkage using TOM min or TOM mean with a cutting sensitivity of 2 or above also showed high IMC compared to other methods (***Figure 4.17***). Moreover, they produced more modules, with complete linkage and a cutting sensitivity of 3 or above, for either TOM min or TOM mean, yielding over 200 modules (***Figure 4.17***).

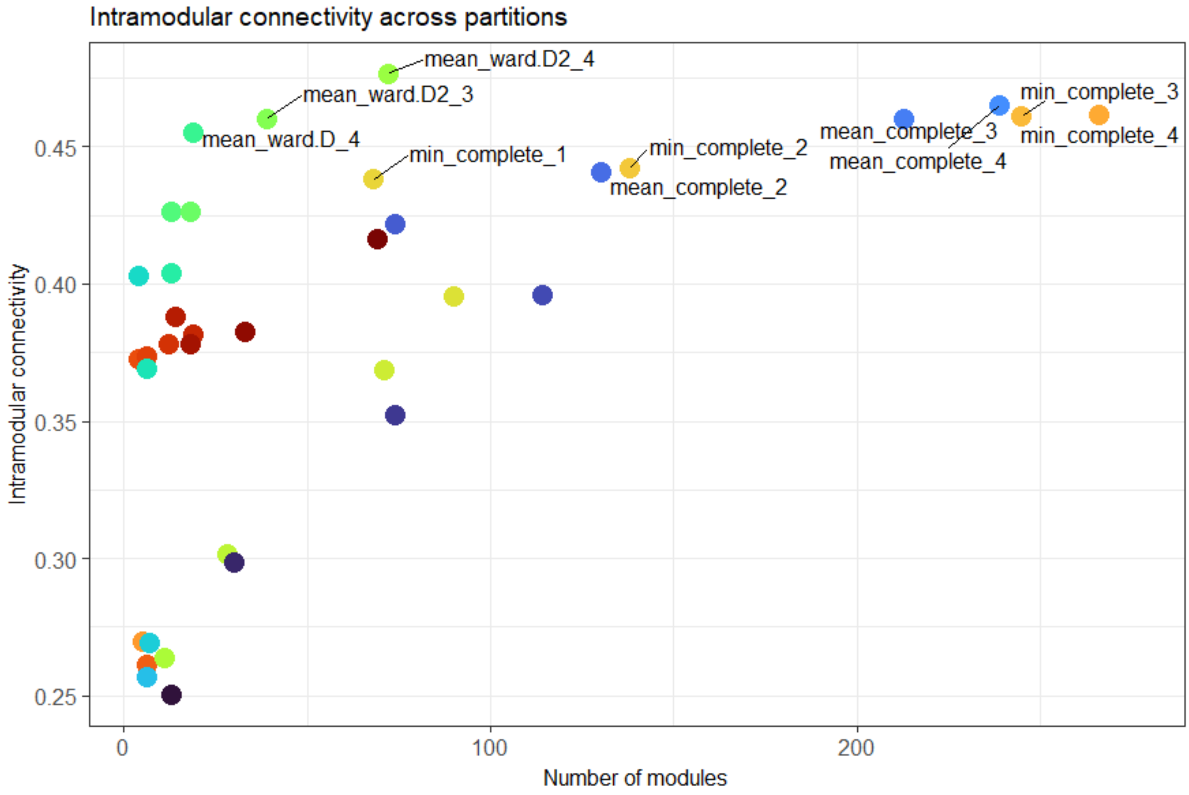


Figure 4.17: Intramodular connectivity across partitions of GCN derived from different GCN construction, clustering and cutting options available in the WGCNA. Higher intramodular connectivity values indicate modules with tighter co-expression.

Assessing the maximum correlation values of kWithin against the GS measure indicated that GCN partitions from the complete-linkage methods provided the highest correlation values for the Clinical GS (***Figure 4.18A***), Dependency GS (***Figure 4.18B***), and DEG GS (***Figure 4.18C***). These high values typically correspond to the GCNs calculated using the TOM min (***Figure 4.18A-B***), except for mean_complete_2 in the DEG GS (***Figure 4.18C***). The minimum values represent the negative correlations, which can be inferred as being not associated with DDLPS. These also indicated that the complete linkage method performed best (***Figure 4.19A-C***), except for mean_average_4 in the DEG GS (***Figure 4.19C***). TOM mean provided lower minimum values for both the clinical GS (***Figure 4.13A***) and DEG GS (***Figure 4.13C***). Using the minimum and maximum values to rank the GCN partitions revealed that the partition derived from the TOM min, complete linkage with a cutting sensitivity of 3 (min_complete_3) provided the overall best partition (***Table 4.10***). This partition contained 245 modules of co-expressed genes.

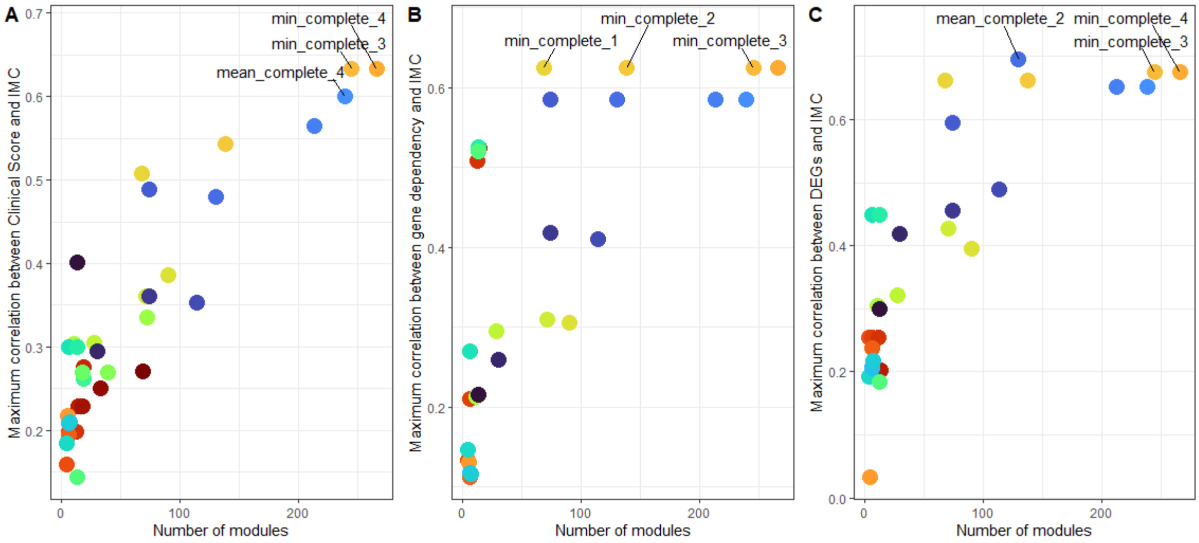


Figure 4.18: Maximum correlation values of GS vs K-within for each partition **A** Clinical score GS **B** Dependency GS **C** DEG score GS.

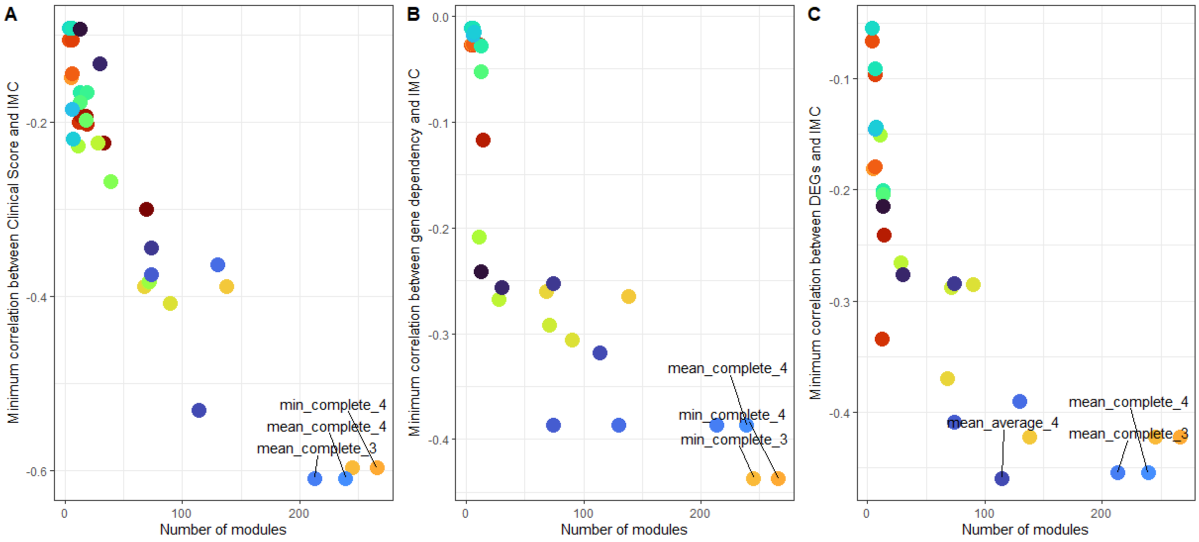


Figure 4.19: Minimum correlation values of GS vs Kwithin for each partition **A** Clinical score GS **B** Dependency GS **C** DEG score GS.

Table 4.10: Top 10 highest ranked partitions. Each of the summaries for GS vs IMC results were ranked. The sum of these ranks were taken (sum) where a lower sum rank indicates a higher rank overall.

partition	Modules	GSvsIMC (min)	GSvsMI (max)	DMvsMI_min	DMvsMI_max	DEGvsMI_min	DEGvsMI_max	sum	Rank
min_complete_3	245	-0.597	0.632	-0.437	0.624	-0.423	0.675	31	1
min_complete_4	266	-0.597	0.632	-0.437	0.624	-0.423	0.675	33	2
mean_complete_4	239	-0.609	0.6	-0.387	0.584	-0.455	0.651	40	3
mean_complete_3	213	-0.609	0.564	-0.387	0.584	-0.455	0.651	43	4
min_complete_2	138	-0.389	0.543	-0.265	0.624	-0.423	0.661	53	5
mean_complete_2	130	-0.365	0.479	-0.387	0.584	-0.391	0.695	55	6
min_complete_1	68	-0.389	0.507	-0.26	0.624	-0.37	0.661	57	7
mean_complete_1	74	-0.376	0.488	-0.387	0.584	-0.41	0.595	60	8
mean_average_4	114	-0.531	0.353	-0.318	0.409	-0.46	0.489	65	9
min_average_4	90	-0.408	0.386	-0.307	0.305	-0.285	0.395	83	10

To demonstrate the overlap between partitions, modules in which MDM2 and CDK4, two key genes in WDLPS/DDLPS tumorigenesis, were sorted into were compared for size (***Figure 4.20A-B***) and co-occurring genes (***Figure 4.20C-D***). Across partitions it is observed that the number of genes in each of the "MDM2" and "CDK4" modules have consistent sizes for more than half of the partitions. "MDM2" module shows a higher concordance of co-expressed genes between partitions as compared to the "CDK4" module (***Figure 4.20C-D***). This suggests that for many partitions the overlap between partitions is higher (single, Ward D, Ward D2) compared to some partitions (complete, average) where there is a drop-off of concordant module genes. However, it is also observed that there are "core" co-expressed genes with both MDM2 and CDK4 that have high occurrence across partitions (***Figure 4.20C-D***).

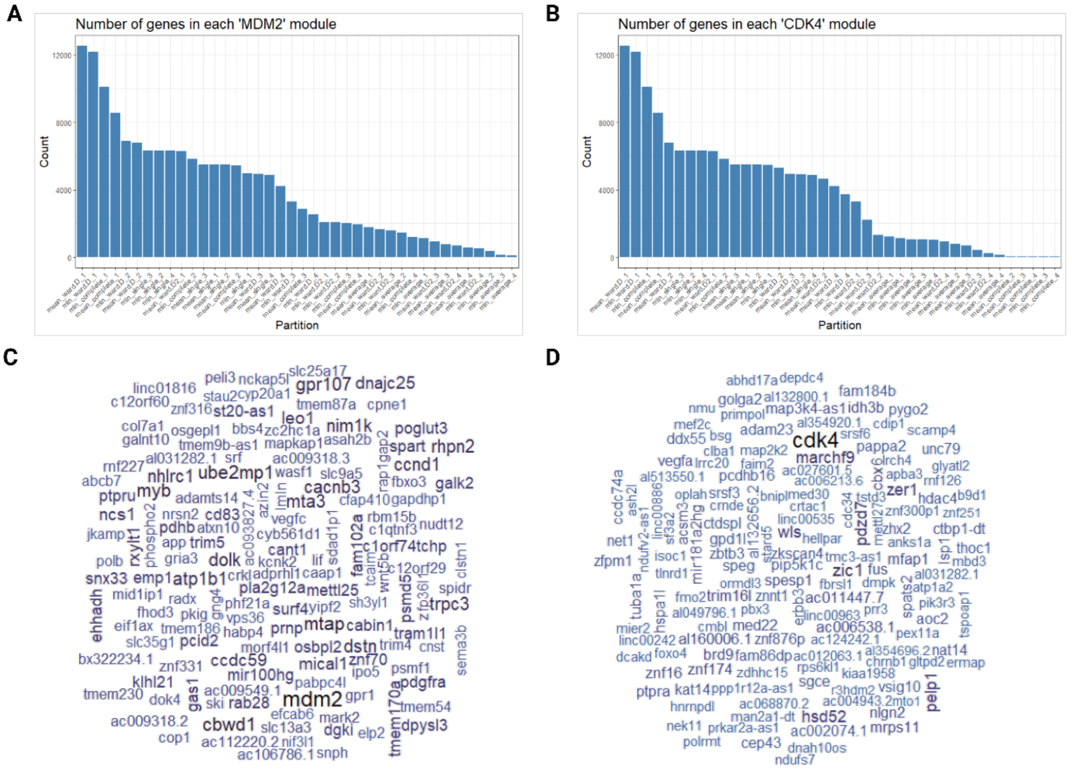


Figure 4.20: Overlap between partitions. Number of genes partitioned into the same module as **A** MDM2 and **B** CDK4 across the different partitions. Wordclouds representing co-occurring genes with **C** MDM2 and **D** CDK4 where darker and larger words have higher occurrence in the same module across partitions.

4.4.2.2 Permutation of module gene assignment

A robust GCN partition would contain underlying patterns of connectivity associated with DDLPS biology and not by random distribution alone. To test this the gene IDs were randomly scrambled, keeping the module number and size constant, creating 200 partitions with randomly shuffled module gene assignments. Comparing the chosen GCN partition (min_complete_3) with the random permutations (***Figure 4.21***), indicated that the chosen partition showed patterns of gene connectivity that are associated with biological relevance and not by random noise alone. These were all tested as significantly higher (*p < 0.05*) in the chosen GCN partition as compared to the random partitions.

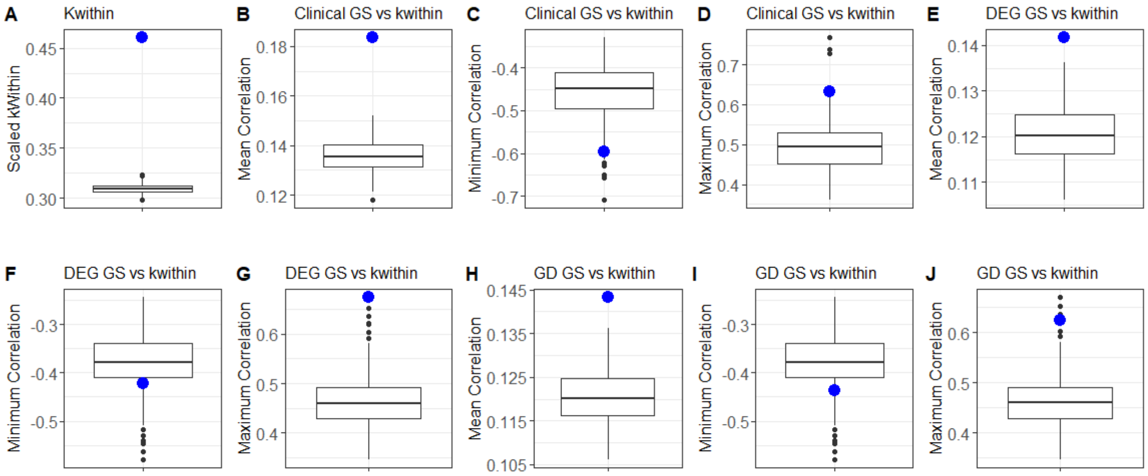


Figure 4.21: Permutation testing for min_complete_3 (blue dot) against random partitions. **A** the intramodular connectivity (k-within). K-within was correlated with each of the three GS measures for each module and summarised according to the max, min, and mean values of their pairwise bicor correlation for **B-D:** Clinical GS **E-G:** DEG GS and **H-J:** Gene Dependency (GD) GS.

4.4.3 Module Preservation Analysis

To validate module co-expression patterns within the TCGA SARC DDLPS GCN module preservation analysis was conducted using the NCC dataset (***Figure 4.22***). To summarise preservation and infer statistical significance, the composite Zsummary preservation score was used. This revealed that in the TCGA SARC DDLPS GCN 199/244 modules (81.5%) with a Zsummary > 2. Of these, 183 (75%) had a Zsummary > 2 < 10 and 16 (6.6%) had Zsummary > 10. There were 45 (18.4%) modules that had a Zsummary < 2 and were not significantly preserved. The gold module had a Zsummary = 1.7, showing that random sampling could not attain preservation levels that were significant suggesting that modules obtaining significance are due to underlying patterns in the data. Module preservation Zsummary shows a weak correlation with module size (Pearson *r* = 0.31), where Zsummary is expected to be dependent on the size of the module.[252]

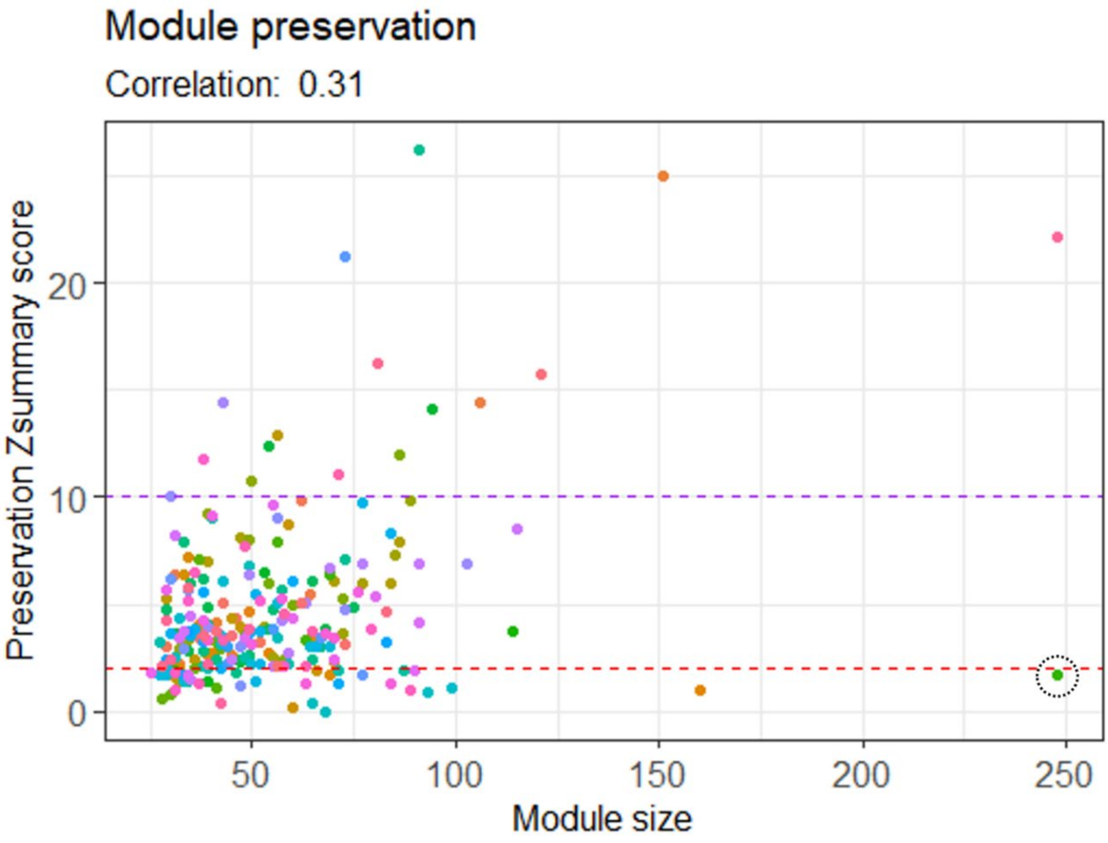


Figure 4.22: Module preservation analysis results as represented by the Zsummary preservation score, a composite score of seven network metrics between the reference (TCGA SARC DDLPS GCN) and the test (NCC) data. Modules are coloured uniquely (n = 244). The black dotted circle represents the “gold” module which is a module of 248 randomly samples genes form the whole network. The red line indicates a 2 < Zsummary < 10 and the purple line indicates a Zsummay > 10. The correlation coefficient (0.31) between Zsummary and module size was calculated using the Pearson correlation.

4.5 Discussion

This chapter defines a robust GCN partition, which will be taken forward in subsequent chapters for module screening. The selected GCN was derived using the TOM min calculation and partitioned using complete linkage hierarchical clustering, with a dynamic tree cut sensitivity of three. The default methodology for WGCNA is to use the TOM min, while the TOM mean is considered an experimental feature.[234] Typically, TOM mean can identify more densely co-expressed modules compared to the TOM min. However, for smaller datasets, TOM min helps protect against outlying or spurious genes by taking the minimum sum of adjacencies.[196]

The observed module numbers across partitions are consistent with the expectations for different linkage types and sensitivity parameters. Both Ward's D1 linkage and single linkages, regardless of sensitivity, yield fewer than 20 unique modules. Single linkage methods, which calculate the minimum distance between clusters, often lead to elongated, chain-like modules. Ward D generates dendrograms like single linkage, as it minimises the sum of squared distances from the cluster mean. Conversely, Ward's D2 typically generates more clusters than D1, as this algorithm calculates distances from cluster centroids rather than the mean. The average method yields a cluster count that falls between that of single and complete linkage. Average linkage was designed to strike a balance between the large clusters seen in single linkage and the reduced sensitivity to outliers observed in complete linkage. Complete linkage, which uses the maximum distance between cluster objects to define clusters, resulted in a GCN partition comprising 245 modules when the tree cutting algorithm's cutting sensitivity was set to three. This outcome is expected, as the maximum distance prevents the formation of large single clusters through object chaining, a phenomenon observed with the single linkage method.[403] However, this approach is more susceptible to errors caused by gene outliers within the pre-cluster, as it calculates the maximum distance.

In assessing the scale-free distribution of gene connectivity among soft-thresholding powers, it was found that using powers greater than 12 resulted in markedly reduced mean and maximum connectivity values. This was expected as higher powers suppress lower correlation values closer towards 0. A power of 12 struck a balance between a more homogenous degree distribution and a scale-free topology with culled gene connectivity values.

Most studies using WGCNA tend to ignore sample network metrics, opting instead to use sample dendrograms followed by a static cut.[249,291] Studies that do consider sample network metrics typically rely solely on the standardised (z-scored) scaled connectivity alone to identify and filter outliers.[404,405] This approach overlooks the clustering coefficient and the pairwise relationship with scaled connectivity leaving valuable network information unevaluated.[346]

A common approach to reduce noise involves to pre-filtering genes used for GCN construction by only taking the top n genes with the highest variance. However, this approach was not adopted here, as it was important to consider all genes passing read count filtering to preserve information regarding gene connectivity. Filtering prior to WGCNA can break the assumptions of a scale-free network, which is why the authors of WGCNA advise against it.[234,256] Additionally, non-uniformity in sample gene expression profiles can alter the results of variance-based filtering, leading to poor gene connectivity distributions among the retained genes.

In this chapter, comprehensive QC is shown to be beneficial. Post-QC the data better fits the assumptions of WGCNA. PCA and hierarchical clustering combined with sample network metrics on TCGA SARC DDLPS gene expression data, effectively identified outliers.

Initial clustering results from HCPC clustering indicated outliers and subsequent exploration identified that cluster 1 was enriched for patients belonging to the M5 methylation cluster from the TCGA publication.[64] In their analysis M5 cluster contained mostly DDLPS and LMS patients that showed poor prognosis. Indeed, after also cross referencing the methylation clusters with the groups identified through sample network metrics revealed that included samples were enriched for this poor prognostic cluster. Immune annotations were also observed in DEGs upregulated in the HCPC cluster 1. Hence methylation data should be inspected in future analyses, especially for any modules retrieving immune annotations.

Outlying samples demonstrated better survival probabilities when corrected for tumour grade and residual margin variables and showed a higher mutational load. Generally, somatic mutation analyses on DDLPS suggest that these tumours are characterised by a low mutational burden compared to other solid tumours.[64] Differences in somatic mutation burden then might be non-consequential and could even stem from inaccuracies/false positives from the calling pipelines used by TCGA SARC. Furthermore, differences in survival outcomes are likely driven by sample size, which makes it difficult to make robust conclusions. All samples here are robustly DDLPS, having not only undergone diagnostic panels, but also an independent review panel for the TCGA study. Removing these outliers significantly improved the underlying scale-free topology of the GCN.[346]

In this project, two key assumptions underpin the decision to remove outlying samples. First, GCNs are assumed to exhibit a scale-free topology. Second, under the first assumption, hub-genes, are believed to play a crucial role in diseased processes.[208,209] Under these assumptions, hub-genes, or other genes associated through co-expression, may represent potential drug targets. By maximising scale-free topology, through the removal of outlying samples, the GCN can better distinguish hubs from non-hubs. This observation is demonstrated in **Figure 4.3** *and* **Figure 4.9,** where a closer adherence to a scale-free topology results in clearer hub definition. Therefore, excluding these samples was deemed beneficial for the goal of identifying drug targets.

In this chapter, several methods for generating GCN partitions are tested on the filtered (DDLPS36) dataset, and the best partition is defined. Disease studies using WGCNA do not typically test node-trait relationships from a QC perspective, nor do they employ random permutations where gene assignments are scrambled to generate random networks.[406-408] Comparing to random networks, a common practice in network science, allows for the comparison of a network to a null model, and helps infer whether networks arise from random stochastic processes or underlying biological processes.

Inherently, scale-free networks display “preferential attachment”, a growth mechanism describing the evolution of such networks. This is often illustrated in social interaction networks, where new nodes entering the network are more likely to attach to existing influential hubs, following a “rich get richer” pattern. In contrast, a random stochastic network should exhibit roughly equal probability for a new node to attach to any existing node.

Assessing random permutations coupled with module preservation analysis, allows us to ascertain the robustness of the network. Here, it is demonstrated that the network partition shows a significantly higher distribution of IMC vs GS and preserves modular patterns of co-expression in an independent dataset.

Finally, module preservation analysis was conducted to validate modular patterns of co-expression in the NCC dataset where it was found that 81.5% of all modules were significantly preserved, and 6.6% showed highly significant preservation. In downstream analysis this will be an important consideration for selecting modules for further investigation and drug target screening. This method was not selected in the assessment of partitions as the primary composite score for preservation, the Zsummary statistic, is dependent on module size. The partitions tested showed various max module sizes (excluding “grey” module) from 1,201 (7.5% of expressed genes) to 12,553 (78.3%) and would not be a fair comparison between partitions.

The main limitation going forward is sample size. This limitation presents a significant challenge to clustering algorithms and hinders our ability to identify the factors driving heterogeneity within the cohort. Despite this limitation, a robust DDLPS GCN partition has been identified and validated using an independent DDLPS disease cohort. Furthermore, using metrics that aim to summarise the module, for example the IMC as used in this chapter to select the GCN partition, or the module eigengene (ME), can aim in reducing noise attributed to low sample number.[251]

Chapter 5 Characterising the underlying biology within the DDLPS GCN

5.1 Introduction

In **Section 4,** a GCN partition with 245 modules of co-expressed genes was constructed using WGCNA on DDLPS RNA-seq data from the TCGA SARC data.[64] WGCNA assumes that modules within this GCN contain functionally related genes with similar biological roles and are under shared regulation.[234] Therefore, it is important to provide biological context to these modules, their genes, and the relationships observed in the GCN.[387]

To understand higher-order relationships between modules, eigengene networks (EGNs) are constructed and the ME-based connectivity (kME), is derived from the correlation of a gene expression profile to the ME expression (see **section 1.2.3**). [251] EGNs have far fewer edges than GCNs (global), providing a streamlined and effective approach to analyse complex gene co-expression data. They are also useful for data integration to explain observed patterns of gene co-expression. A common approach is the integration of functional enrichment analysis (e.g., Gene set enrichment analysis - GSEA, overrepresentation analysis – ORA, etc) which can be displayed using module-term annotations, or using quantified measures of enrichment (e.g., enrichment scores and/or statistical significance). These annotations can be used to label modules and sort them into communities within the EGN in a supervised manner, or alternatively communities can be detected using unsupervised clustering (e.g., Louvain Community Detection) relying on network characteristics (e.g., scaled intramodular connectivity and modularity gain) or other clustering approaches (e.g., k-means).[291,308,409-411]

Modules are selected for further inspection based on a characteristic of interest, called a **gene significance (GS)** measure, which pertains to the significance of a gene in the context of the disease. In **Section 4.1,** three GS measures were proposed and used in **Section 4.4** to identify a robust GCN partition through correlation with intramodular connectivity (IMC). These GS measures are:

1. Correlation of gene expression values to the deviance residuals from the OS multivariate Cox PH model (**section 3.4.3.1**).
2. LogFC signed -log10 FDR-adjusted p-value from differential gene expression analysis of a DDLPS vs WDLPS vs adipose contrast.

3. The mean gene effect scores from DepMap CRISPR-Cas9 inferred cancer vulnerabilities in DDLPS cell lines (LPS141, LPS853, LPS510).[313]

Other studies (as discussed in **Section 1.3**) with similar aims integrate the results from several computational experiments and have been conducted across cancer types.[249,291] Such studies have been conducted in various cancer types.[248,307-310]l[247,311,312]. Two studies in DDLPS have been conducted using such approaches.[249,291] These studies leverage information from various aspects of disease biology to pinpoint key elements, allowing hypotheses on the functional relevance of patterns observed. However, these studies typically only consider the positive correlations to GS measures describing adverse disease phenotypes, pro-tumour, or poor clinical outcomes. Both signs of the correlation are important and describe processes that interplay and result in the disease phenotype.

Recently, a single-cell RNA sequencing analysis in DDLPS (**Section 1.1.3)**[143] revealed transcriptomic profiles in WD/DDLPS similar to adipose stromal progenitor cells (ASPCs), which are pluripotent cells in white-adipose tissue (WATs). This indicates a shared progenitor for WDLPS and DDLPS, through conserved genomic alterations such as *MDM2* amplification. Key differences between the WDLPS and DDLPS TMEs were noted. DDLPS exhibited stronger signals for angiogenesis, proliferation, and metastasis, whereas WDLPS showed stronger signals for adipogenesis. DDLPS was found to possess a more tumour permissive TME, with higher TGF-β signalling and a higher presence of immunosuppressive immune cells. This dataset can be overlayed with the GCN derived from bulk-RNA sequencing to validate and deconvolute gene signatures that correspond to the tumour and its stroma. Furthermore, co-expression of genes and their programmes can be driven by other cells within the microenvironment due to cell-cell communication.[412] These patterns are captured by bulk RNA-sequencing but can be further delineated using scRNA-seq data to infer cell types.

In this chapter, GCN modules are first explored for significant enrichment in biological processes and molecular pathways to provide a broad biological context to the underlying functions in the GCN. The three GS measures are used to rank modules to screen out modules of interest and those corresponding to a favourable outcome are then scrutinised. Subsequently, single-cell data from the Institute Curie will be leveraged to deconvolute the TCGA DDLPS GCN and further explain and validate the observed modular cell type signature associations.[143]

5.2 Chapter aims and objectives.

Hypothesis: A weighed gene co-expression network partition contains modules that represent biologically relevant functions within the tumour and its microenvironment, contributing to disease biology.

Chapter Aims and Objectives:

Aim 1 – Characterisation of DDLPS GCN modular relationships and biology

Objective 1.1: Use the Louvain Community Detection algorithm to highlight inter-module relationships.

Objective 1.2: Use available human gene sets (e.g., from MSigDB) to provide biological context to modules and overlay with results from community detection to identify concordance between approaches.

Objective 1.3: Integrate single-cell RNA-sequencing data to identify cell type associations within the GCN partition.

Objective 1.4: Conduct enrichment analysis on an independent RNA-sequencing dataset to identify any concordance between these findings.

Aim 2 – Identify modules of interest using GS measures and elucidate modular mechanisms.

Objective 2.1: Explore the modular distribution of GS measures in the GCN partition.

Objective 2.2: Use these GS measures to rank and select modules of interest for further exploration.

Aim 3 – Inspect modules associated with a favourable outcome.

Objective 3.1: Perform an in-depth modular enrichment analysis on modules associated with favourable outcomes.

Objective 3.2: Integrate cell type clusters identified in scRNA-seq data to highlight cell type interplay in top-ranked modules.

5.3 Methods

5.3.1 GCN construction and module detection

The WGCNA package and its uses are described in-depth in **Section 1.3.3.** The WGCNA R package was used for the gene co-expression network (GCN) analysis of bulk RNA-sequencing data (TCGA SARC and NCC datasets – **see section 3.4.1**). Filtered (see **section 2.8**) and normalised gene expression counts from the TCGA SARC data[64] were used as input to WCGNA. Quality control was conducted using sample dendrograms, principal component analysis, and sample network metrics as described in **section 4.3.2**. Justifications for the choice of the following WGCNA parameters are outlined in detail in **section 4.4.2.** Gene adjacencies were calculated using the 'adjacency' function, specifying a signed network using the Spearman correlation coefficient and a power of 12. Spearman correlation was chosen to account for gene outliers and account for non-linear gene expression relationships. To build a weighted (see **section 1.3.3**) GCN the adjacency was used as input to the 'TOMsimilarity' function specifying a "signed" network, and a TOMdenom of "min", which is the minimum between adjacency values.[413] The TOM dissimilarity (1 – TOM) was calculated and clustered using a hierarchical clustering algorithm with complete linkage. Modules were defined using a dynamic tree cutting algorithm with a sensitivity parameter (DeepSplit) of four (justifications for these options are provided in **section 4.4.1-4.4.2**).

Given differences in RNA-seq library preparation methods, an additional normalization step was applied to the NCC dataset[114] to ensure compatibility with the TCGA SARC data for downstream analysis. The gene expression matrix was subset according to matching Ensembl gene IDs in the TCGA DDLPS data, resulting in 14,441 genes compared to the 16,032 genes in the TCGA SARC data post-filtering. Subsequent WGCNA methods were identical to those described for the TCGA DDLPS data.

5.3.2 Module eigengene calculation, correlation, and adjacency

MEs were calculated using the 'ModuleEigengenes' function giving a ME expression for each module per patient (for the TCGA DDLPS GCN this derived 245 MEs). The kME (see **section 1.3.3**) was calculated by computing the Spearman correlation between gene expression values and the ME expression. To construct the signed adjacency matrix for the EGN, adjacency values were calculated using the formula $0.5 + correlation(MEi, MEj)^2/2$, where *i* and *j* are modules represented by their ME. The diagonal of the ME adjacency was zeroed to remove self-adjacency values of 1. This adjacency is the eigengene network (EGN) represented as a matrix.

The EGN matrix was then exported as an edge data frame using the ‘exportToVisant’ function (WGCNA package) using an adjacency threshold of 0.8 and subsequently converted to an igraph object using the ‘graph_from_data_frame’ function. The hard threshold of 0.8 for the adjacency was chosen to retain only positive correlations of a high value whilst preserving network structure. The EGN was then projected as described in **section 2.12**, using the VisNetwork R package (version 2.1.2).

5.3.3 Louvain community detection

The Louvain community detection algorithm within the igraph R package (version 2.0.3), using the ‘cluster-louvain’ function, was employed to identify communities within the EGN.[345] This step enables the detection of communities containing modules of shared function and can be paired with downstream results (e.g., GSEA) for interpretation of modular relationships. Initially, nodes are assigned to independent communities and moved locally to a community that maximises modularity (the optimal number of edges or the highest degree density). This process is then repeated at the community level, where nodes represent communities, merging them if it maximises modularity. The process continues until maximum modularity gain is reached. The Igraph implementation processes nodes randomly, making it a stochastic process.[414] The Igraph implementation was favoured over setting a seed for the Louvain algorithm due to its balance of fast computation time and accuracy.[414] In this application, degree represents the ME connectivity, meaning communities represent MEs that are maximally connected. Under the assumptions made by WGCNA, modules represent genes with similar underlying biological functions and shared transcriptional regulation, so ME communities represent modules of correlated function and shared regulation.

5.3.4 Module gene set/pathway annotations

To identify enriched biological functions within network modules, a pre-ranked gene set enrichment analysis (GSEA) was conducted using the fgsea R package[415] (version 1.31.0) and gene sets from MSigDB.[416] These gene sets included “hallmark – C1”, “Reactome – C2:CP subset”, and “GOBP”. These gene sets were chosen based on them describing multiple molecular functions, pathways and biological processes that are often pertinent to disease. The gene list used as input for testing against gene sets included all GCN genes named according to HGNC symbols and ranked according to the kME for each module. Ranking using the kME was preserved and detected fuzzy module-gene relationships and related annotations.

In addition to pre-ranked GSEA, gene overrepresentation analysis (ORA) was conducted using the ToppFun (Transcriptome, Ontology, Phenotype, Proteome, and Pharmacome Annotations based gene list Functional Enrichment Analysis) online tool available as part of the ToppGene Suite (toppgene.cchmc.org).[417] The multiple protein query and annotations tool at STRING (Search Tool For Recurring Instances of Neighbouring Genes), available on the STRING website (string-db.org), was also utilised.[227] These tools allowed for quick assessment of gene lists derived from the module gene assignment (using the module labels to define module gene membership), genes overlapping between modules and single-cell clusters, genes differentially expressed between subsets, and top methylated probes.

5.3.5 Module screening

Module screening was conducted using the GS measures described in sections **4.1** and **5.1.** The Intramodular connectivity (IMC) was calculated as described in **section 4.3.6** and correlated with each of the three GS measures. Justifications for this approach are detailed in sections **4.1** and **4.3.6.** The Spearman correlation value of the IMC versus GS was used to rank the modules. Modules with a higher absolute correlation between intramodular connectivity and GS measures are inferred to have higher biological significance.[234] The selected GS measures detail inferred tumour gene dependencies, transcriptomic reprogramming (via differential gene expression), and insight into the correlation of gene expression with the risk of death over time. Hence, top ranking modules could be seen as “unfavourable”, representing molecular or clinical signatures linked to poor disease outcomes such as increased proliferation (e.g., cell cycle). Conversely, modules negatively correlated with the GS measures may indicate inverse characteristics (e.g., immune activation or apoptosis) and can be seen as “favourable”.

5.3.6 Module network concepts

After identifying modules of interest using the module screening strategy these were explored using gene connectivity network concepts; the maximum adjacency ratio (MAR), the scaled connectivity (SC) and the clustering coefficient (CC) which are recommended by the authors of WGCNA for identifying hub-structures within the module.[387] This was calculated using the ‘fundamentalnetworkconcepts’ function from the WGCNA R package. Using these metrics can highlight patterns of connectivity within the module and allude to the module hub gene by degree centrality. Intramodular connectivity patterns are important to network analysis as genes with higher connectivity may be of importance to the biological function represented in the module.

5.3.7 Methylation data

Methylation data from the Illumina 450k platform and methylation probe annotations available in the TCGA SARC were acquired using the TCGA biolinks R package.[328] Methylation expression and annotation objects were filtered first by a mapping quality of >60 at both mapped sites, where a mapping quality of 60 represents unique mapping[401]. Probes were selected based on having annotated gene names that matched the TCGA SARC DDLPS processed RNA-seq data (**see section 2.6**) from available annotations. Probes with a mapping quality of < 60 and no matching gene annotation were removed from further analysis. Differentially methylated probes were identified according to the 100 most variable probes among the samples.

To integrate the methylation data with the GCN, the absolute Spearman correlation between the probes CpG methylation value and the MEs was calculated, identifying both positive and negative correlates. The top 200 correlated probes with MEs were then filtered to those having a matching HGNC gene symbol annotation. These top variable and correlated probes were inspected for the relationship between gene expression and methylation. A function was created to calculate the Spearman correlation between gene expression and CpG methylation values, along with a p-value. Additionally, the sequence to which these probes maps (start and end – available in the annotations) was extracted, and a scatter plot was generated for visualisation.

5.3.8 Genomic contexture of methylation probes

To identify the genomic position of the 450k methylation probes, and assess whether these are within enhancer/promoter regions in genes of interest the NCBI Genome Viewer online tool (available at https://www.ncbi.nlm.nih.gov/gdv/browser/genome/) was used.[418] Probe positions were taken from the gene annotation data (hg38) available from the TCGA SARC project.[64] The beginning and end positions were manually tagged on the NCBI GDV. The Following tracks were loaded:

Intronic coverage: describes the relative coverage of genomic DNA reads mapping to these positions.

Exonic coverage: describes the relative coverage of genomic DNA reads mapping to these positions.

CpG islands track: displays known high frequencies of CpG sites.

Enhancer regions: identifies potential intragenic regulatory elements that may be methylated to regulate gene expression.[419]

5.3.9 ScRNA-seq analysis, module overlap and exploration.

10X Genomics single-cell RNA-sequencing data on 11 DDLPS patient samples was kindly provided by Sarah Watson at the institute-curie, including 28029 cells prior to filtering.[143] Data retrieved included cell type annotations as described in the published material and were used in this analysis. Single cell read data was processed and filtered as described in **section 2.16**. All single-cell data was analysed using the Seurat R package (version = 5.1.0) [348] Red blood cells were removed from the cell pool leaving 27446 cells. This scRNA-seq data is the first available for human patient DDLPS samples and is a novel opportunity to deconvolute the DDLPS GCN and identify cell-specific co-expression patterns. To integrate the single-cell and WGCNA results, the overlap between cell cluster genes and WGCNA module genes were identified. This was achieved using an overrepresentation test and the reported Fishers exact statistic via the GeneOverlap R package (version 1.38.0). If the intersection between gene sets was found to be less than two genes, the significance was manually zeroed to prevent significant results with low levels of overlap. A -log10(p-value) of 1.3 was considered significant. Results from this analysis were visualised using the pheatmap R package (version 1.01.2).

Gene expression for the given genes in the single-cell data were visualised using the ‘DotPlot’ and ‘Featureplot’ functions from Seurat, with the assay set to “SCT” for the featureplot function. To inspect the genes expression of genes overlapping between modules and single-cell data, the ‘intersect’ function from base R was used and passed to DotPlot or FeauturePlot for visualisation. For genes of interest, tumour cells were extracted from the single-cell data using the provided annotations, and cells were subset according to high or low gene expression values. High or low expression values were decided based on the distribution of gene expression values as inspected on a histogram. Differentially expressed genes (putative cluster markers) between these subsets were identified using the ‘FindMarkers’ function in

Seurat, employing the default Wilcoxon rank sum test. Only those passing a threshold of 0.5 logFC were considered. Genes were further filtered according to an average logFC > 2 and a Bonferroni adjusted p-value of 0.01, ensuring that only genes highly expressed with high significance in a specific cluster were identified.

5.3.10 Cell signatures

Fibro-adipogenic progenitor (FAP) (skeletal muscle) signatures were obtained from a systematic literature review.[420] These were split into two types, embryonal and adult FAP signatures. Both positive marker sets were tested. Adult FAPs included *CD34*, *PDGFRA, CD201 (PROCR), CD166* (*ALCAM*)*, CD105 (ENG), CD90* (*THY1*), *CD73* (*NT5E*), *CD15 (*FUT4*), COL1A1, TCF4.* Embryonal signatures were *PDGFRA, DCN, FN1, LUM, OSR1, POSTN, FAP, THY1, VIM, NT5E, COL1A1, COL1A2, COL3A1, PTN, OGN,* and *FBLN5*. Tumour associated endothelial and pericyte signatures were derived from multiple sources in the literature. TEC signatures were taken as (*CD276, TEM7 (PLXDC1), CD31 (PECAM1), VEGFR2 (KDR), VWF, NESTIN (NES), CD133 (PROM1)).*[421] Transdifferentiation/vascular mimicry was assessed based on the expression of CD31 (*PECAM1*), CD34, and *CDH5*.[422,423]

Endothelial-to-mesenchymal transition signatures were taken as *ACTA2, FSP1* (*AIFM2*)*, COL1A1, COL1A2, FBN1, Vimentin* (*VIM*)*, and MMP2.*[424] It was noted that CD31 and VWF and VE-cadherin are decreased in endothelial-to-mesenchymal transition (EndoMT) although negative expression markers were not assessed. CAF markers were taken as *ACTA2, FAP, S100A4, PDGFRA, PDGFRB,* and *Vimentin (VIM)*.[425] Classical adipose-derived MSC signatures were taken from previous studies and include CD73 (NT5E), CD90 (THY1), CD105 (ENG), CD166 (ALCAM), and CD44.[426,427]

5.3.11 Single-cell Co-expression analysis

The co-expression tables analysis (COTAN) R package was used to calculate gene co-expression within the TME.[428] COTAN utilises contingency tables to compute the gene co-expression across cells using raw RNA read counts. The ‘proceedToCoex’ function was used on the raw RNA counts. The ‘calculateCoex’ function was used to calculate gene co-expression setting ‘actOnCells’ as FALSE as gene co-expression values were desired and not cell-cell correlations. The diagonal was set to zero to remove self-correlations in the ‘getGenesCoex’ function.

5.4 Results

5.4.1 Characterising GCN communities

5.4.1.1 Cancer hallmarks are enriched within the DDLPS GCN

The MEs were calculated for the TCGA DDLPS GCN revealing seven Louvain communities (***Figure 5.1A***). Correlation among MEs, in accordance with the Louvain communities, reveals clusters of Mes that may be enriched for similar biological processes (***Figure 5.1B***). To explore the biological processes/pathways these modules recapitulate, pre-ranked (by kME) GSEA was conducted using Reactome, GOBP and Hallmark gene sets. An overview of the top ten most enriched Reactome pathways (by NES) is displayed in ***Table 5.1.*** The cell cycle emerges as a strongly enriched process particularly in M201 and M100 (***Table 5.1***). In general, these processes were heterogenous although repeated elements were noted (***Appendix A.4***)

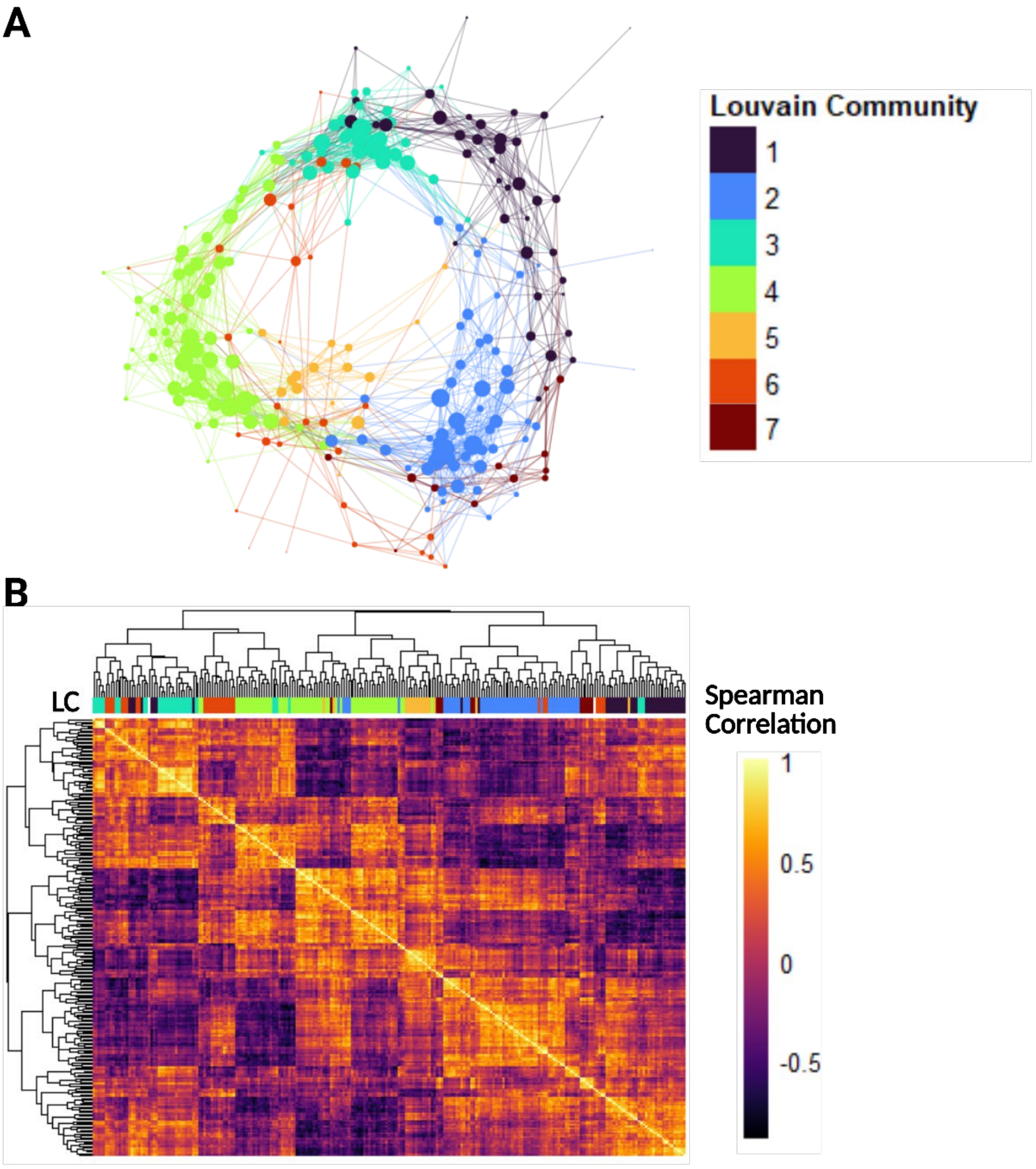


Figure 5.1: Module eigengene (ME) analysis. **A:** EGN with results from calculations to detect Louvain communities within the network, as depicted by the assigned colour. ME edges were defined according to having an adjacency > 0.80. **B:** Hierarchical clustering of Spearman correlation values for MEs with annotations for associated Louvain communities (LC).

Table 5.1: Top ten Reactome pathways enrichments

Module label	Reactome ID	Reactome name	Size	-Log10(p-value)
M201	R-HSA-69278	Cell cycle mitotic	502	87.796159
M100	R-HSA-69278	Cell cycle mitotic	502	81.875461
M178	R-HSA-1640170	Cell cycle	618	80.545186
M142	R-HSA-69278	Cell cycle mitotic	502	78.697535
M86	R-HSA-168249	Innate immune system	838	77.445364
M56	R-HSA-156842	Eukaryotic translation elongation	89	74.947877
M106	R-HSA-72613	Eukaryotic translation initiation	116	73.127866
M141	R-HSA-72766	Translation	288	73.016042
M53	R-HSA-156842	Eukaryotic translation elongation	89	68.577138
M150	R-HSA-72766	Translation	288	64.633429

M – Module, size – the size of the gene set. -Log10(p-value) is the Benjamini-Hochberg adjusted p-value.

The top five enrichments were visualized onto the EGN for the Reactome (***Figure 5.2A***), GOBP (***Figure 5.2B***), and hallmark (***Figure 5.2C***) gene sets, identifying a strong concordance between annotations and Louvain ME communities. The enrichments across these gene sets suggest an importance for translation, immune, extracellular matrix (ECM) remodelling, molecule transport, cell cycle, and metabolism. These functions are expected in cancer biology and outline many well-defined cancer hallmarks.[429] Furthermore, the most frequent amplifications in DDLPS cluster mostly according to their known gene functions (***Appendix A.5***).

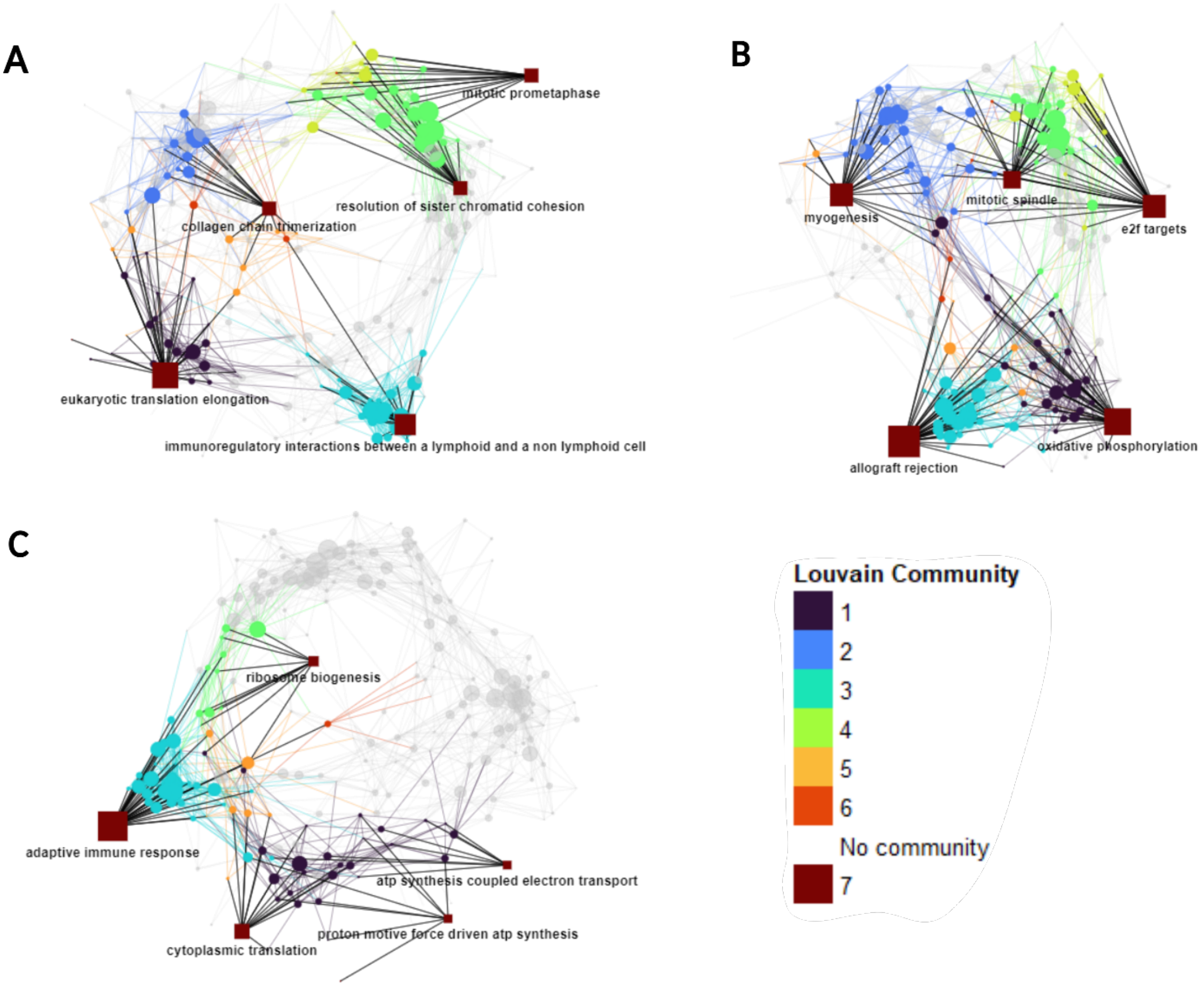

A
mitotic prometaphase
resolution of sister chromatid cohesion
collagen chain trimerization
eukaryotic translation elongation
immunoregulatory interactions between a lymphoid and a non lymphoid cell
B
myogenesis
mitotic spindle
e2f targets
oxidative phosphorylation
allograft rejection
C
ribosome biogenesis
adaptive immune response
atp synthesis coupled electron transport
proton motive force driven atp synthesis
cytoplasmic translation
Louvain Community
1
2
3
4
5
6
No community
7

Figure 5.2: Top five most frequent module enrichments from **A:** Reactome. **B:** Gene Ontology Biological Processes (GOBP) and **C:** Hallmark gene sets. Module eigengenes (MEs) are indicated by a circle, whose size is proportional to the ME degree connectivity, and colour indicates Louvain community. Squares indicate an enriched gene set, where size is proportional to the number of modules that term is enriched.

WGCNA conducted independently on NCC RNA-seq data revealed a weighted GCN that was partitioned using the same parameters as the TCGA DDLPS network into 243 modules. MEs were calculated and Louvain communities were detected based on an adjacency >0.8, revealing four communities and projected as an EGN (***Figure 5.3A***). Network structure indicates a relationship which appears more binary (nodular) as compared to the TCGA DDLPS, confirmed when looking at the correlation between all MEs (to rule out this was not an effect of hard thresholding the adjacency) (***Figure 5.3B***). Subsequently, enrichment analysis on the NCC GCN partition retrieved far fewer significant modular enrichments (***Table 5.2***). For GSEA on GOBP gene sets, only 11 modules retrieved 22 significant enrichments (top positive and negative) with five unique gene sets. The most frequent GOBP gene set was found to be neuron projection regeneration (n = 10) most enriched in NCC M61 (NES = 2.16).

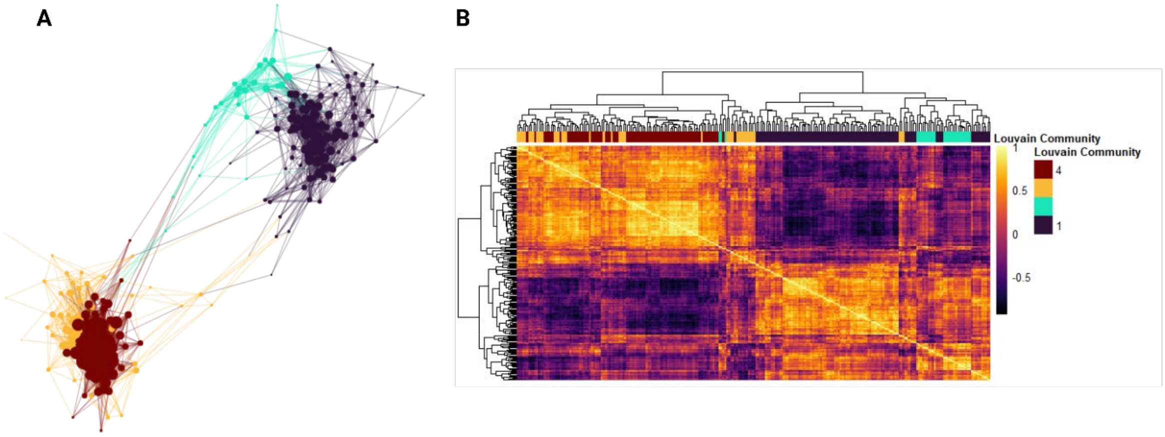


Figure 5.3: **A** Eigengene network (EGN) using results from WGCNA analysis on NCC RNA-seq data passing an adjacency of 0.8 and coloured according to Louvain community. **B** Heatmap of ME Spearman correlation coloured according to Louvain community detection algorithm.

Table 5.2**:** Most frequent enriched top gene sets for the NCC GCN.

Reactome	Reactome ID	Frequency	GOBP	GOBP ID	Frequency
Integration of energy metabolism	R-HSA-163685	14	Neuron projection regeneration	GO:0031175	10
Regulation of Insulin secretion	R-HAS-422356	12	Positive regulation of intrinsic apoptotic signalling pathway	GO:0008630	6
Class A 1 rhodopsin-like receptors	R-HAS-373076	4	Biomineral tissue development	GO:0031214	2
-	-	-	Ephrin receptor signalling	GO:0048013	2
-	-	-	Organic acid metabolic process	GO:0006082	2

GOBP – Gene ontology biological process gene sets.

5.4.1.2 Inferred enrichment for single-cell cluster marker gene overlap in GCN shows high concordance with GSEA and LC results

Single-cell RNA sequencing data on 11 DDLPS samples was leveraged to better understand and contextualise module relationships with the wider **TME** (***Figure 5.4A***).[143] Single-cell overlap analysis indicated 66/245 modules showing significant overrepresentation (Fisher's exact) for cell cluster marker genes in the TCGA DDLPS GCN (***Figure 5.4B***) and the NCC DDLPS GCN (***Figure 5.4C***). These analyses reveal a rich TME with multiple cell types and suggest that gene co-expression profiles within the GCN correspond with those cell type signatures, which line up with gene set annotations (***Figure 5.2***). The most frequent cell type represented in the GCN was tumour cells across 22 modules, followed by myeloid cells in 14 modules and endothelial cells (ECs) in 12 modules. The overlap among Louvain community analysis, gene set enrichment analysis, and single-cell cluster overrepresentation indicates strong concordance. Like GSEA results, performing this analysis on the NCC GCN (***Figure 5.4C***) reveals a more dichotomised relationship among ME correlation with cell types.

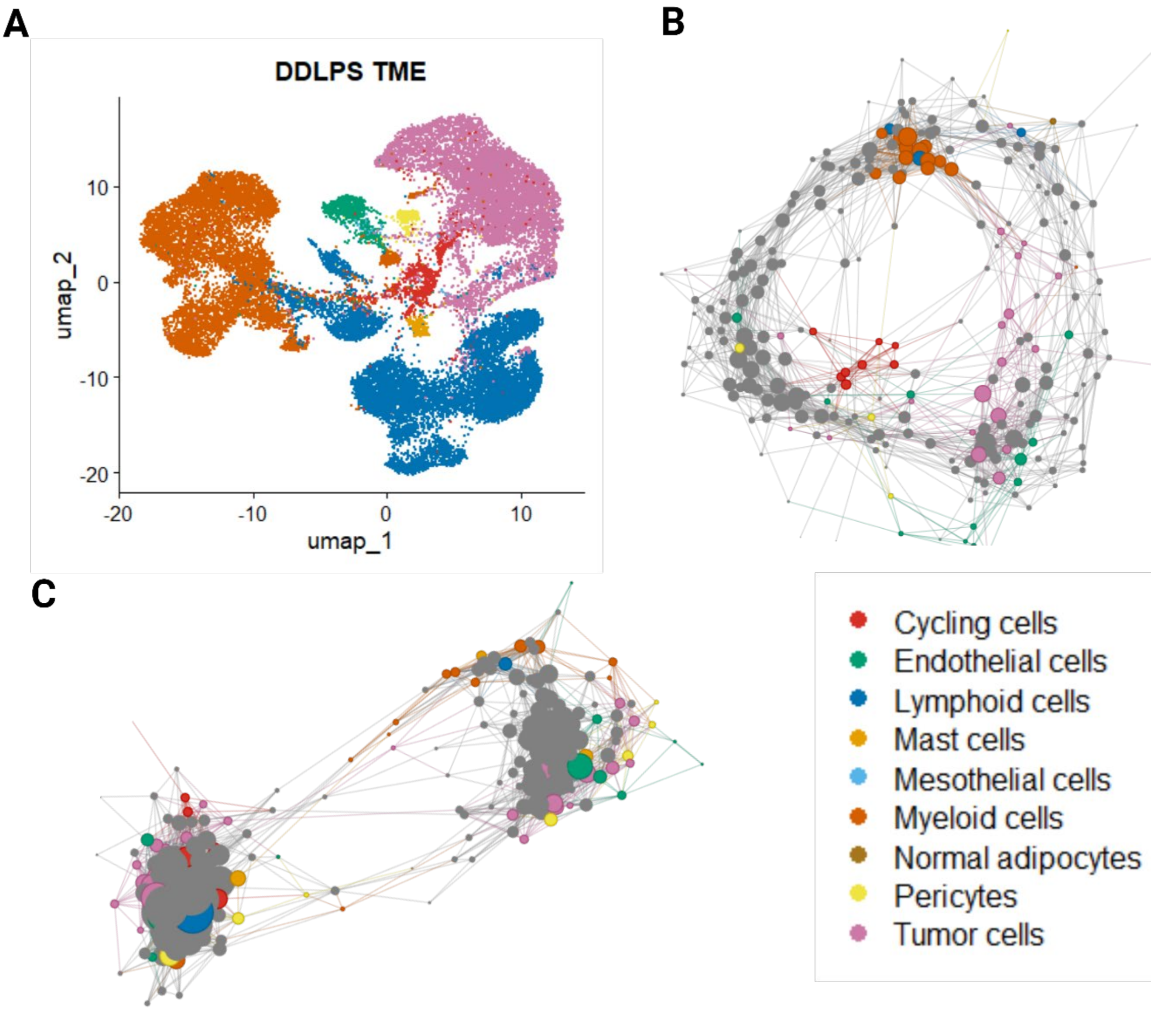


Figure 5.4: DDLPS single-cell cluster marker integration with the TCGA DDLPS GCN.**A** Projection of the first two dimensions of a UMAP reduction for 27446 cells. **B-C:** EGN with significant and highest module enrichment for cell cluster marker genes for **B** TCGA DDLPS and **C** NCC DDLPS. node size is proportional to the degree centrality. Colour indicates the cell type annotation according to broad cell type designations.

5.4.2 Module screening using GCN metrics

Modules were screened using the correlation between intramodular connectivity (IMC) and GS measures (the CorDeviance as described in **section 4.3.3- *Figure 5.5A***), the logFC signed -log10(p-value) from a differential expression data (**see section 4.3.4 – *Figure 5.5B***) and the gene effect score from DepMap (***Figure 5.5C***). Modules were then selected based on a combined ranking metric (sum rank of each metric) (**Figure 5.5D; Appendix A.6**). The highest-ranking module was M241, a module associated with cell cycle processes, which along with other top-ranked (combined ranking metric) modules are discussed in detail in **section 6**.

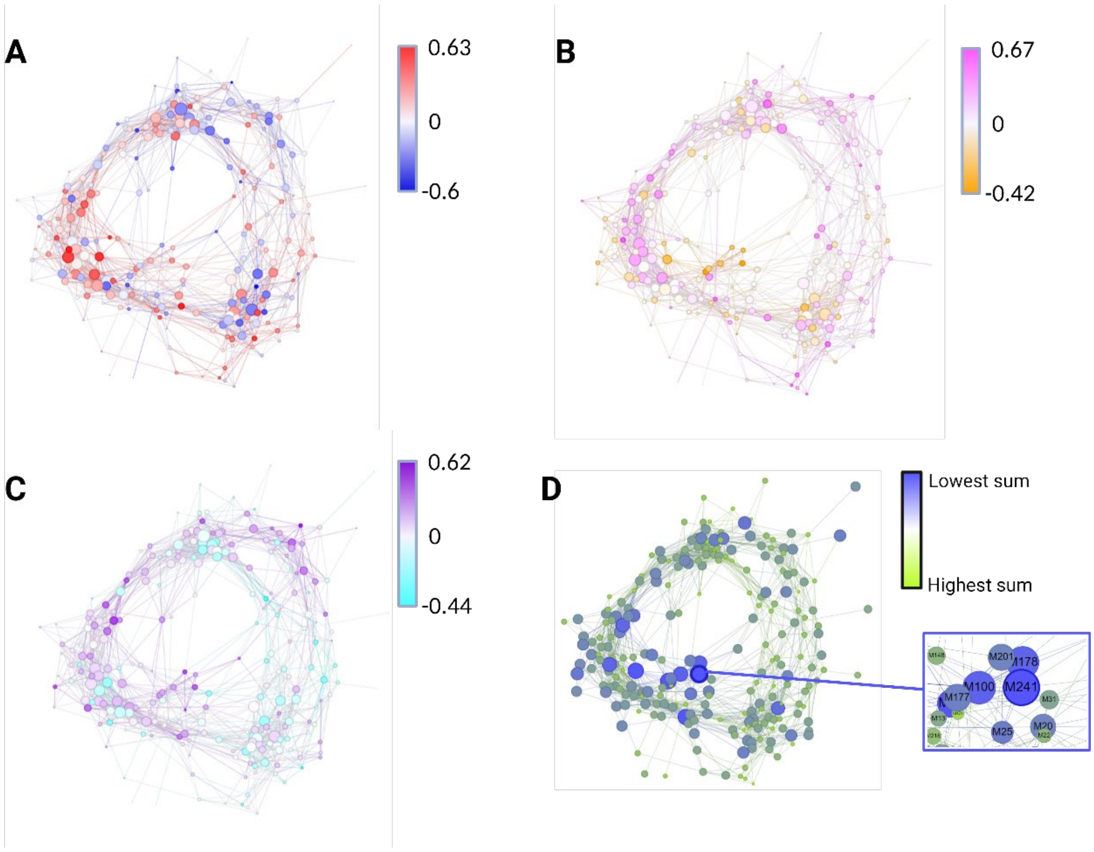


Figure 5.5: Module screening to search for module of interest where intramodular connectivity (IMC) correlated positive or negative with **A:** The correlation of each gene expression to the deviance residuals from the TCGA DDLPS Cox proportional hazards model derived in **section 3**. **B:** The logFC signed -log10(FDR adjusted p value) from a differential expression experiment on DDLPS vs WDLPS and **C:** The mean gene effect scores from three DDLPS DepMap cell-lines (LPS141, LPS510, and LPS853) depicting gene essentiality. The ranks of these individual metric were summed to find modules of interest. The highest ranked modules are displayed inside the zoom box in **D** where M241 showed the highest overall rank (1st), M173 (not shown) was the lowest ranking module with a rank of 245 (the number of modules).

5.4.3 Lowest ranking modules

5.4.3.1 A diverse range of disease processes are enriched in the lowest ranked modules

Herein, the lowest-ranking modules will be referred to as favourable modules, of which the top ten (from highest to lowest GS rank) are M173, M236, M234, M45, M194, M54, M98, M2, M131, and M107 (***Table 5.3***). All these modules, except for M194 (Zsummary = 1.64), showed significant preservation (Zsummary > 2) of co-expression patterns between the TCGA DDLPS and the NCC DDLPS data. M173 had the highest Zsummary preservation score of the favourable modules of 10.13. Hence, these modules show conserved transcriptional programmes (and by inference their downstream translated biological processes) between independent DDLPS gene expression data.

Table 5.3: Top favourable modules according to the module screening using the correlation between the intramodular connectivity (IMC – kwithin) and the GS measures.

Module label	kWithin (IMC)	Clinical GS	DepMap GS	DD vs WD GS	Sum of GS ranks (overall rank)	Zsummary
M173	0.4351	-0.3412	-0.2598	-0.1522	669	10.13
M236	0.4984	-0.1237	-0.2273	-0.3051	658	2.01
M234	0.4960	-0.0990	-0.0569	-0.3363	587	4.02
M45	0.4575	-0.2421	-0.1564	-0.0696	580	5.89
M194	0.5619	-0.3174	-0.0865	-0.0732	568	1.64
M54	0.4986	-0.5329	-0.2028	0.0282	564	2.43
M98	0.3654	-0.2610	-0.2179	0.0003	562	4.62
M2	0.5036	0.0518	-0.1395	-0.2998	561	3.55
M131	0.4383	-0.3421	-0.1566	0.0131	554	4.69
M107	0.4269	-0.1277	-0.0164	-0.1621	546	7.41

IMC – intramodular connectivity. GS – gene significance, Zsummary – a Z-test composite of preservation statistics permuted with random sampling.

To identify the enriched functions, GSEA was conducted on module genes ranked by the kME. The enrichments identified were found to be diverse (as demonstrated for the top ten enrichments) (***Figure 5.6A***; ***Table 5.3***). The most significant and strongly enriched gene set (ranked by NES and *p-value*) was the GOBP adaptive immune response (*NES* = 3.48, *-log10p* = 73.8) in M45. In total, three of the ten modules (M173, M45 and M54) were shown to have enrichments in immune-related processes (***Figure 5.6A; Table 5.3***). This shared enrichment may also explain why these three modules also display higher connectivity between them compared to other modules grouped within this GCN sub-graph (***Figure 5.6A***).

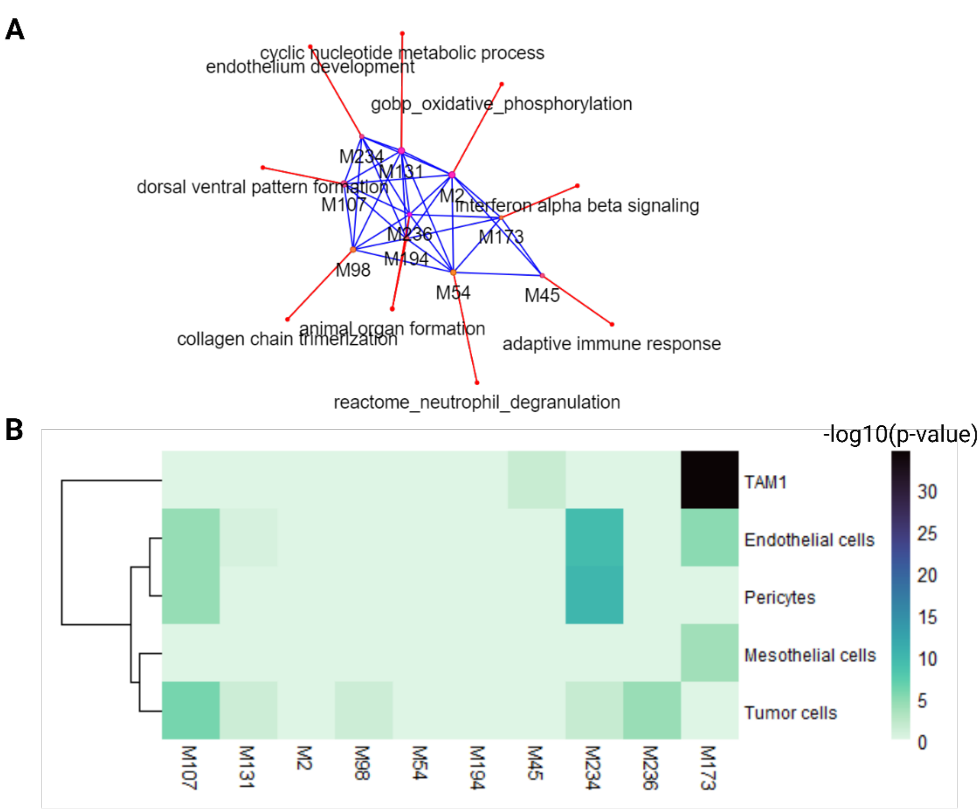


Figure 5.6: A EGN of the top ten favourable modules and their strongest (by NES) GOBP and Reactome enrichment. Blue edges indicate ME connections and red edges indicate their top pathway. **B** Overrepresentation analysis for module gene overlap with single-cell cluster gene markers. Colour indicates the -log10(p-value) of the overrepresentation test.

Leveraging scRNA-seq data reveals significant enrichment for inferred single-cell cluster markers (***Figure 5.6B***), which reveals tumour associated macrophage M1 (TAM1) related gene expression signatures in M173 and M45, endothelial cells (M107, M173, M234, and M131), pericytes (M107 and M234), mesothelial cells (EC) (M173), and tumour cells (M107, M131, M98, M234, and M236). TAM1 signatures suggests an inflammatory environment (IFN signalling) which would be consistent with these modules being favourable due to the signature for immune activation.

Notably, there is a close association between tumour cells and cells of the vasculature within the same modules (e.g., M107, M234, and M131 – see ***Figure 5.6B***) indicating a tight co-expression between genes expressed in different cell types. This could be attributed to tumour angiogenesis through either endothelial sprouting or trans-differentiation/vascular mimicry processes, which have been reported in WDLPS.[430] Further to this, the gene expression profile of pericytes bares a strong similarity to tumour cells particularly in DDLPS characteristic SCNAs including *MDM2*, *CDK4*, *HMGA2* and *FRS2* (***Figure 5.7***). Assessing markers typically associated with DDLPS, mesenchymal cells like fibroblasts (particularly an activated fibroblast-like phenotype), MSCs, and tumour ECs (TECs) reveals several observations (***Figure 5.10***). Firstly, tumour cells possess high expression for MSC/CAF markers and express CD34, a vascular EC/haematopoietic stem cell marker also used to identify cancer stem cells (CSC). Secondly, ECs express high levels of known EC markers but also show expression for mesenchymal markers, indicating endothelial to mesenchymal transition (EndoMT).

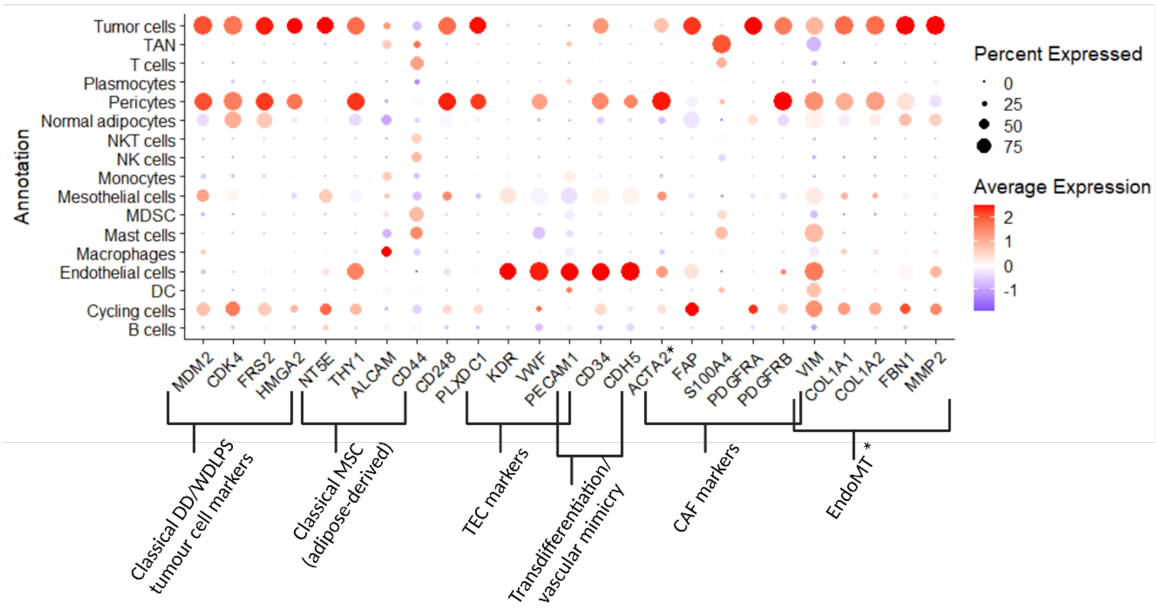


Figure 5.7: Single-cell RNA sequencing expression of gene signatures corresponding to different disease/differentiation processes. *- Indicates endothelial to mesenchymal transition (EndoMT) marker. Figure created using BioRender.com.

kME (***Figure 5.8A***) and ME correlations (***Figure 5.8B***) among top ten favourable modules show distinct clusters of modules that are connected through conduit modules (e.g., M45). Single-cell enrichment displays differing cell type signatures between these modules (with some overlap), implicating heterogenous cell roles in the TME (***Figure 5.6B***). M173, M234 and M107 all contained genes that are expressed in ECs but only M107 and M234 show significant correlations (***Figure 5.8B***). This could represent heterogenous cell types in the TME or modular biological pathways.

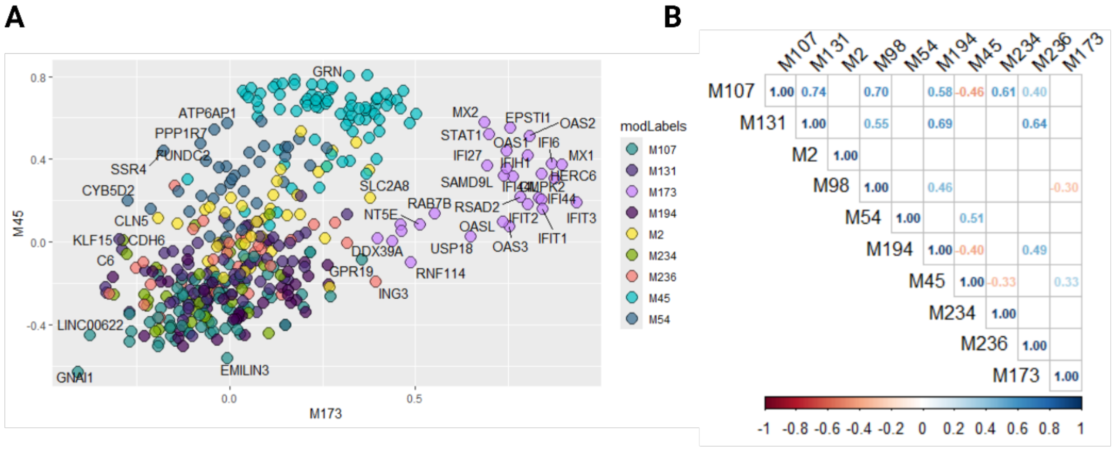


Figure 5.8: **A** Plot detailing kME – gene module membership values for genes within the top ten positive modules. Colour indicates module label. **B** Significant Spearman correlation between module eigengenes. Colour indicates the correlation values. Non-significant Spearman correlations are not shown.

The top-ranked favourable module is M173 (***Figure 5.9A***) which is most strongly enriched for Interferon signalling of both type I - α-β (NES = 3.22, -*log10p* = 21.4) and type II -γ (NES = 3.03, -*log10p* = 18.89) pathways, among other immune-related pathways/processes (***Figure 5.9B)***. M173 contains 15 genes in the interferon type I α-β signalling, and 5 genes for the IFN-γ response. The co-expression of these genes suggests that this module represents IFN-induced transcriptional programmes.

M173 is correlated with M45 (*rho* = 0.44, *p* = 0.008 – ***Figure 5.8B; Figure 5.9C***), which is enriched for adaptive immune response (NES = 3.48, -*log10p = 73.87*) and antigen presentation processes (***Figure 5.9D***), specifically 'immunoregulatory interactions between a lymphoid and non-lymphoid cell' (NES = 3.43, -*log10p* = 32.00). Additionally, *C-type lectin domain family 4 member E* (*CLEC4E*)*,* a gene within M45, is typically associated with a sterile inflammation response that is notably higher in M1-like macrophages.[431]

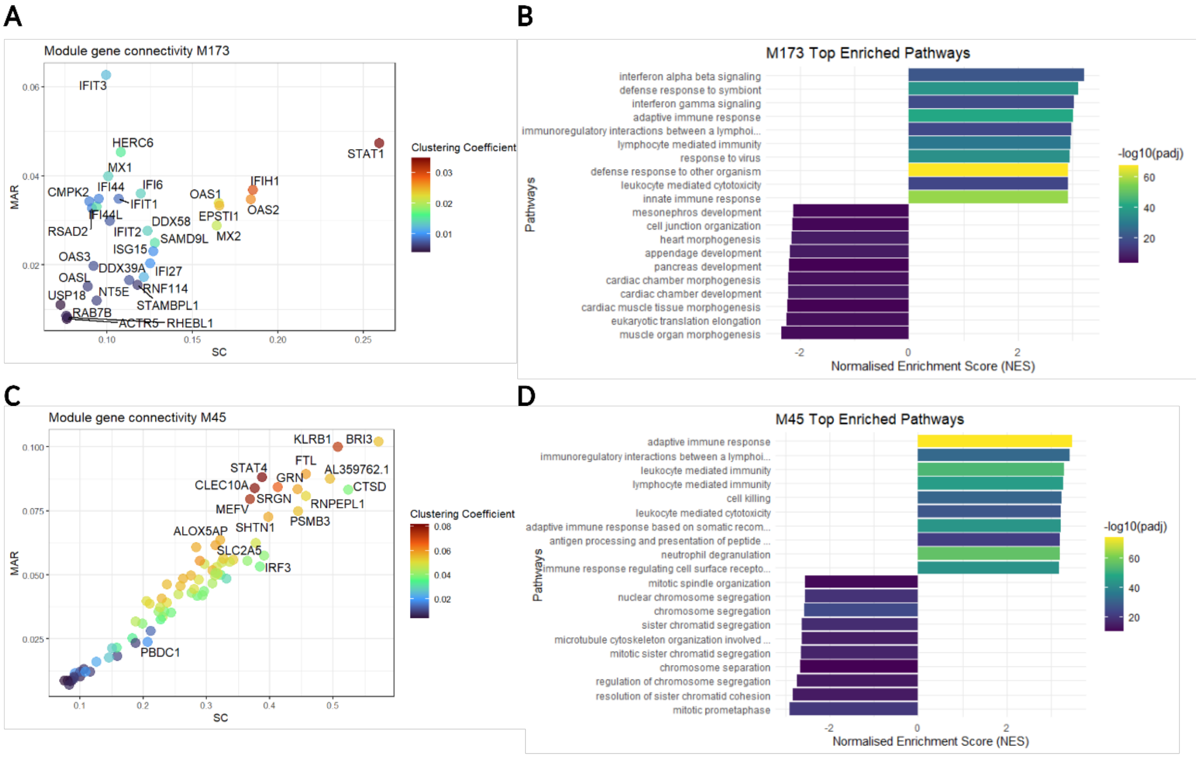


Figure 5.9: Module analysis results for M173 (**A-B**) and M45 (**C-D**). Module connectivity for **A** M173 and **C** M45 detailing the SC – scaled connectivity and the MAR – maximum adjacency ratio within the module coloured by the clustering coefficient. Combined Reactome and GOBP (Gene Ontology Biological Processes) gene set enrichment results for **B** M173 and **D** M45. Bars are coloured by the -log10 adjusted p-value (filtered for -log10 adjusted p > 1.3) and sorted according to the top and bottom ten processes by NES – normalised enrichment score.

There is no strong association between M173 and modules annotated with T-cell markers M12 (*CD3, FOXP3, GZMB, PD1*) (*rho = 0.37, p = 0.026*) and M164 (*ITGAL, GZMK, ITK, CD8, CD247*) (*rho = 0.38, p = 0.023*), although still significant. Single-cell data suggests M173 and M45 (albeit with a lower value) contain genes that are most prominently expressed in M1-like tumour associated macrophages (TAM1) (***Figure 5.6B***). Other favourable modules indicated ECM (M98), cholesterol metabolism and transport (M194), tumour-associated vasculature/angiogenesis (M236, M234), differentiation processes (M131, M107) and EndoMT signatures (M234) (***Appendix A.6***).

5.4.3.2 TME inflammation signatures correlated to cholesterol/lipid metabolism and angiogenesis

There was a strong interferon signature identified in the top favourable modules most notably in M173 through IFN response genes including *IFIT1, IFIT2, IFIT3, IFI27* in a variety of cell types (***Figure 5.10A***). These cell types include macrophages (of the TAM1 phenotype) but also in tumour cells (e.g., *IFI27, IFI6, IFI44*), ECs (e.g., *IFI27, IFIT1, IFIT2,* and *IFIT3)* mesothelial cells (e.g., *IFI6, IFI27, IFIT3*), and pericytes (*IFI27*). This module alone indicates inferred IFN inflammation present within the TME.

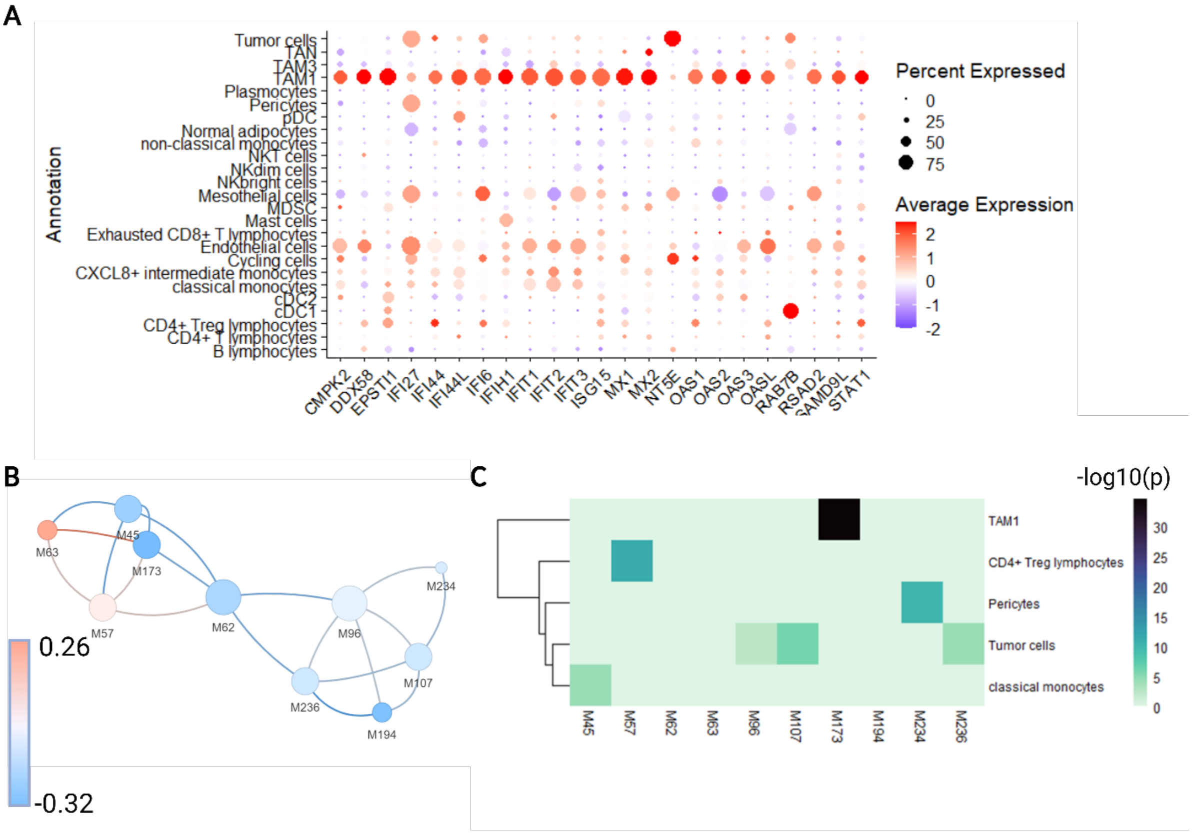


Figure 5.10: **A** Overlap of genes between M173 and single-cell data, average expression is indicated by colour, and the size of the dots indicate percentage of expression in that cell type. **B** EGN sub-graph of top five favourable modules and their five immediate neighbours. Colour indicates, **C** Heatmap for enrichment of cell types in modules where colour indicates the -log10(p-value).

To explore this relationship further, the top five favourable modules were used as seeds in a random walk with restart (RWR) approach to identify the five most immediate neighbouring modules (***Figure 5.10B; Figure 5.10C***). These five nearest neighbouring modules are M62, M63, M96, M107 and M57. From previous analysis M45 was found to be enriched for antigen presentation processes (***Figure 5.9D***), detailing active antigen presentation and a pro-inflammatory signature. Antigen presentation was also identified in M57 (***Figure 5.11A, Figure 5.11B***) which contains several MHC class 1 related genes (*HLA-A, HLA-B, HLA-C, HLA-E, HLA-F, TAPBP* etc) (***Figure 5.20C***). The most enriched Reactome pathway for M57 was Adaptive immune response (NES = 3.49, -log10padj = 76.91). M234 and M236 were enriched for functions in cell differentiation and EndoMT processes respectively with expression of genes in vascular cells (***Appendix A.7***).

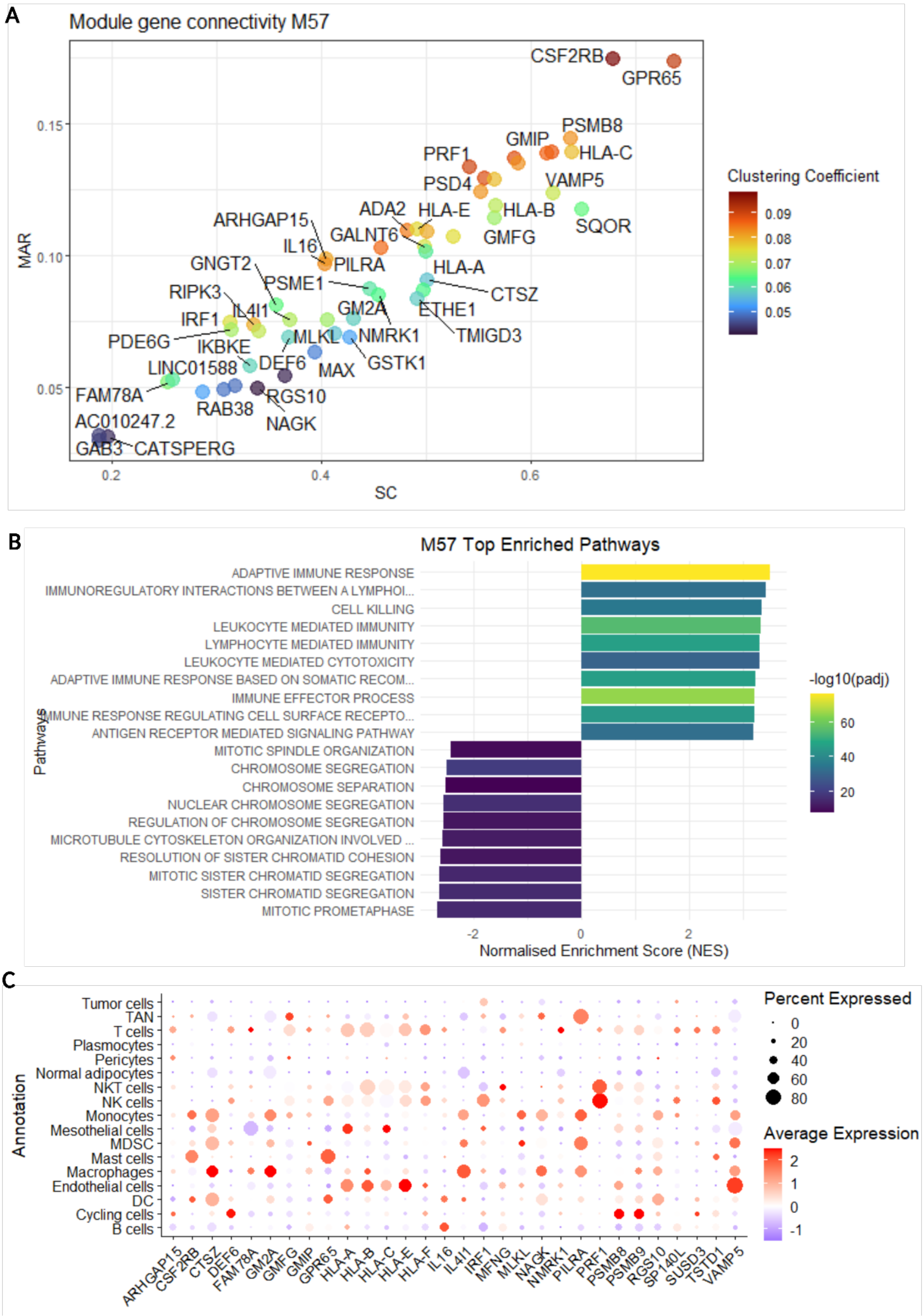


Figure 5.11: M57 module analysis. **A** Module connectivity graph showing the SC – scaled connectivity, and MAR – Maximum adjacency ratio, coloured by the clustering coefficient. **B** Reactome and GOBP enrichment analysis results passing a -log10 adj p > 1.3, displayed are the top ten positive and negative results ordered by NES – normalised enrichment score, coloured by -log10 adj p value. **C** Intersect of M57 genes with single cell RNA-seq cell clusters, labelled according to “global” annotations. Figure created in BioRender.com.

Another module in the M173 neighbour network (***Figure 5.10B***) is M62 (***Figure 5.12A***) as a conduit between modules associated with IFN-responses or antigen-presentation and lipid/cholesterol metabolism, along with ECMO and angiogenesis signatures (e.g., M107, M236). M62 was found to be enriched for IFN α-β response (NES = 2.49, -log10p = 7.09) (***Figure 5.12B***) among other immune-related gene sets using the kME as a gene list ranking metric. M62 by partition contained no genes included in the Interferon signalling gene set, suggesting the ME expression of M62 is correlated with interferon response gene expression within other modules.

Genes within M62 suggest a role in cholesterol and lipid homeostasis and metabolism (e.g., *CETP*, *CEBPA*, *NR1H3* (LXRα), *GPX4, GPBAR1*), prostaglandin synthesis (*PTSG1 (COX1)*) immune cell migration and infiltration (e.g., ADGRE5 (CD97) and *ADRB2*) and angiogenesis/ECM interactions (e.g., *CD248*).[432] *NR1H3* is important for the upregulation of cholesterol efflux molecules, including ABCA1 and ABCG1 (M238 and M155 respectively), to which M62 shows a strong correlation (*rho = 0.59, <0.001 and rho* = 0.54, p < 0.001 respectively). M155 contains genes that are highly expressed in tumour associated neutrophils, corresponding to the role of cholesterol metabolites in recruiting immunosuppressive cells.[432] *CETP* encodes a cholesterol ester transferase, which was found to show highest gene expression in vascular cells (ECs and pericytes) (***Figure 5.12C***).

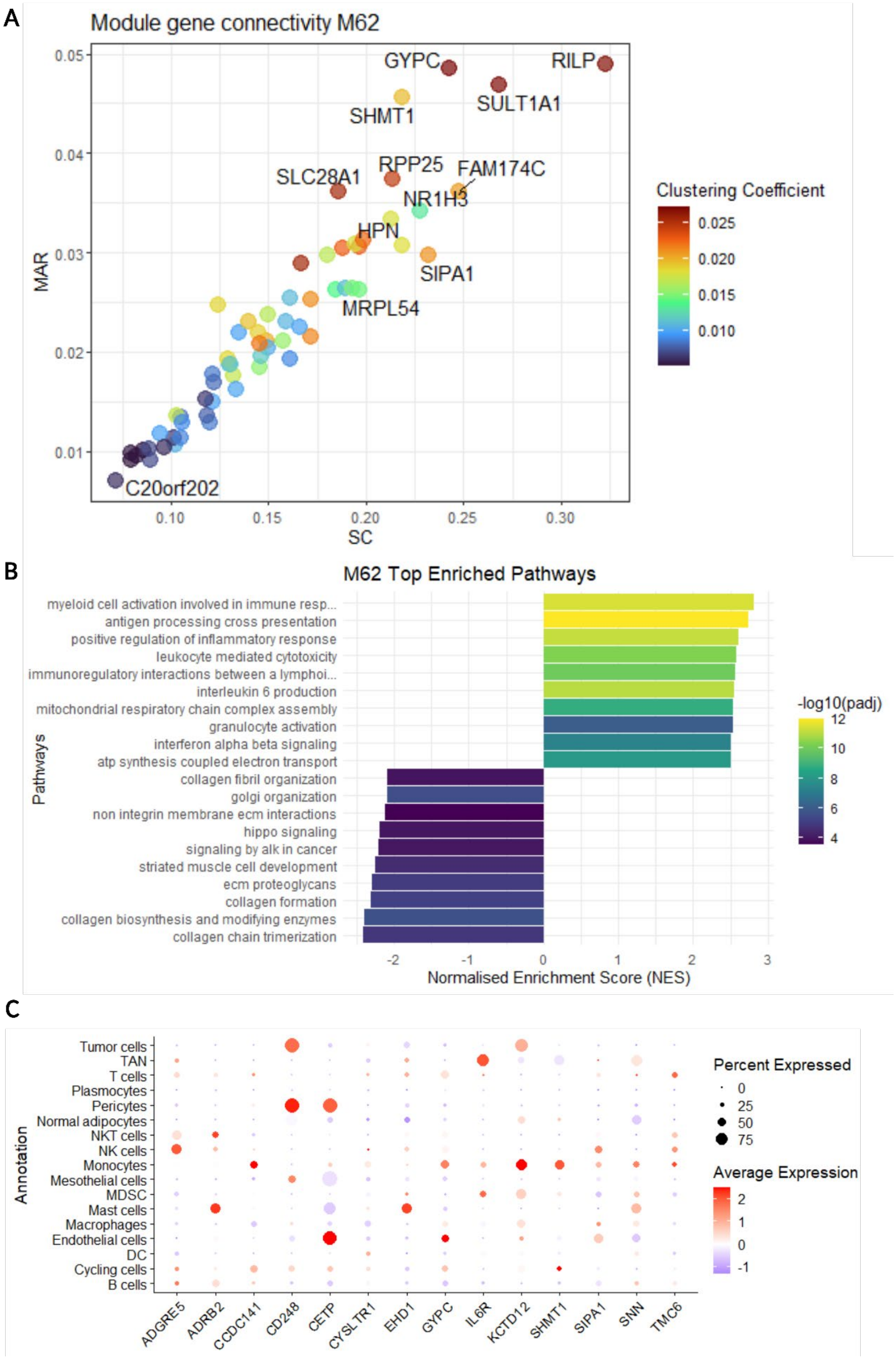


Figure 5.12: M62 module analysis. **A** Module connectivity graph showing the SC – scaled connectivity, and MAR – Maximum adjacency ratio, coloured by the clustering coefficient. **B** Reactome and GOBP enrichment analysis results passing a -log10 adj p > 1.3, displayed are the top ten positive and negative results ordered by NES – normalised enrichment score, coloured by -log10 adj p value. **C** Intersect of M62 genes with single cell RNA-seq cell clusters, labelled according to "global" annotations. Figure created in BioRender.com.

A direct neighbour to M62 in the GCN sub-graph (***Figure 5.10B***) is M96 (***Figure 5.13A***), which showed enrichments (***Figure 5.13B***) in DNA strand elongation (NES =2.15, -log10padj = 3.33), regulation of endothelial cell differentiation (NES = 2.02, -log10p =2.5) and regulation of cell migration involved in sprouting angiogenesis (NES = 1.87, -log10p =2.13). Overlapping genes (***Figure 5.13C***) with single-cell data reveals ABC transporters (*ABCA10, ABCA6, ABCA9*) expression in tumour cells among other genes implicating this relation to lipid homeostasis, adipogenesis and angiogenesis (e.g., *EBF2* and *NRP1*) and a correlation to the function of M62 (including GPBAR1). Together (*ARDB2* and *NRP1* may contribute to modulating endothelial cell function and causing vasodilation, which would further drive inflammation within the TME.

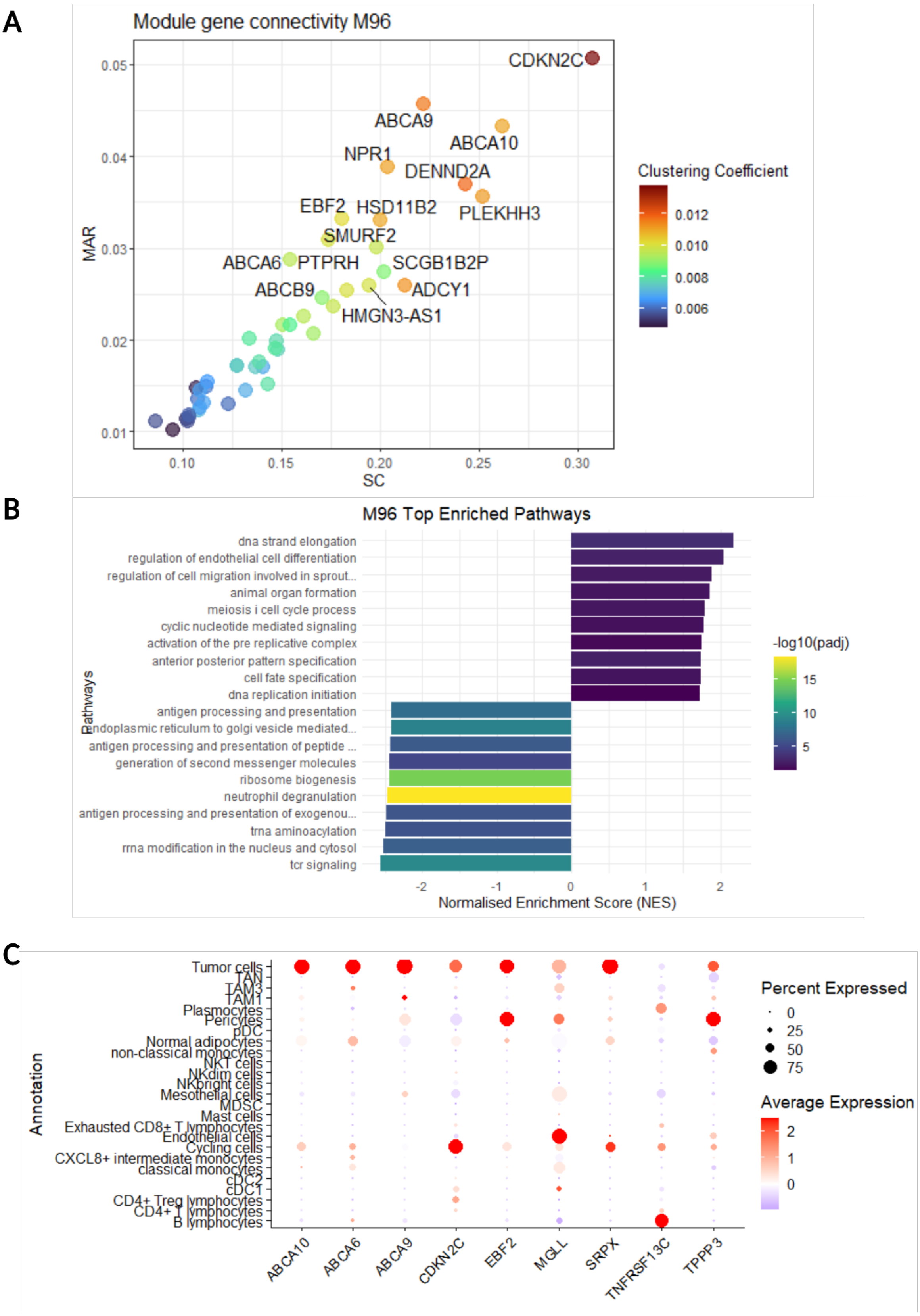


Figure 5.13: M96 module analysis. **A** Module connectivity graph showing the SC – scaled connectivity, and MAR – Maximum adjacency ratio, coloured by the clustering coefficient. **B** Reactome and GOBP enrichment analysis results passing a -log10 adj p > 1.3, displayed are the top ten positive and negative results ordered by NES – normalised enrichment score, coloured by -log10 adj p value. **C** Intersect of M96 genes with single cell RNA-seq cell clusters, labelled according to "global" annotations. Figure created in BioRender.com.

M96 shows a positive correlation to M61 (*rho = 0.41*, *p = 0.014*), which contains several genes that are highly expressed in ECs (***Figure 5.14A, Figure 5.14B***), involved in cholesterol metabolism (influx) (*SCARF1* and *PCSK9*) and regulation of angiogenesis (*CLEC14A*, *FLT4, TIE1*) or other endothelial functions/phenotypes (*VWF*). Top enrichments for M61 indicate GOBP endothelium development (*NES* = 2.59, -*log10p*adj = 12.4).

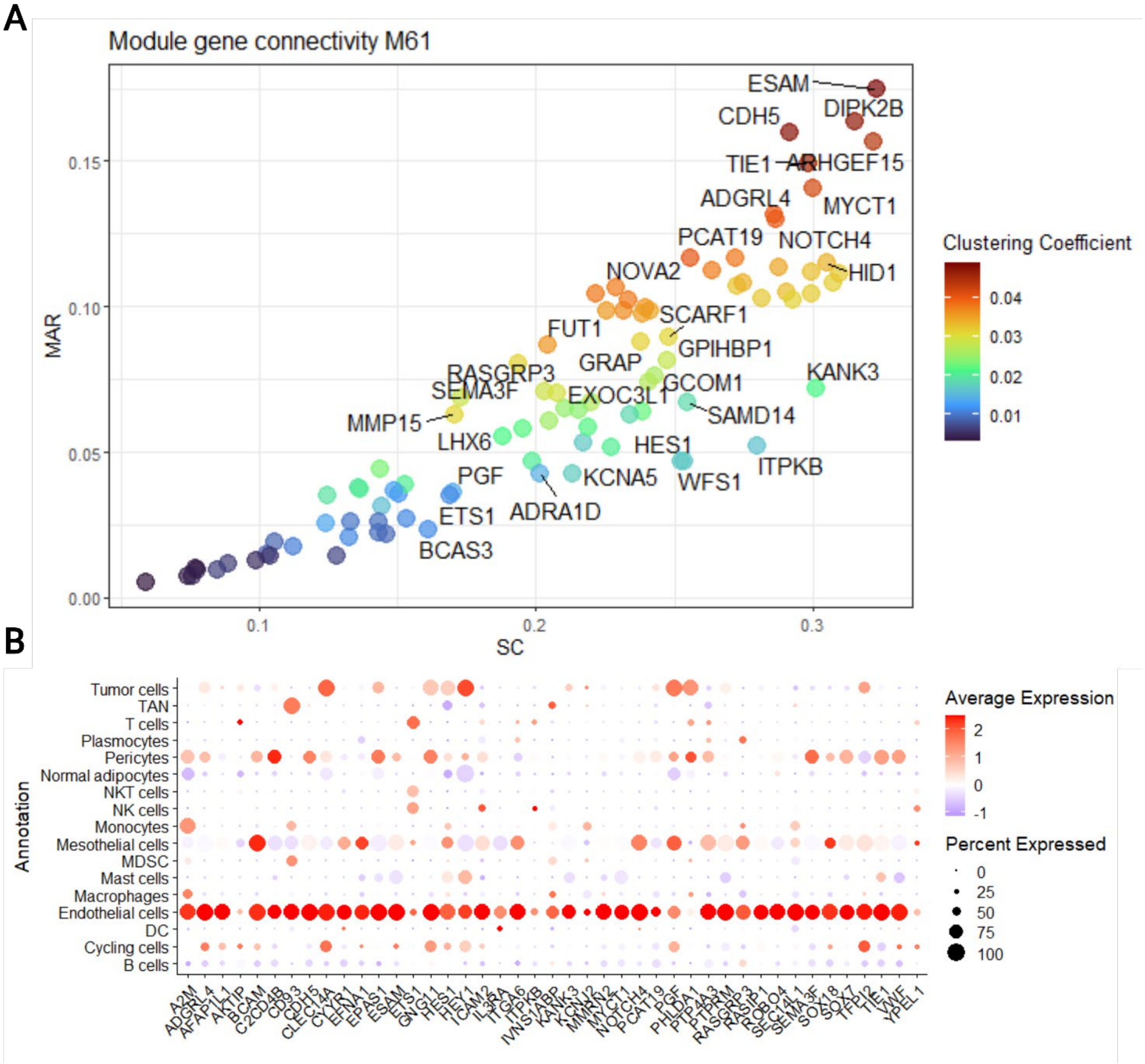


Figure 5.14: M61 module analysis. **A:** Module connectivity graph showing the SC – scaled connectivity, and MAR – Maximum adjacency ratio, coloured by the clustering coefficient. **B:** Intersect of M61 genes with single cell RNA-seq cell clusters, labelled according to “global” annotations. Figure created in BioRender.com.

Another module in the GCN sub-graph is M63, which contains the *ACAT1* gene encoding an enzyme required for cholesterol esterification and thus the production of cholesterol esters. Notably, M63 is the only module with a weak positive correlation to the clinical GS score (*rho* = 0.26), inferred as being correlated to a poor outcome (***Figure 5.10B***). M63 shows a moderate positive correlation to M62 (*rho = 0.39, p = 0.018*).

VEGFA is a potent angiogenic molecule and was clustered into M97 (***Figure 5.15A***), a module which is not present in the GCN sub-graph, but is correlated to M62 (*rho = 0.53, p* < *0.001*). M97 contains enrichment for DNA strand elongation (*NES = 2.38, -log10padj* = 4.42) but also contains genes involved in fatty acid uptake and generation of free fatty acids (e.g., *CD36 (FAT), LIPE, SLC27A3*) and adipogenesis/lipid metabolism (notably *PPARG*) (***Figure 5.15B***). Corresponding to this M97 also showed enrichment for GOBP fatty acid beta oxidation (*NES = 2.09, -log10padj = 3.63*). M107 (***Appendix A.7***) was found to be enriched for genes related to angiogenesis (***Appendix A.7)***. It was also found to be enriched for genes in aquaporin mediated transport (NES = 1.59, *log1-padj* = 1.89) (***Figure 5.16A***), where aquaporins are known to be associated with angiogenesis and fibrosis. Gene expressed are largely inferred to be from tumour, EC and pericyte clusters (***Figure 5.16B***).

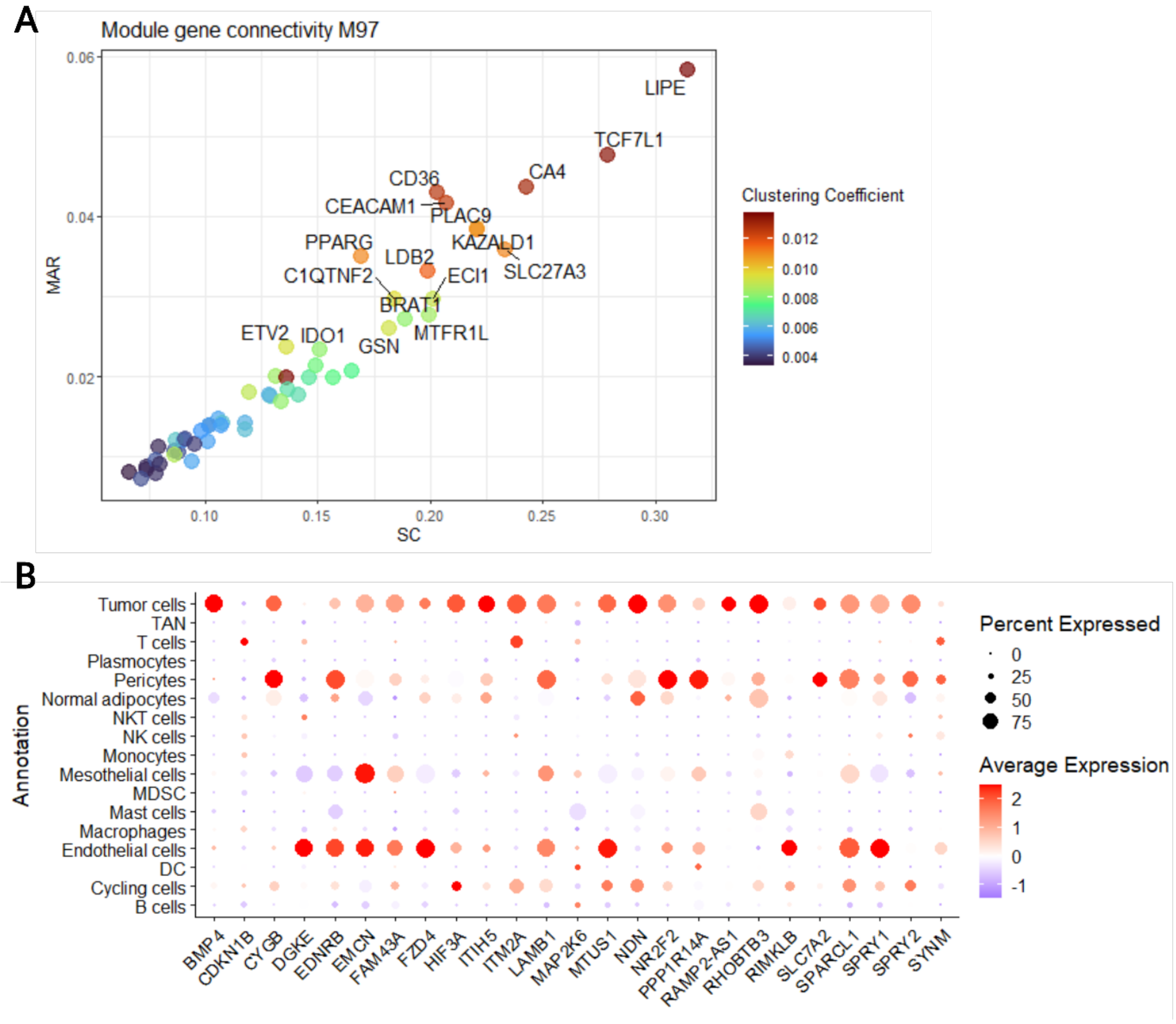


Figure 5.15: M97 module analysis. **A:** Module connectivity graph showing the SC – scaled connectivity, and MAR – Maximum adjacency ratio, coloured by the clustering coefficient. **B:** Intersect of M97 genes with single cell RNA-seq cell clusters, labelled according to "global" annotations. Figure created in BioRender.com.

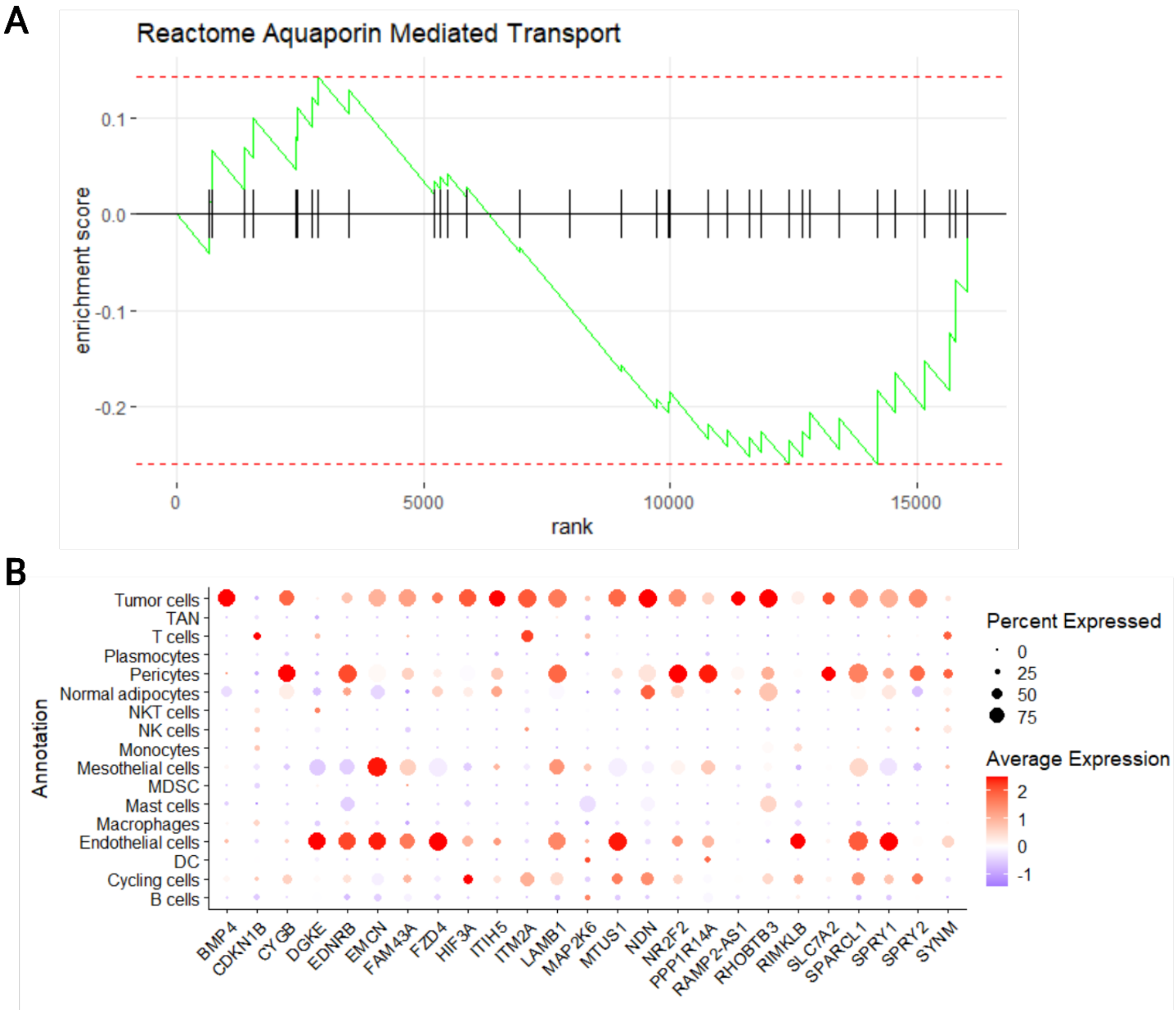


Figure 5.16:**A:** Enrichment plot for pre-ranked GSEA for GCN genes using the Reactome Aquaporin Mediated Transport pathway gene set. Genes were ranked according to their kME for M107 **B:** Intersect of M107 genes with single cell RNA-seq cell clusters, labelled according to "global" annotations. Figure created in BioRender.com.

Regions of the network sub-graph (M96, M234, M236, M194 and M194 – containing *LCAT*) and their correlations (e.g., M61, M97, M63 and M162) are enriched for cholesterol/lipid metabolism genes along with angiogenesis and ECM remodelling, expressed in various cells across the TME, particularly ECs, pericytes, and tumour cells. These signatures are linked, through M62, to interferon response signatures observed in M173 and direct neighbours (M45, and M57) primarily in tumour cells, ECs, pericytes and macrophage (TAM1-like).

Increased intake/biosynthesis of lipids, including cholesterol, promote an inflammatory environment. High-levels of cholesterol can induce ER-stress, activating *XBP1 (*M162*)* target genes, including *CTLA4* and *PDCD1*.[432] There is a positive correlation between M62 and M162 (*rho = 0.45, p = 0.006*) and a weak, non-significant correlation with M12 (*rho = 0.26, p = 0.12*).

However, M162 is significantly correlated with M12 (*rho = 0.58, p < 0.001*). These results suggest that WD-like fatty regions of the DDLPS tumour, through increased metabolite generation, contribute to a chronically inflamed TME.

5.4.3.3 Tumour fibrosis driven by IFN signalling

Stimulation through is known to mediate a multitude of molecular programmes. To explore the transcriptional effect of IFN response in the TME, several of the IFN response genes noted in M173 for their expression levels in tumour cells (***Figure 5.17A***) were further examined. Most notable of these were *Interferon alpha inducible protein* 6 (*IFI6*) and *Interferon alpha inducible protein 27* (*IFI27*). *IFI6* has been studied across a variety of cancers and has been implicated in resistance to immunotherapy, with mesenchymal-like ESCC tumours expressing IFI6 showing worse overall survival and extensive ECM remodelling.[433] *IFI6* (***Figure 5.17B***) displayed a notable cut-off for high and low expression levels corresponding to the 95th percentile, where populations of tumour cells show high expression levels.

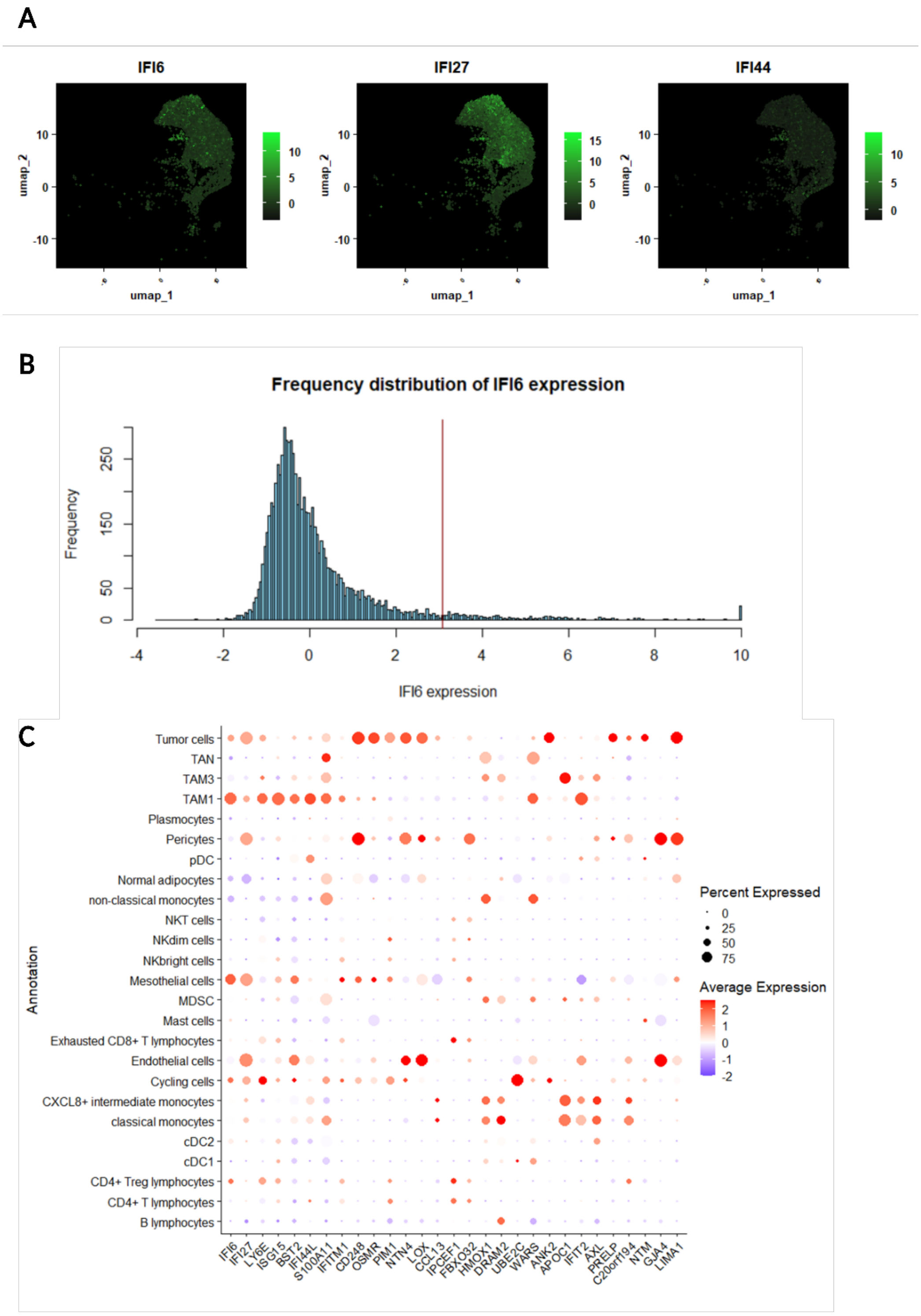


Figure 5.17: A: Expression of IFN response genes from M173 showing highest expression in tumour cells. **B:** IFI6 distribution of expression among tumour cells. Red line indicates the 95th percentile cut point used to subset the data. **C** Differentially expressed genes according to the IFI6 subset. Colour indicates the average expression and size of the dot indicates the percent of cells expressing this gene.

Partitioning tumour cells by *IFI6* expression levels (cut by the 95th percentile of expression) and filtering to highlight top differentially expressed genes (logFC > 1.5 and *padj < 0.01*) identifies genes enriched for ECM organisation (***Figure 5.17C***). These include *CD248* (M62)*, FAP, GSN*, *COL1A1, LAM1*, *LUM*, and TGFβ receptors (*TGFBR3*). TGFβ signalling has been postulated to drive phenotypic change in adipocyte progenitor cells towards a DDLPS phenotype, primarily through the inhibition of adipogenic programmes and the depletion of cholesterol/lipid metabolites.[143] TGFβ signalling is known to upregulate *CD248* expression in stromal cells[434] and lead to intracellular cholesterol accumulation via MEK-ERK1/2 signalling phosphorylating *SREBP2*. Cholesterol is a factor in a positive feedback loop for TGFβ internalisation (via lipid raft mediated endocytosis) and increased TGFβ expression.[435]

COTAN gene co-expression analysis reveals co-expression of both *IFI6 (***Figure 5.18A***)* with ECM organisation and remodelling genes. Twenty-one of the top 100 genes were included in the Reactome Extracellular Matrix Organisation gene set, which was the most significantly enriched pathway (*NES* = 2.11, -log10p = 30.36; ***Figure 5.18B***). This suggests extensive ECM remodelling with downstream proteins involved in matrix deposition (e.g., *PCOLCE,* which encodes a C-endopeptidase enhancer 1 and processes procollagen into mature triple helical collagen fibrils, and *GSN,* Type 5 and 6 *COL* genes, *FBN1*, among others). ECMO genes are co-expressed with the inflammation response.

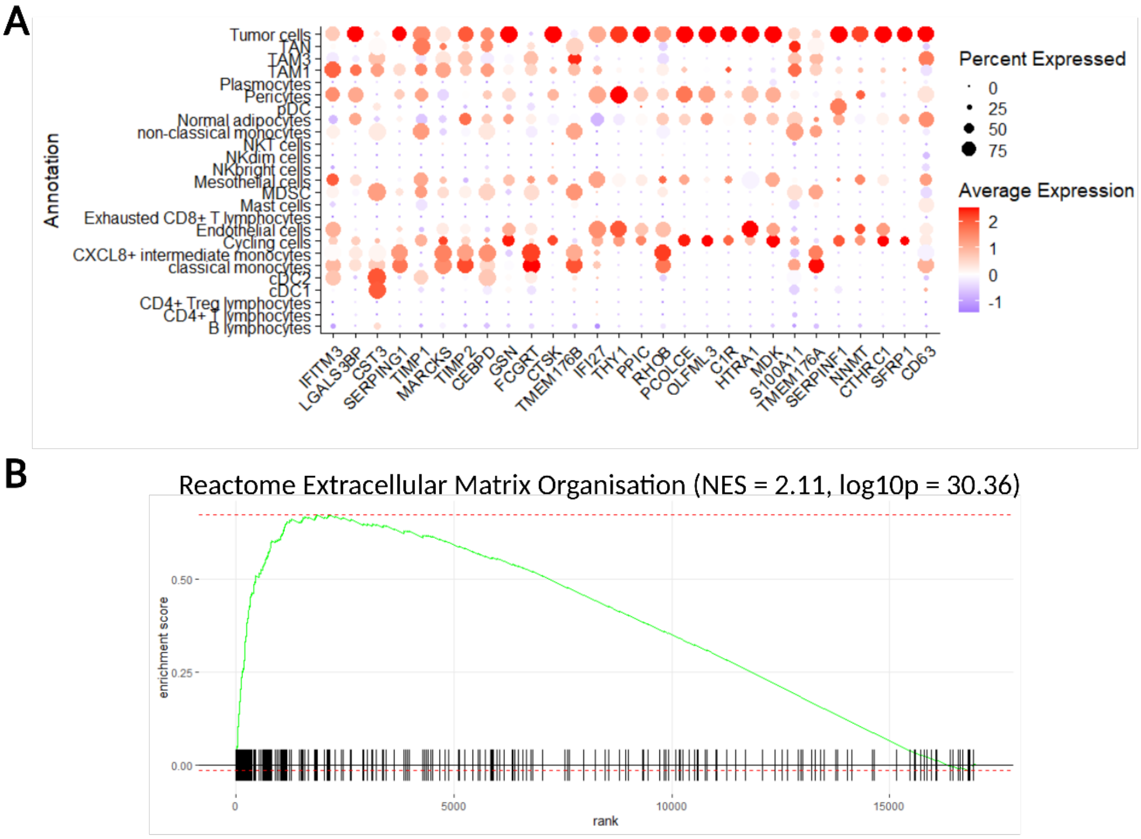


Figure 5.18: A: Single-cell COTAN gene co-expression analysis showing the top 30 genes co-expressed with IFI6 across cell types. **B**: Pre-ranked (by co-expression value) GSEA Enrichment plot for Reactome Extracellular Matrix Organisation. NES: Normalised Enrichment Score. The p value is a BH adjusted p-value.

The source of fibrotic/ECMO gene expression appears to be primarily the tumour itself, indicated in the GCN via modules M121, M107, and M236, which show the highest overrepresentation of genes in inferred single-cell tumour cell clusters (***Figure 5.19A***) and the expression of ECMO genes as evidenced by their enrichments (***Appendix A.7***). Furthermore, tumours express multiple fibrotic/ECM genes (***Figure 5.20A-B***), such as *FBN1, LUM, FAP, VIM, COL1A1, COL1A2,* and *COL3A1*. This suggests that tumour cells treated with interferons are contributing to tumour fibrosis, indicating a strong signature for fibrosis within the TME, in line with our cytological understanding of DDLPS and other dedifferentiated/undifferentiated mesenchymal tumours.

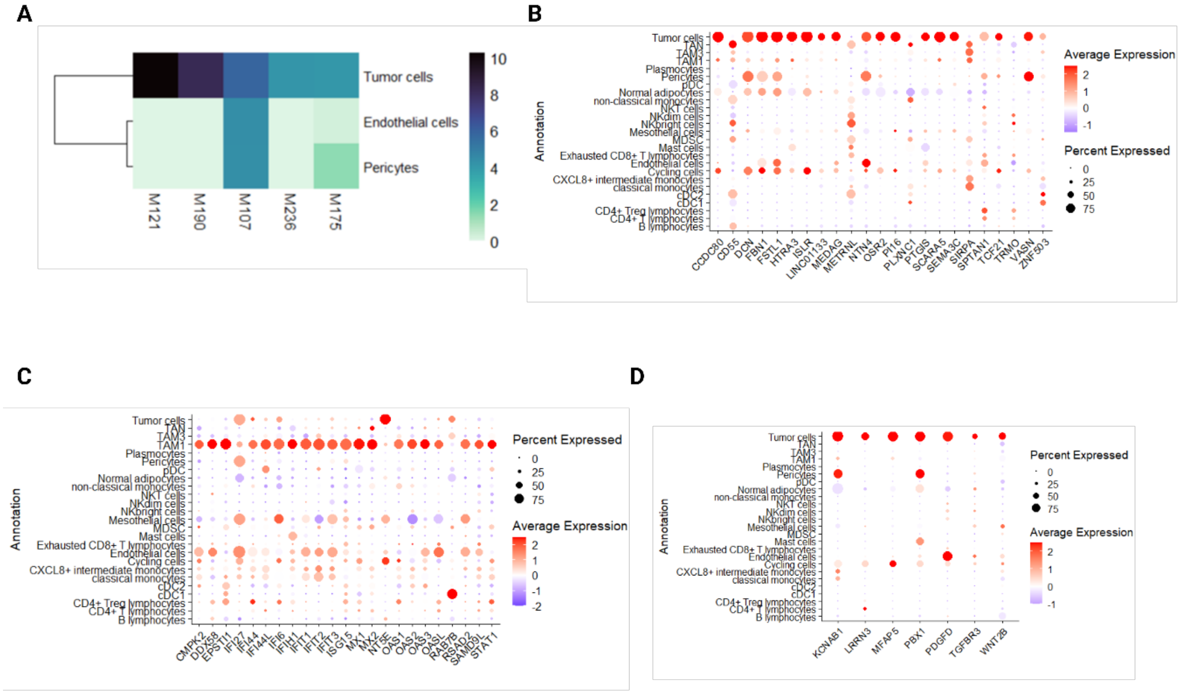


Figure 5.19: **A** Modules with the highest tumour signatures in single cell data where colour indicates the -log10(pvalue) from a Fishers exact test. **B** M121 gene overlap **C** M107 gene overlap and **D** M236 gene overlap with single cell data. Colour indicates the average expression and size of the dot indicates the percentage of cells expressed.

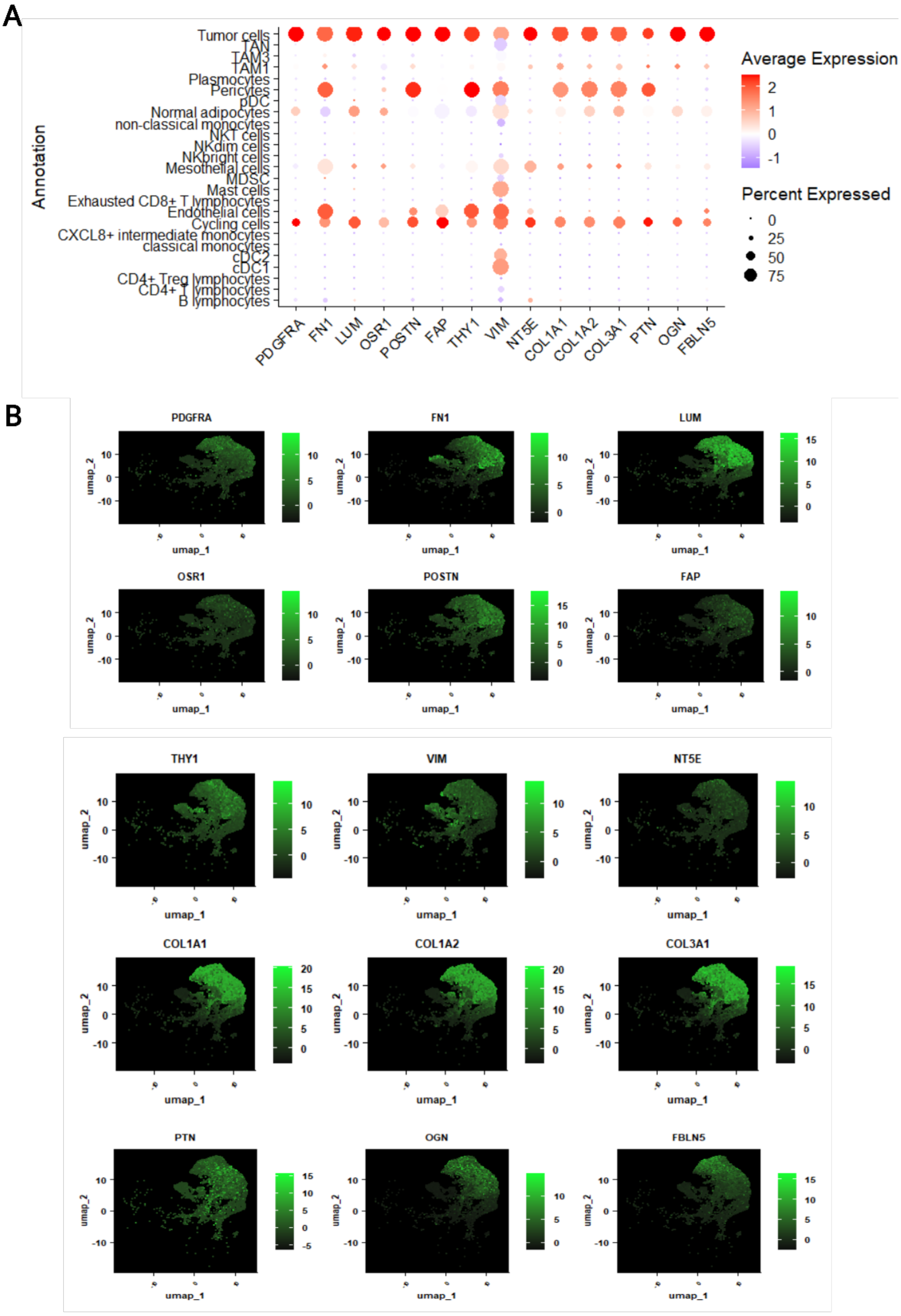


Figure 5.20: **A** Gene expression of embryonic FAP markers in DDLPS single-cell data. Colour indicates the average expression; size of the dots indicates the percentage of cells expressing. **B** UMAP (DDLPS cells) representation of FAP genes. Colour indicates average expression.

There is no strong indication of an association between hypoxia and angiogenesis, as inferred from the negative correlations among M236 (*rho* = -0.66) and M107 (*rho* = -0.44) with *HIF1A* M127 modules, suggesting other routes for angiogenesis. There was a moderate correlation between M107 and M197 (IL6) module (*rho = 0.34*), but no significant correlation was found between M236 and M197. Furthermore, in the TCGA DDLPS cohort, there does not seem to be a significant correlation between the gene expression of CD248 (M62) with *HIF1A* (M137) (*rho* = -0.11, *p* = 0.5054) or *HIF3A* (M107) (*rho* = 0.14, *p* = 0.4113), *PECAM1* (M186) (*rho* = -0.06, *p* = 0.7106), and *CD34* (M4) (*rho* = 0.27, *p* = 0.1181). However, there was a moderate and significant correlation with *CD105* (*END* - M183) (*rho* = 0.42, *p* = 0.0123) and *VEGFA* (M97) (*rho* = 0.38, *p* = 0.0221).

There is no indication that angiogenesis in DDLPS is linked to hypoxia as M137 (containing *HIF1A*) is negatively correlated with angiogenic modules M97 (containing *VEGFA*) (*rho* = -0.64, *p* < 0.001) and M183 (containing *CD105 – END*) (*rho* = -0.41, *p* = 0.014). There is a positive correlation between angiogenic modules and M107 (containing *HIF3A*); for example, M107 shows a strong positive correlation (*rho = 0.70*, *p* < 0.001) with M4 (containing *CD34*-). Conversely there is a negative correlation between M137 and M107 (*rho* = *-0.44, p* < 0.008).

5.4.3.4 Epigenetic modulation corresponds to differential inflammation response across observed the GCN

Gene expression markers suggest a tumour endothelial cell (TEC) phenotype within the TME (***Figure 5.7***). ECs are known to exhibit heterogenous epigenetic profiles that are aberrant in TECs.[436] Epigenetic modulation upon induction of inflammatory immune signals has been reported in the literature. Hence, methylation data from TCGA DDLPS was leveraged.[64] Methylation data was sorted by the top variable methylation probes across samples. These probes were correlated to module eigengenes to identify those associated with modules containing IL/IFN response genes. The NES from respective Reactome gene sets (Reactome – signalling by interleukins and Reactome interferon signalling) was overlayed (***Figure 5.21A***). A clear pattern emerges for the correlation of ME expression with hypo or hypermethylation that corresponds to the IFN and IL response gene enrichments.

There were 886 *PTPRN2* annotated methylation probes available in the TCGA SARC methylation data. Of these probes, 485 (54.8%) were found to possess a negative Spearman correlation coefficient of ≥ 0.2 with the M173 ME expression. Only 15 (2%) probes were found to have the inverse relationship. Assessing the methylation of *PTPRN2* and its gene expression reveals an inverse methylation-to-expression relationship where methylation probes (***Figure***

5.21B) show increased gene expression for increased methylation proportion correlating to lower M173 ME expression values (***Figure 5.21C***). IFN/IL-induced signalling is associated with lower gene expression and hypomethylation for *PTPRN2* and the M173 ME. *PTPRN2* was partitioned into M224 (***Figure 5.22A***) which was found to be enriched (***Figure 5.22B***) for Eukaryotic Translation Elongation according to pre-ranked GSEA (NES = 2.78, -log10padj = 11.89). Genes other than *PTPRN2* in M224 include aquaporins and their regulators in blood vessel (*AQP1* and *KLF2*), endothelial nitric oxide synthase (eNOS) trafficking (*NOSTRIN*), and lipid transport (*FABP4*). *KLF2* is present within this module, is enriched in endothelial cells, and regulates endothelial genes known as key regulators of vascular function.[437]

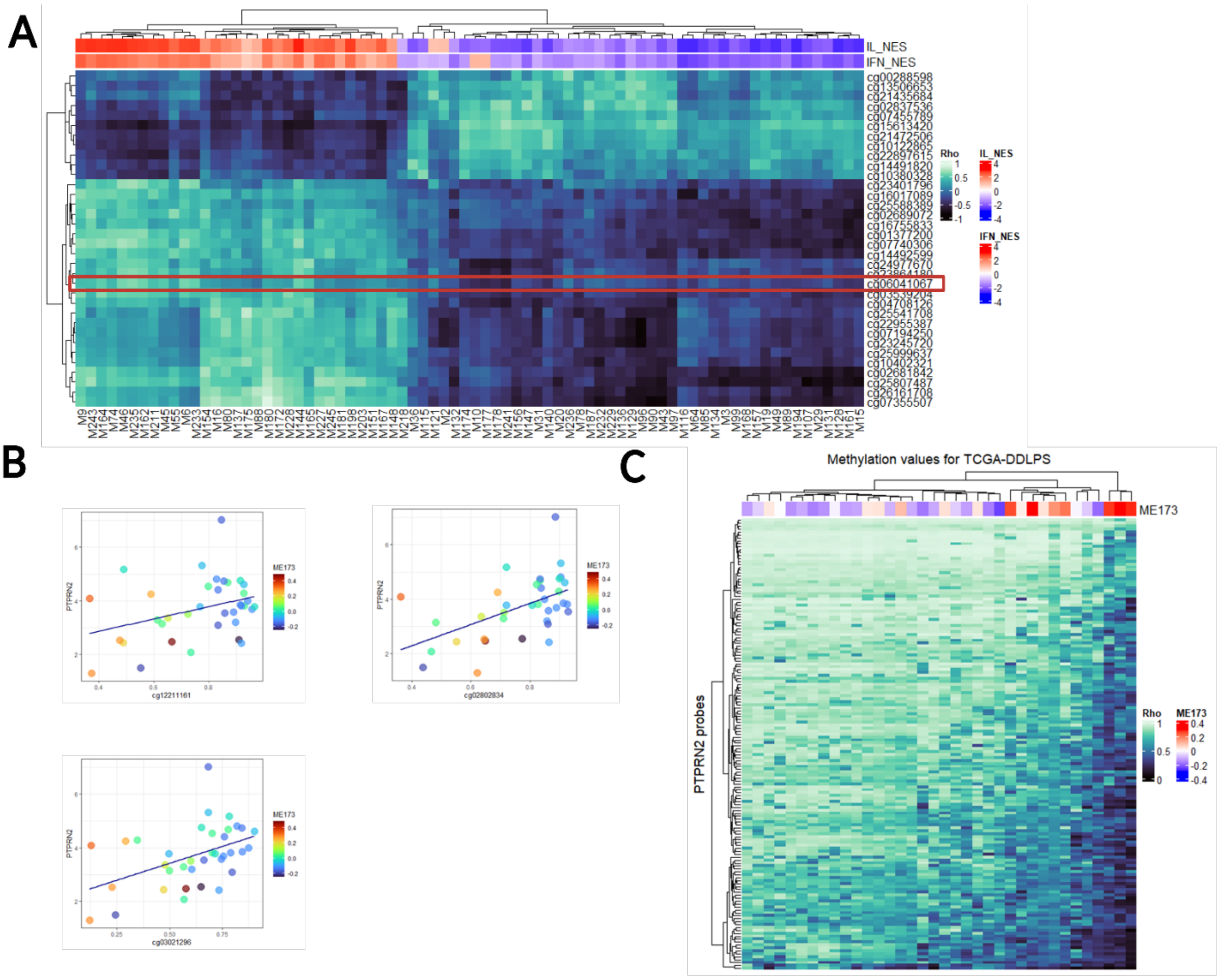


Figure 5.21: A Correlation of top variable methylation probes to module eigengenes. Column annotations are the results from a Gene set enrichment analysis conducted on the signalling with interleukin and Interferon signalling Reactome gene sets. Red box indicates the PTPRN2 probe. **B** PTPRN2 gene expression in TCGA DDLPS data against annotated probes coloured according to the ME M173 expression. **C** All PTPRN2 probes passing a mapping quality of >60 annotated according to the ME M173 expression.

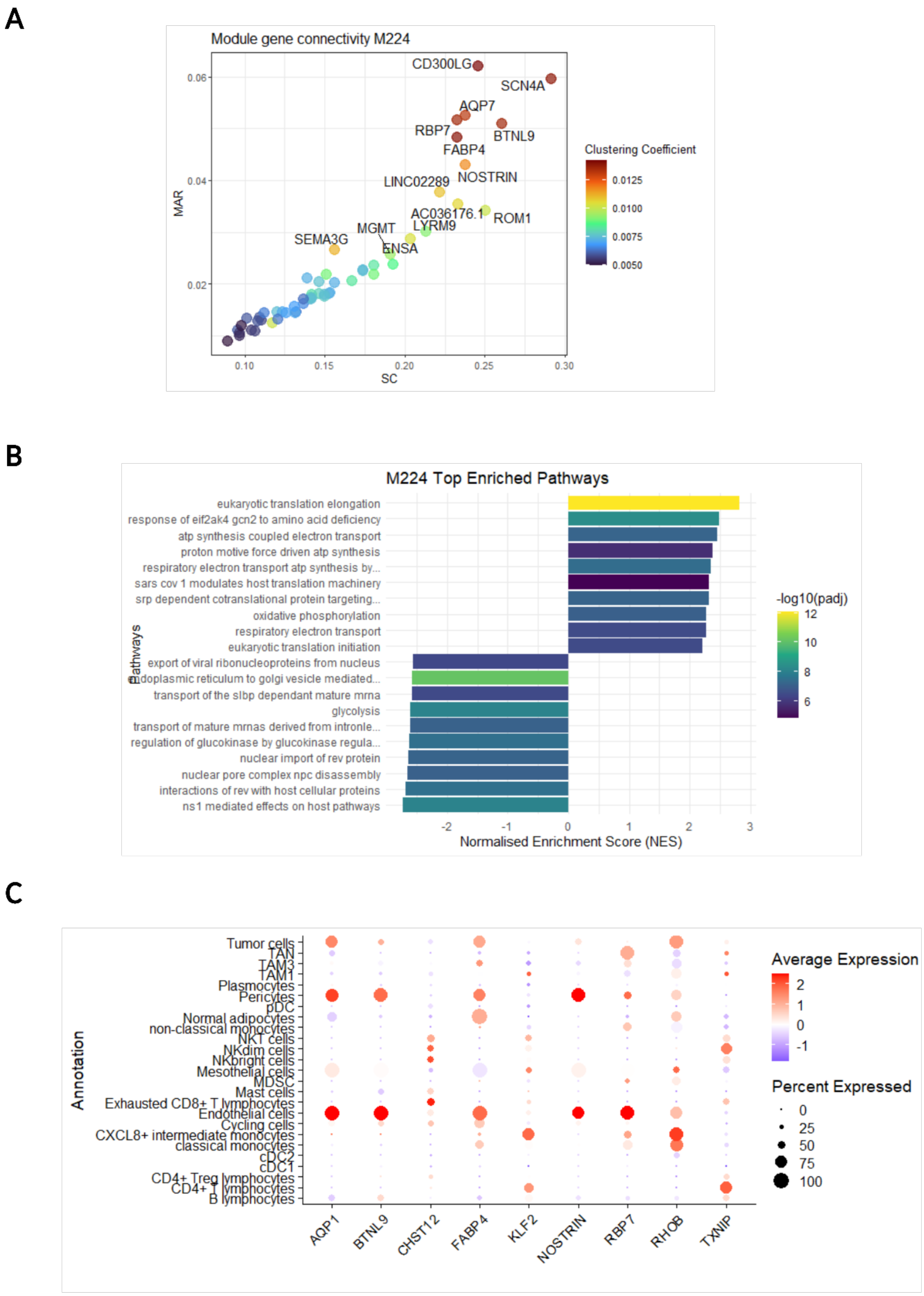


Figure 5.22: M224 module analysis. **A** Module connectivity graph showing the SC – scaled connectivity, and MAR – Maximum adjacency ratio, coloured by the clustering coefficient. **B** Reactome and GOBP enrichment analysis results passing a -log10 adj p > 1.3, displayed are the top ten positive and negative results ordered by NES – normalised enrichment score, coloured by -log10 adj p value. **C** Intersect of M224 genes with single cell RNA-seq cell clusters, labelled according to "global" annotations. Figure created in BioRender.com.

The top three *PTPRN2* methylation probes (cg12211161, cg09450352, and cg00596819) were found to correspond to intronic regions where there are annotated CpG islands, high GC content, and gene enhancer regions (***Figure 5.23***). This corresponds to the observed increase in *PTPRN2* gene expression with hypermethylation at enhancer sites.

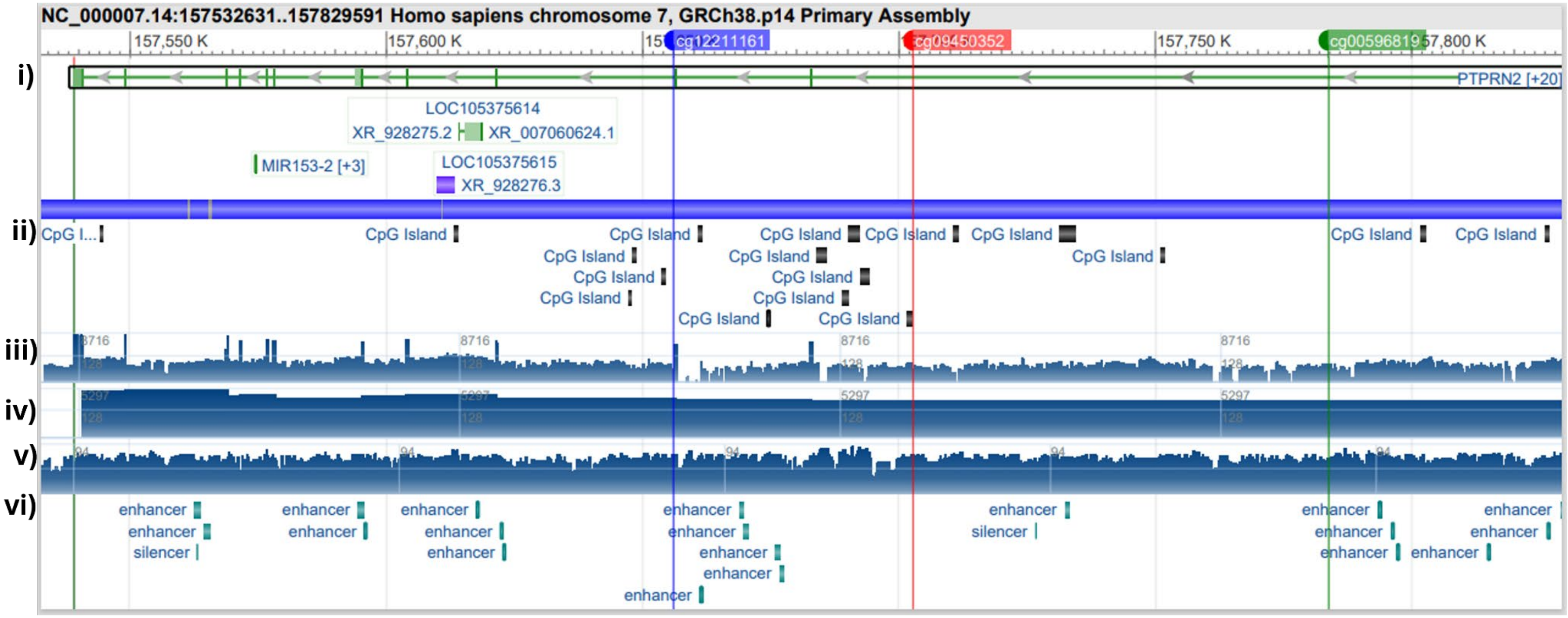


Figure 5.23: National Center for Biotechnology Information (NCBI) Genome Data Viewer (GDV) online web tool available at (https://www.ncbi.nlm.nih.gov/gdv/browser/genome/). The top three negatively correlated methylation probes (cg12211161 – blue, cg09450352, green – cg00596819). The following tracks were loaded **i)** NCBI RefSeq Annotation GCF_000001405.40-RS_2024_08, **ii)** The CpG Islands track, **iii)** RNA exon and **iv**) RNA intron spanning reads aggregate (filtered) Release 110, **v)** The GC content, and **vi)** The biological features (enhancer, silencer, promoter regions, etc)

5.5 Discussion

In this chapter, a range of enriched biological processes and molecular pathways in the GCN modules were identified and GS measures were used to rank modules to identify modules of interest with low correlation. In these shortlisted modules a heterogenous interferon signature was notable and correlated with tumour fibrosis (ECM remodeling) and lipid metabolism. IFN response genes corresponded to differential methylation across the GCN altering epigenetic programmes in DDLPS samples – notably *PTPRN2*. Such a signature may correspond to vascular heterogeneity. Furthermore, it was observed that tumour cells and vasculature cells show a striking similarity in their gene expression profiles which highlight tumour-associated vasculature processes.

Enriched processes/pathways include cell cycle, immune, metabolic, and development processes, describing well-documented pan-cancer hallmarks.[429,438,439] There was a notable overlap between EGN communities, GSEA enrichments and single-cell cell marker enrichment. This suggests that the GCN has successfully recapitulated biological processes and MEs display variation among gene co-expression that are likely meaningful. However, reproducing these results in the NCC dataset was not successful, this could be due to differences in gene expression profiles, mRNA enrichment methodologies, and filtering strategies.[114]

M173 and M62 displayed heterogenous IFN α-β and γ responses, which are crucial in anti-tumour immunity inhibiting tumour growth, angiogenesis, and recruiting pro-inflammatory immune cells.[440] It was noted that M173 also contained *STAT1* as a well-connected gene where it is typically regarded as a pro-inflammatory transcription factor in orchestrating downstream IFN (and immune) signaling.[441] The high expression of MHC class I molecules in M57 suggests that antigen presentation is active within the DDLPS TME, corresponding to the comparatively higher immune cell infiltration observed in DDLPS versus other LPS/sarcomas.[37,381,442] This notion is further supported with scRNA-seq results inferring multiple immune cell types in the TME.[143] However, chronic exposure to interferons can promote a tumour permissive environment through immune suppression and escape mechanisms.[443,444]

TGFβ signaling has been reported to be a key factor in promoting the DDLPS phenotype by blocking adipogenic differentiation in stem cells.[143] TGFβ is also known to be a key factor in tissue fibrosis, perhaps most notably through promoting ECM deposition, primarily from activated fibroblasts.[445] In this analysis tumour cells expressing high levels of IFN response genes most notably *IFI6* also highly expressed ECM genes which were also co-expressed. *CD248* is a notable example of one such gene, known to be upregulated in inflammation and fibrosis and enhances TGFβ activity.[446-448] *CD248* encodes Endosialin, a transmembrane glycoprotein. Endosialin was found to be lowly expressed in cells of mesenchymal origin and upregulated in fibroblasts, pericytes of tumour-associated vasculature, activated macrophages, and a range of tumours.[449,450] Upregulation of *CD248* enhances ECM deposition of fibroblasts through attenuating TGFβ and *CD248* knockdowns in human adipocytes corresponded to reduced tissue fibrosis and inflammation.[448,450] Inactivation of *cd248* in mice rescued pro-inflammatory chemokines, atherosclerosis and produced a more contractile environment blood vessels.[446] Increased *CD248* expression was found in adipose tissue compared to other tissues and correlated with obesity, inflammatory pathway annotations, ECM remodeling, and angiogenesis, potentially through augmentation of the hypoxic response (*HIF-1a* and *HIF-2*).[450,451]

In this study, there was no direct correlation between *CD248* and hypoxia-inducible genes *HIF1A, HIF3A*, neoangiogenic/endothelial cell markers *PECAM-1* (*CD31*), *CD34*, or *VEGFA*, but there was a weak yet significant association with *CD105* (*ENG*). This suggests that CD248 is not acting through hypoxia-induced genes. Therefore, another source is likely, which may be chronic IFN signals, growth factor stimulation, or mechanical shear stress due to the dense fibrotic TME.[452]

Endosialin has been shown to interact with ECM fibronectin and collagens, and be highly expressed in metastatic lesions (including STS).[453,454] Endosialin expression could distinguish undifferentiated and poorly differentiated mesenchymal tumours from other undifferentiated malignancies.[454] A phase 1 trial for ontuxizumab (MORAb-004), an anti-Endosialin monoclonal antibody, reported some stable disease responses in select sarcomas (myxoid chordrosarcoma, chondrosarcoma, undifferentiated pleomorphic sarcoma, rhabdomyosarcoma, and a unspecified sarcoma occurring in the uterus).[455] However, a follow-up phase II trial showed no evidence of increased response or PFS over placebo plus gemcitabine and docetaxel (including unspecified LPS).[456]

The fibrotic nature aligns with the cytological understanding of DDLPS, which shows a dense matrix and tightly packed stem-cell-like cellular structures throughout.[44] Gene expression signatures for tumour-associated vasculature showed that pericytes had a highly similar gene expression profile to DDLPS tumour cells. Pericytic vascular mimicry has been reported in WDLPS.[430]

Tumour angiogenesis is typically associated with hypoxia, either from increased metabolic demand of the tumour or TME density in fibrotic/dense tumours.[457] In the GCN, there was no correlation between hypoxia-inducible factors (*HIF1A*) and angiogenesis or activated endothelial cells. Instead, angiogenesis appeared to be associated with lipid homeostasis and directly with the gene expression of the tumour itself. Normally healthy ECs undergo a state of quiescence that is actively maintained until pericytes can cover the vessel, maturating it. These ECs must then be activated for angiogenesis. TECs, on the other hand, with their abnormal phenotype, are always activated and exist in a state of chronic inflammation, interacting with tumour and pro-inflammatory cells via adhesion molecules (ICAM-1, VCAM-1, and E-selectin).[458] Notably, the phenotype of endothelial cells within the DDLPS TME appears to be tumour-associated (TECs) according to the markers expressed (e.g., CD105 – ENG, Von Willebrand factor – VWF, and stem-cell marker CD34).[459]

In this analysis, *Protein Tyrosine phosphatase receptor type 2 (PTPRN2*), a gene within M224, was found to be differentially methylated, with hypomethylation patterns corresponding to increased expression of ME of M173, where IFN/IL6 response signatures were strongest. DNA methylation probes corresponding to enhancer regions suggest that these intronic methylation sites correspond to increased expression. *PTPRN2* is a phosphatase, membrane-bound on insulin-containing dense-core vesicles important in insulin secretion in the presence of glucose stimuli.[460] It is generally expressed in the nervous system and pancreatic endocrine cells and is overexpressed and differentially methylated in various tumour types.[461-466] PTPRN2 lacks classical PTP activity but possesses PIP activity. In colon cancer, it has been found to predict poor prognosis, and knockdowns (via shRNA) inhibited cell invasion, migration, and colony formation.[461]

A study in breast cancer found that TCGA data showed overexpressed levels of PTPRN2 with no obvious genomic alterations, suggesting epigenetic modulation.[462] PTPRN2 was found to predict poor prognosis. The immature mRNA of *PTPRN2, proPTPRN2*, was found to be exclusively expressed in breast cancer cells. *ProPTPRN2* expression corresponded to increased tumour size in mouse xenograft models. The isoform of PTPRN2 forms a complex with TRAF2,a RING domain E3 ubiquitin ligase, suppressing apoptosis.[462,465] A study in colorectal cancer (pre-print) produced similar results, showing that in addition to HOXD13 promoter binding, nuclear S100A10 binds to the 3' UTR region of PTPRN2, increasing its expression.[465]

A genome-wide methylation study in select soft-tissue sarcomas revealed that across leiomyosarcoma samples, *PTPRN2* was the most variable gene.[467] PTPRN2 was shown to modulate actin dynamics along with increasing migration and metastasis.[463] It has also been found to predict poor survival in patients treated with immune checkpoint blockade (anti-CTLA4 and anti-PDL1) in prostate cancer.[464] *PTPRN2* was highly upregulated in colorectal cancer cell lines after 5FU and oxaliplatin treatment, being associated with clinical stage and poor prognosis.[465] SiRNA knockouts indicated that MSC markers decreased with *PTPRN2* silencing. It has been considered a marker for EMT in metastatic prostate cancer.[468] Silencing impeded G1/S phase transition through correlation between *PTPRN2* and *Cyclin D1*. Furthermore, phosphorylated STAT3 and cyclin D1 decreased in silenced *PTPRN2*. *PTPRN2* silencing may cause cell cycle arrest through inactivation of the STAT3/cyclin D1 pathway and PTPRN2/cyclin D1 interaction. PTPRN2 promotes activation of the TRAF2/MAPK/MMP7 pathways, potentially through promotion of Lysine-63 linked ubiquitination of TRAF2, which is required for the recruitment of TAK1, phosphorylating MAPKs and the IKK complex to initiate MAPK cascades, possibly potentiating the cell cycle.

A study found that PTPRN2 was differentially methylated in correlation with module gene expression (M173) associated with IFN and IL signaling (ref 537). In human umbilical vein endothelial cells (HUVECS), IL-6 signaling induces insulin resistance in ECs, decreasing the activation of Akt/eNOS and stabilizing STAT3 phosphorylation, thereby reducing angiogenesis via decreased activity of DNA methyltransferases leading to hypomethylation.[469] The same study found that PTPRN2 showed changes in methylation patterns upon IL6 treatment, suggesting that PTPRN2 may play a role in promoting angiogenesis and nutrient delivery. Given this evidence, PTPRN2 has the potential to be a putative target in DDLPS, perhaps to potentiate current immune checkpoint inhibitors.

There is a strong signature for lipid metabolic pathways in the DDLPS TME which likely corresponds to molecular signatures within the WD-like/adipocytic portions of the tumour; the composite ranking metric included genes differentially expressed in WDLPS versus DDLPS, that is lowest ranking modules contain genes significantly overexpressed in WD portions.[333] Lipid metabolism is multi-faceted, complex, and extensively rewired in cancer, involving a range of synthesized or imported lipids to meet the high energy/nutrient demand for proliferation.[470,471] Fatty acids (FAs) can be leveraged to produce acetyl-CoA via β-oxidation for subsequent ATP generation or used to derive anabolic metabolites (e.g., phospholipids and triglycerides).[472,473] In non-adipogenic tissues, FAs are largely extracted from the microenvironment via transporters such as CD36 (present in M97) this is instead of producing FAs through *de novo* FA biosynthesis.[474] In cancer, in addition to an observed increase in FA (lipid) uptake through upregulation of transporters (e.g., *CD36* is highly expressed in multiple cancer types and has been correlated to EMT), they also conduct *de novo* biosynthesis of FAs.[471,475] FAs are synthesized from Acetyl-CoA (from citrate generated from glycolysis-TCA using glucose) and malonyl-CoA by the Fatty acid synthase (FASN) enzyme.[472] Through a series of biochemical reactions, FAs can then be converted to a range of molecules, including phospholipids, signaling lipids, glycerol, diglycerides or stored as triglycerides.[473]

Cholesterol is a key constituent of cellular membranes and involved in the synthesis of sex and steroid hormones, among other roles.[476] Like FA and other lipids, cholesterol can be acquired through uptake or *de novo* synthesis via the oxidative mevalonate pathway in the endoplasmic reticulum using acetyl-CoA.[477] FAs are not directly required for cholesterol biosynthesis, but can be used to derive acetyl-CoA for cholesterol. Cholesterol is taken into the cell by various routes, primarily through the low-density lipoprotein (LDL) to LDL receptor (LDLR) interfaces at the cell surface.[470,471]

Cholesterol biosynthesis genes have been shown to be expressed and retrieved in pathway analysis in CSCs for several cancer types.[478,479] CSCs display metabolic heterogeneity and can prefer various sources of metabolites and energy, including glycolysis, OXPHOS, and lipid metabolism.[480] With increased ECM remodeling and inflammation, the energy demand on cells is high, and therefore cells switch to glycolytic processes and lipid metabolism to meet energy demand.[481] Dysregulated lipid metabolism can promote the EMT to a fibroblastic phenotype and the release of profibrotic factors (e.g., TGFβ, vimentin, collagen I and III, and CD248) to activate myofibroblasts in fibrotic disease and cancer.[482-484] Fibrosis is also associated with the depletion of lipids in cells and accumulation of fatty acids (e.g., palmitic acid in pulmonary fibrosis) and glycerol from lipolysis of triacylglycerides.[483,484] Furthermore, fibrosis has been associated with an increase in extracellular cholesterol levels along with metabolites.[485] Signatures for cholesterol export in the GCN were identified, with gene expression in vascular and tumour cells.

Increased cholesterol metabolism has been noted to play a role in sustaining cancer stemness in a range of cancers.[479,486] This is not only for meeting energy and nutrient requirements but also in determining the cell fate of cancer stem cells (CSCs). FA β-oxidation can be promoted by stem cell markers, including NANOG, facilitating a switch from OXPHOS to β-oxidation.[487] Furthermore, lipogenic genes, including *PPARG,* are shown to be correlated with *NANOG* expression. Mevalonate pathway cholesterol intermediates sustain pluripotency in colon CSCs spheroids by blocking the rate-limiting enzyme HMGCR.[488] The blockade of cholesterol synthesis enzymes has been shown to reduce the growth of CSCs.[486]

CETP (M62) is required for the transfer of cholesterol esters and triacylglycerides (TAG) between high-density lipoprotein (HDL) and V/LDL lipoproteins to maintain intracellular and extracellular cholesterol homeostasis.[489] M62 also contains the *NR1H3* gene encoding the Liver X receptor- α (LXR-α), a transcription factor for cholesterol efflux genes, including the ATP-binding cassette family of proteins (ABC). LXR-α is activated through high intracellular concentrations of cholesterol to reduce cytotoxicity due to lipid reactive oxygen species.[432] ABC proteins, present in several modules across the GCN, (e.g., ABCA1) are involved in the transport of substrates across the plasma membranes of cells.[471] These are often downregulated in cancer, as cancers upregulate uptake and biosynthesis pathways.[490] Upregulated LXR-α and ABCA1 inhibit cancer growth in murine prostate cancer xenografts.[491] *Cyclooxygenase 1* (*PTGS1*) encodes an enzyme required for prostaglandin biosynthesis from cholesterol and is also involved in cholesterol efflux, where inhibition of COX1 will decrease efflux and promote accumulation via downregulation of associated ABC protein genes.[492] Furthermore, M62 was correlated with M63, which contained *Acyl-CoA: cholesterol acyltransferase (ACAT1),* encoding a protein involved in cholesterol esterification which is subsequently stored in lipid droplets.[432] Esterification is a means by which free cholesterol can be safely stored and reduce cytotoxic effects.

To conclude, this chapter identified top-ranked modules that can now be screened for hub-genes that could be putative drug targets in the subsequent **chapter 6**. Bottom-ranked "favourable" modules retrieved inflammation, lipid metabolism, ECM and angiogenic signatures. It is hypothesized that a high content of cholesterol and lipid metabolites in the TME could be providing the DDLPS tumour with nutrients and energy, nurturing CSCs, contributing to fibrosis, and suppressing differentiation programs in DDLPS. Furthermore, interferon responses corresponded to differentially methylated genes, notably *PTPRN2*, which may correspond to vascular heterogeneity warranting further exploration.

Chapter 6 Identifying robust hub genes and candidate drug targets

6.1 Introduction

In **section 5.4.4,** the biological enrichments within the DDLPS GCN were assessed, and then modules were ranked according to the correlation of IMC to GS measures. Then subsequently the lowest correlating modules were explored. This chapter examines the top ranked modules whose IMC correlate with higher GS values. These modules are inferred to be more biologically significant to unfavourable disease characteristics, associating with poorer survival and DDLPS specific molecular programs that have high essentiality according to DepMAP.[234] Modules associated with these features may represent pathways that are tumour vulnerabilities. A core hypothesis of this project, as discussed in **sections 1.3.5**, **sections 4.1 and 4.5**, is that hub genes are crucial to their networks. Therefore, hub genes in modules inferred to have high biological relevance may represent crucial genes and hence may make attractive targets.

Common centrality indices are explored in **section 1.3.1**. In brief, there are multiple, and contextual, ways to identify hubs.[213,317] Eigencentrality evaluates a node's influence by considering the connectivity of its neighbours, unlike degree centrality which is a direct measure of a nodes connectivity.[493,494] Eigencentrality is a measure of the influence of a node in a network where a high eigencentrality indicates a node that is connected to hubs, assigning a higher score of importance according to how well-connected neighbouring nodes are. Where the degree can be likened to a person's popularity, eigencentrality measures the relative importance of genes it is connected to. For example, a gene co-expressed with many genes that show a low relative co-expression with other genes will have a low eigencentrality. Those that are co-expressed with genes that have many co-expression interactions will have a high eigencentrality.

Subnetwork analysis is the examination of smaller local structures of a larger network. This saves on computation time and it allows the inspection of local network structures with higher similarity and functional relevance.[234,236] WGCNA provides a platform for subnetwork analysis through the clustering of genes into modules.[234] Furthermore, this inspection is not restricted to using the module gene assignment alone, but also through the kME.[251]

Section 4.4.3 demonstrated that TCGA and NCC DDLPS GCNs had over 80% of modules significantly preserved, and incorporating this at the subnetwork level allows for interpretation of concordant co-expression edges. The Random walk with restart on multiplex heterogenous (RandomWalkRestartMH) R package incorporates a random walk with restart (RWR) algorithm for multiplexing based on local structures.[282] In this algorithm a random walker starts at a seed node and moves to neighbouring nodes within and between layers, either restarting at the seed node with a set probability or moving to another neighbour. This process is repeated for multiple iterations, and nodes are scored based on the frequency of visits, indicating proximity within and between network layers.

The result is a probability distribution that equates to a measure of node similarity. This benefits subnetwork analysis as the adjacencies, and subsequently the TOMs between GCNs are normalised.[282] RWR requires the selection of a seed node, which intuitively would be a gene known to be significant in disease (e.g., *MDM2* and/or *CDK4* in DDLPS) or a hub-gene in one of the layers, although this can be difficult given the context of networks being multiplexed. Another limitation of this approach is that it explores local structures (199 nodes plus the seeds) in the network. Although, such limitations are not detrimental to exploring local structures.

Hub genes can be identified from subnetworks and presented with more confidence due to known preservation of co-expression patterns. After hub-genes are identified, drug target information can be leveraged to identify putative targets and inhibitors.[495,496] For this purpose there is a vast range of databases providing information on drugs, their targets and disease indications.[319,320] The Therapeutic Target Database (TTD) and ChemProbesPortal (CPP) provide information for targets, current status, indications, and cross-referencing to external databases for each drug.

For studies applying such a WGCNA-based approach to DDLPS, has identified modules associated with the DDLPS disease phenotype.[249,291] Then genes that show tight association with the modules identified were taken given a set threshold[291] and cross-referenced with other analyses including differential expression and/or Least Absolute Shrinkage and Selection Operator (LASSO) regression – screening out those that may be predictive of outcome and thus interesting from a clinical perspective.

In this chapter, the top ten ranked modules will be explored, assessing the enriched functions using gene annotations available from MsigDB. Then the GCNs from both the TCGA DDLPS and the NCC DDLPS GCN will be multiplexed, and graph-based analysis will be used to identify hub-genes. PPI data will be leveraged to identify pathway/interactor information for each hub. Next to identify drugs and their targets, drug target data will be leveraged from the TTD and CPP to identify and propose candidate targets.

6.2 Aims & Objectives

Hypothesis: Modules-of-interest describe crucial DDLPS biological processes, and their hubs are key molecules that represent putative drug targets.

Chapter Aims and Objectives:

Aim 1 – Evaluate enriched biological processed and pathways in top-ranked modules

Objective 1.1: Perform GSEA using MsigDB gene sets to identify enriched biological processes and pathways.

Objective 1.2: Confirm cell type associations using an independent single-cell RNA-sequencing dataset.

Objective 1.3: Evaluate the relationship between top modules and neighbouring "pathways" (modules).

Aim 2 – Identify hub-genes in multiplexed subnetworks

Objective 2.1: Use the Random Walk With Repeat algorithm to identify module subnetworks containing concordant gene co-expression patterns.

Objective 2.2: Use the degree and eigencentrality indices to Identify hub-genes within subnetworks.

Aim 3 – Identify hub interactors and candidate drug targets

Objective 3.1: Integrate PPI data from the STRING database to identify interactions of hub genes.

Objective 3.2: Retrieve and search drug-target interactions for the hub-genes and identify drug-targets.

6.3 Methods

6.3.1 Gene annotation signatures

Human gene annotations for GO biological process (GOBP) and REACTOME pathways were retrieved from MsigDB (available at: https://www.gsea-msigdb.org/gsea/msigdb/).[416] To infer the biological processes and molecular pathways in each module GSEA was conducted. To do this all genes within the TCGA DDLPS GCN were taken (16,032) and ranked according to their gene module membership (kME) for each modules and used an input to the 'fgsea' function from the fgsea R package[415] (version 1.31.0), using the GOBP and REACTOME gene sets. For the fgsea function the minimum size of included gene sets was set to 30, the maximum size was set to 1000, and the number of processes was set to 1, eps for p-value estimation was set to 0. Results from fgsea were filtered according to a -log10(BH-adjusted *p-value*) > 1.3 and the top ten positive and negatively enriched pathways were taken according to the normalised enrichment score (NES). The NES is an enrichment score that has been normalised to the mean enrichment of a random sampling equivalent to the input gene list size.

6.3.2 Eigengene network neighbourhood analysis

The EGN was constructed by calculating the adjacency using the correlation between MEs (1 + cor(ME)/2), the adjacency was converted to a data frame using the 'exportNetworkToVisANT' from the WGCNA package (version 1.72-5) with an adjacency threshold of 0.7 (only MEs with an adjacency of >0.7 were retained) and converted to an igraph object using the igraph package (version 2.0.3) using the 'graph_from_data_frame' function.[234,345] The igraph object was then used as input to the RandomWalkWithRestartMH R package (version 1.22.0) using functions 'create.multiplex', 'compute.adjacency.matrix' and 'normalize.multiplex.adjacency'. For neighbourhood analysis.[282] To identify module relationships among the top ranked modules, the RandomWalkWithRestartMH R package (version 1.22.0) was used to conduct a RWR on the EGN. The seeds were set according to the top five intramodular connectivity (IMC) vs GS ranked modules (M241, M10, M66, M35 and M100). The restart probability *r* was set to 0.1 to allow for network exploration where higher values of *r* reduce the number of nodes a random walk visits. For extracting results, *k = 5* was chosen extracting the top five neighbours in the 'create.multiplexNetwork.topResults' function. These top results were then converted to a tbl_graph object using the 'as_tbl_graph' function from the tidygraph R package (version 1.3.1).

6.3.3 Single cell data integration & analysis

10X Genomics single-cell RNA-sequencing data on 11 DDLPS patient samples was kindly provided by Sarah Watson at the institute-curie, including 28029 cells prior to filtering.[143] Data retrieved included cell type annotations as described in the published material and were used in this analysis. Single cell read data was processed and filtered as described in **section 2.16**. All single-cell data was analysed using the Seurat R package (version = 5.1.0) [348] Red blood cells were removed from the cell pool leaving 27446 cells. This scRNA-seq data is the first available for human patient DDLPS samples and is a novel opportunity to deconvolute the DDLPS GCN and identify cell-specific co-expression patterns. To integrate the single-cell and WGCNA results, the overlap between cell cluster genes and WGCNA module genes were identified. This was achieved using an overrepresentation test and the reported Fishers exact statistic via the GeneOverlap R package (version 1.38.0). If the intersection between gene sets was found to be less than two genes, the significance was manually zeroed to prevent significant results with low levels of overlap. A -log10(p-value) of 1.3 was considered significant. Results from this analysis were visualised using the pheatmap R package (version 1.01.2).

Gene expression for the given genes in the single-cell data was visualised using the 'DotPlot' and 'Featureplot' functions from Seurat, with the assay set to "SCT" for the featureplot function. To inspect the genes expression of genes overlapping between modules and single-cell data, the 'intersect' function from base R was used and passed to DotPlot or FeauturePlot for visualisation. For genes of interest, tumour cells were extracted from the single-cell data using the provided annotations, and cells were subset according to high or low gene expression values. High or low expression values were decided based on the distribution of gene expression values as inspected on a histogram. Differentially expressed genes (putative cluster markers) between these subsets were identified using the 'FindMarkers' function in Seurat, employing the default Wilcoxon rank sum test. Only those passing a threshold of 0.5 logFC were considered. Genes were further filtered according to an average logFC > 2 and a Bonferroni adjusted p-value of 0.01, ensuring that only genes highly expressed with high significance in a specific cluster were identified.

6.3.4 Identifying subnetworks and RWR-preprocessing

To identify subnetworks, the gene module membership (kME) was used as a method to filter genes correlating to a given module, satisfying a kME > 0.5. Typically, a kME > 0.7 is selected for identifying important genes. However, for subnetworks a kME >0.5 struck a balance between excluding non-correlating genes but expanding the network to include valuable information on an expansive GCN. Genes satisfying the threshold were then subset from the TOM of both the TCGA and NCC. Separate networks were then constructed and filtered using the igraph and tidygraph R packages (as described in **6.3.3**). To remove lowly co-expressed edges which are less robust (those with a low TOM) and filter the edges a TOM threshold was set. TOM threshold was set on a per subnetwork basis starting from TOM > 0.01 and then increasing the TOM to >0.1 if network connectivity allowed. For some subnetworks with low co-expression values (TOM) amongst genes, a high TOM would make a sparse or empty network that is non informative. TOM thresholds for the TCGA and NCC were kept the same. The processed TCGA and NCC subnetworks were then used in multiplexing. In some networks, where there was evidence of a strong overlap between gene kME values (e..g, M10 and M241) based on ranked gene lists returning the same genes, a decision was made to identify network genes based on the module gene assignments alone. This was done to prevent the identification of networks with no unique hubs.

6.3.5 Network multiplexing and integration

To investigate the co-expression network dynamics a multiplex network approach was used to combine the TOMs from the TCGA and NCC subnetworks. These networks were used as input into the RWR algorithm via the RandomWalkRestartMH[282] (*version* 1.22.0). The seeds of the analysis were set based on the most connected genes identified in the TCGA DDLPS subnetwork. The 'create.multiple' function was used to create a multiplex object for which the adjacency of layers were calculated using the 'compute.adjacency.matrix' function, which was then normalised using the 'normalize.multiplex.adjacency'. The random walk with restart was conducted using the 'Random.Walk.Restart.Multiplex' function using the seeds as the start nodes, and a restart probability "r" of 0.70, as is recommended by the authors of the RandomWalkRestartMH package. The *tau* setting, which controls the weighting of layers for the RWR algorithm, was kept as default (1:1) as there was no evidence to suggest substantial differences in GCN network densities. The results were retrieved using the 'create.multiplexNetwork.topResults' setting the number of top result 'k' as 199 (199 + the number of seeds in total).

The tidygraph R package (*version 1.3.1*) was then used to build a graph object and perform centrality indices calculation and network feature filtering. To identify concordant edges between the two layers (network types) of the stacked network, the number of loops (an edge being present in both TCGA and NCC) were calculated and those with ≤1 loops were removed (removing non-concordant co-expression edges). Then the degree and eigencentrality measures were calculated using igraph functions. Genes with only one connection were removed according to a degree centrality of >1, this was done to remove orphan nodes.

To visualise the network, in the first instance the ggraph R package (version 2.2.1) was used. This was useful for large networks as the computation time for ggraph is low and it contained multiple options for faceting by edge and node features (e.g., type of edge/node). Elsewhere the visNetwork (*version 2.1.2*) R package was used to generate multiplexed subnetworks.

6.3.6 Identifying hub genes

Hub genes were identified based on two centrality indices. The first is the degree centrality which describes the total number of connections a node has.[212,213,493] The second measure was the eigencentrality, which is the number of connections adjusted to the number of highly connected neighbours.[212,213] Nodes that are high in eigencentrality are often more robust and show a lower level of redundancy and are more likely to be essential in ordering of network topology. Therefore, targeting eigenhubs may disrupt networks more effectively than hub genes.

To identify hub genes the igraph R package was used (version 2.0.3) where the 'centrality-degree' function was used to calculate the degree and the 'centrality_eigen' function was used to calculate the eigencentrality, using the edge list as input.[345] These values were then ordered from largest to smallest to highlight the hub genes.

6.3.7 Retrieving and filtering STRING protein-protein interaction data

STRING is a curated database containing information on protein-protein interactions (PPIs) that is commonly used in cancer network studies seeking to identify candidate biomarkers/targets.[227] As is discussed in **section 1.** For each PPI STRING provides several types of evidence-based confidence scores derived from different sources: Text-mining, Experiments, Databases, Co-expression, Neighbourhood, Gene Fusion, and Co-occurrence. STRING combines these scores into an "overall confidence" score, for which the creators of STRING suggest using a minimum of value of 400 which is described as a "medium confidence level".

Protein-protein interaction data (9606.protein.links.full.v12.0.txt) along with associated annotations (9606.protein.info.v12.0.txt) were retrieved from the STRING website (available at : https://string-db.org/), using the version 12.0 (2024) release.[227] PPI data was imported into R studio and protein-protein interactions were filtered based on an experimental evidence score of ≥ 700, which left a total of 90673 protein interactions across 6270 proteins. This threshold was selected to prioritize high-confidence, experimentally validated interactions, minimizing the inclusion of false positives. Only experimental evidence was used for several reasons: It directly reflects experimental validation of molecular interactions, avoids redundancy with co-expression analysis already performed and focuses on identifying interactors of key hub genes.

6.3.8 Drug-target data retrieval and integration

Drug-target information was retrieved from both the Therapeutic Target Database (TTD; available at, https://idrblab.net) and the Chemical Probes Portal (CPP; available at, https://www.chemicalprobes.org/).[319,320] The following files were retrieved from the TDD: The 'P1-07-Drug-TargetMapping.csv' which mapped TTD drug IDs to target IDs, providing a mode of action and the highest drug status (e.g., clinical trial phase 1, pre-clinical etc) and the 'P1-01-TTD_target_download_edited.txt' containing, and the "ChemicalProbesPortal-26_07_2024.csv" file was retrieved from the CPP. Files were processed in R and were formatted into a data frame object using the tidyr (version 1.3.0) and dyplyr (version 1.1.4) R packages. Files were then left joined using the drug names provided by TTD and CPP. Drug-target interactions were then filtered to remove drugs that were discontinued, terminated, or withdrawn from market. The mode of action for the drug was filtered to be inhibitory against the target based on available data annotations from these sources.

6.3.9 TCGA SARC ploidy scores

To assess the relationship between gene expression and tumour ploidy scores, which is a measure of genomic instability, the metadata from TCGA SARC was retrieved as set out in **section 2.4.** TCGA SARC ploidy scores were derived in the analyses conducted by the cancer genome consortium sarcoma project[64] In their work ploidy scores were calculated using the ABSOLUTE algorithm.[497] The ploidy scores estimate the average DNA copy number per cancer cell.

Key genes were considered as those known to be biologically significant in DDLPS which include, MDM2, CDK4, FRS2, YEATS4 and CPM. Targets identified in the subsequent results sections will also be regarded as key genes.

Ploidy values were binned into categories reflecting increasing levels of aneuploidy and genome doubling. Samples were classified as near-diploid (ploidy ≤ 2.2), moderate aneuploid (ploidy > 2.2 and ≤ 2.7), near-triploid (ploidy > 2.7 and ≤ 3.2), near whole-genome duplication (near-WGD) (ploidy > 3.2 and ≤ 3.8), or whole-genome duplicated (WGD) (ploidy > 3.8).[498] Differences in gene expression across tumour ploidy classes were assessed using the Kruskal–Wallis test, a non-parametric method for comparing multiple groups. Statistical significance was adjusted for multiple testing where applicable using the Benjamini-Hochberg (BH) procedure.

6.3.10 CINSARC 67 gene signatures

The Complex Index in Sarcoma (CINSARC) is a prognostic gene expression signature of 67 genes developed specifically for soft tissue sarcomas. The CINSARC details chromosomal instability paired with dysregulated cell cycle functions in cancer including assembly of the spindle apparatus, chromosome segregation, DNA replication and repair.[499,500] The CINSARC has become a well-recognised signature in the STS disease space since its conception.[501] It is notable that high CINSARC 67 gene expression is typical in DDLPS and corresponds to poor outcome.[502]

Enrichment of the CINSARC 67 score was assessed for each module using two approaches. The first was to identify enrichment by a pre-ranked gene set enrichment analysis (GSEA) using the fgsea R package[415] (version 1.31.0) where genes were ranked according to their kME value, as is also described in **section 6.3.1**. The second method was to use the GeneOverlap R package (version = version 1.38.0) along with the 'newGeneOverlap' and 'testGeneOverlap' functions. These were used in a similar method as set out in **section 2.17**. However, gene sets tested were the module genes by partition, and the CINSARC 67 gene signature. A p-value of 0.05, or -log10 transformed value of 1.3 were deemed statistically significant overlaps.

CINSARC 67 genes were taken from the published article by Lesluyes et al.[500] These genes are; *ANLN, ASPM, AURKA, AURKB, BIRC5, BUB1, BUB1B, C13orf34, CCNA2, CCNB1, CCNB2, CDC2, CDC20, CDC45L, CDC6, CDC7, CDCA2, CDCA3, CDCA8, CENPA, CENPE, CENPL, CEP55, CHEK1, CKS2, ECT2, ESPL1, FBXO5, FOXM1, H2AFX, HP1BP3, KIAA1794, KIF1, KIF14, KIF15, KIF18A, KIF20A, KIF23, KIF2C, KIF4A, KIFC1, MAD2L1, MCM2, MCM7, MELK, NCAPH, NDE1, NEK2, NUF2, OIP5, UBE2C, PBK, PLK4, PRC1, PTTG1, RAD51AP1, RNASEH2A, RRM2, SGOL2, SMC2, SPAG5*, *SPBC25*, *TOP2A*, *TPX2*, *TRIP13*, *TTK*, *ZWINT*

6.4 Results

6.4.1 Top ranked modules enrichments

In **section 5.4.2** modules were ranked according to the correlation of IMC with GS. Here the results for the top ten modules associated with high IMC vs GS values are presented (***Table 6.1***). These top ten module are (in decreasing order): M241, M10, M66, M35, M100, M23, M178, M225, M94, and M89. All these top modules are preserved among DDLPS datasets according to the Zsummary score (>2), M241 and M10 are highly preserved (>10) where the top ranked module M241 showed the highest Zsummary (15.76).

Table 6.1: Module ranking summary for modules positively associated with GS measures.

Module	IMC	IMC vs Clinical GS	IMC vs Deg GS	IMC vs Dep GS	Rank	Preservation Zsummary
M241	0.37	0.269	0.649	0.441	1	**15.76**
M10	0.35	0.284	0.675	0.281	2	**14.46**
M66	0.47	0.318	0.255	0.367	3	2.02
M35	0.44	0.499	0.444	0.191	4	4.35
M100	0.58	0.175	0.409	0.251	5	7.96
M23	0.32	0.507	0.063	0.238	6	7.22
M178	0.46	0.144	0.433	0.241	7	6.35
M225	0.39	0.543	0.078	0.216	8	2.11
M94	0.53	0.394	0.021	0.369	9	5.99
M89	0.48	0.272	0.193	0.128	10	2.43

IMC: intramodular connectivity. **GS**: gene significance score. **Clinical GS:** GS derived from ICM correlated with the corDeviance as described in **DEG GS:** The Differentially expressed gene GS. **Dep GS** : The DepMap DDLPS cell lines dependency GS score. **Preservation Zsummary:** The Zsummary value from a module preservation analysis using the NCC dataset, a **bold** value indicates a highly significant preservation of Zsummary ≥10. **Rank:** The total rank of the GS measure summed together.

Enrichment analysis revealed a strong signature for cell cycle processes in these modules, with strong overlap between module gene membership and cell cycle functions (***Figure 6.1A***). Notably, M241 showed the most significant enrichment for the GOBP mitotic sister chromatid segregation, *NES* = 3.49, *-log10p* = 42.32) and showed strong enrichment for many cell cycle gene sets (***Figure 6.1B***). Six of the ten top ranked modules were significantly enriched for cell cycle related processes including, mitotic sister chromatid separation GOBP term (M241, M10, M100, M178), sister chromatid segregation (M35) and regulation of chromosome segregation (M66). One module (M94) was enriched for DNA replication. Overall, seven modules showed enrichment for cell cycle processes. This was also observed in the single-cell data, where five of the top ten modules showed cycling cell gene expression signatures (***Figure 6.1C***). Together with the module preservation Zsummary scores (***Table 6.2***), these results indicate that M241, M10, M100, M66, M35, M178, M94 represent preserved mitotic cell cycle programmes. Most notably M241 and M10, which showed high levels of preservation (***Table 6.1***). M241 (***Figure 6.2A***) had 57 GOBP cell cycle annotated genes out of the 127 within the module and M10 (***Figure 6.2B***) had similar observations, containing 54 genes with GOBP cell cycle annotations.

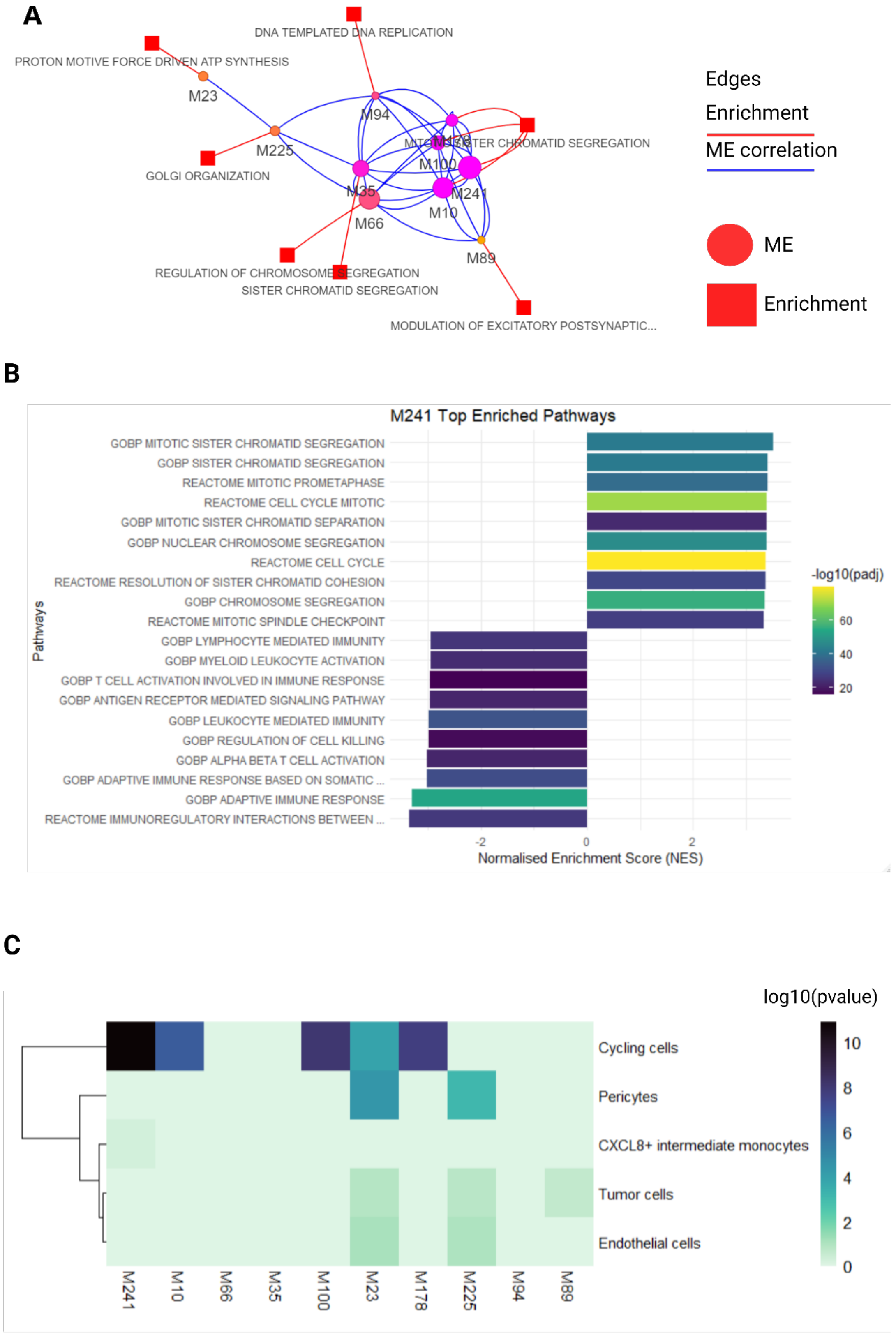


Figure 6.1: Module enrichment results. **A:** Enrichment-module network showing the module-eigengenes (MEs - circles) and their gene enrichment (both REACTOME and GOBP terms - squares). Blue edges indicate a correlation between MEs, and red edges indicate the top enrichment term for that module (as represented by its ME). **B:** M241 expanded enrichment analysis results detailing the ten top and bottom pathways by normalised enrichment score (NES) and passing a -log10(BH adjusted p-value) > 1.3 threshold. **C:** Heatmap detailing top overlap of module genes with top cell type marker genes via an overlap test (overrepresentation, Fishers exact).

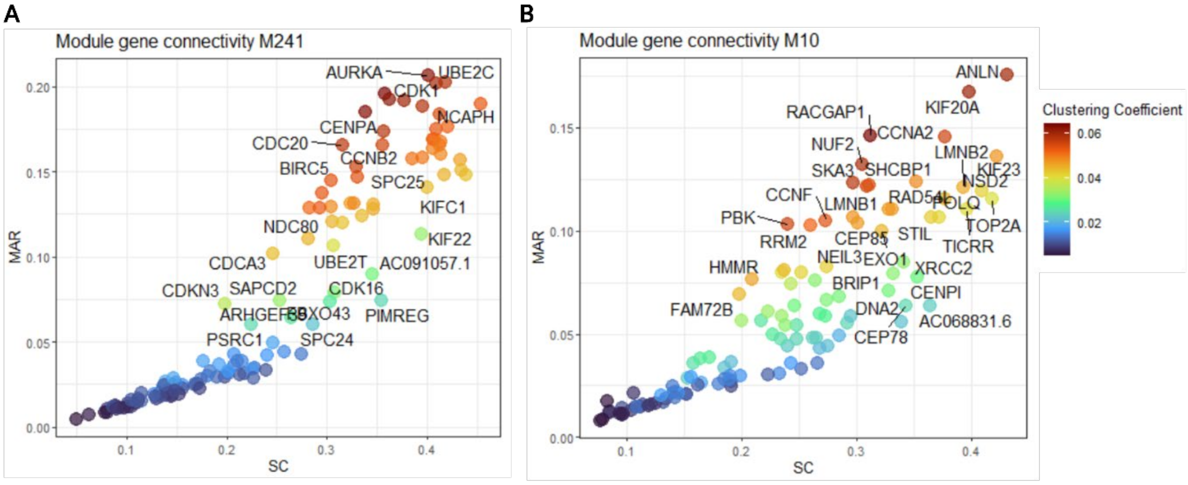


Figure 6.2: Module gene connectivity for **A** M241 and **B** M10. The SC – scaled connectivity measures how well-connected genes are in respect to others (gene connectivity scaled to the maximally connected gene in that module) and the MAR- maximum adjacency ratio, is the highest value of adjacency between two genes indicating a connection strength. Colour indicates the clustering coefficient, details how well genes cluster together.

In addition to the cell cycle modules (M241, M10, M66, M100, M178, M94, M35, M89) there were two modules (M225, and M23) that were not enriched for cell cycle functions. M225 (***Figure 6.3A***) showed a strong enrichment for Golgi organisation, and several enrichments pertaining to intracellular vesicular transport (***Figure 6.3B***). M23 (***Figure 6.4A***) showed an enrichment for protein synthesis (translation initiation and elongation), protein exportation and transportation (SRP-dependent cotranslational protein targeting to membrane), and amino acid stress response (***Figure 6.4B***). Notably, M23 contained many mitochondrial genes (***Figure 6.4A***). Further supporting them being distinctive modules with a moderate and significant negative Spearman correlation in the MEs of M241 with M225 (*rho* = -0.32, *p* = 0.06) and M23 (*rho* = 0.44, *p* = 0.007) with M241 (***Figure 6.5A***). M225 and M23 are uncorrelated (*rho* = 0.006, *p* = 0.972) (***Figure 6.5A***). The remaining modules were increasingly correlated with M241 and M10 (***Figure 6.5B***), M10 and M241 displayed a high and significant positive Spearman correlation (*rho* = 0.89, *p* < 2.2e-16).

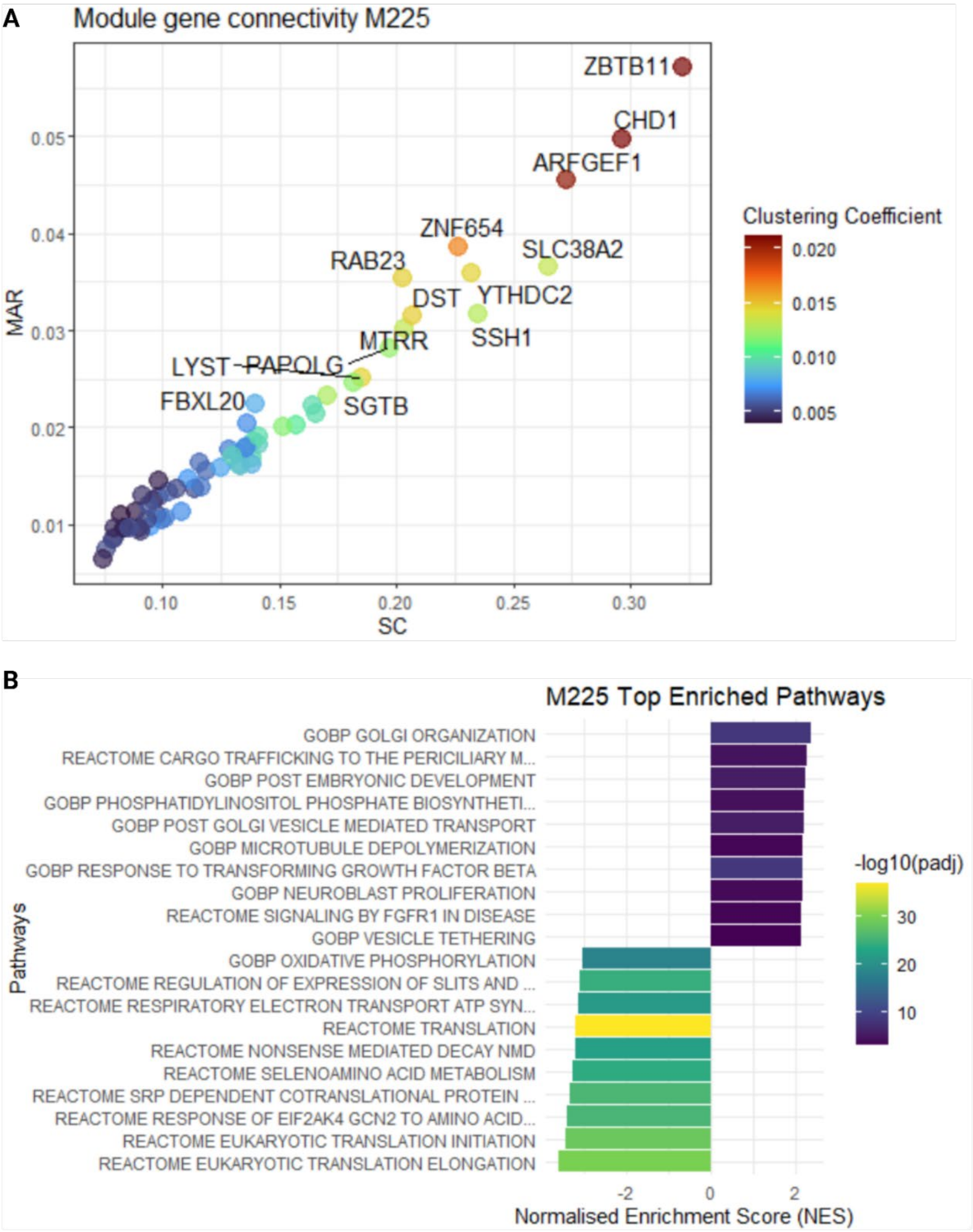


Figure 6.3 A: The module gene connectivity for M225. The SC – scaled connectivity measures how well-connected genes are in respect to others (gene connectivity scaled to the maximally connected gene in that module) and the MAR- maximum adjacency ratio, is the highest value of adjacency between two genes indicating a connection strength. Colour indicates the clustering coefficient, details how well genes cluster together. **B:** M225 enrichment analysis results detailing the ten top and bottom pathways by normalised enrichment score (NES) and passing a -log10(BH adjusted p-value) > 1.3 threshold

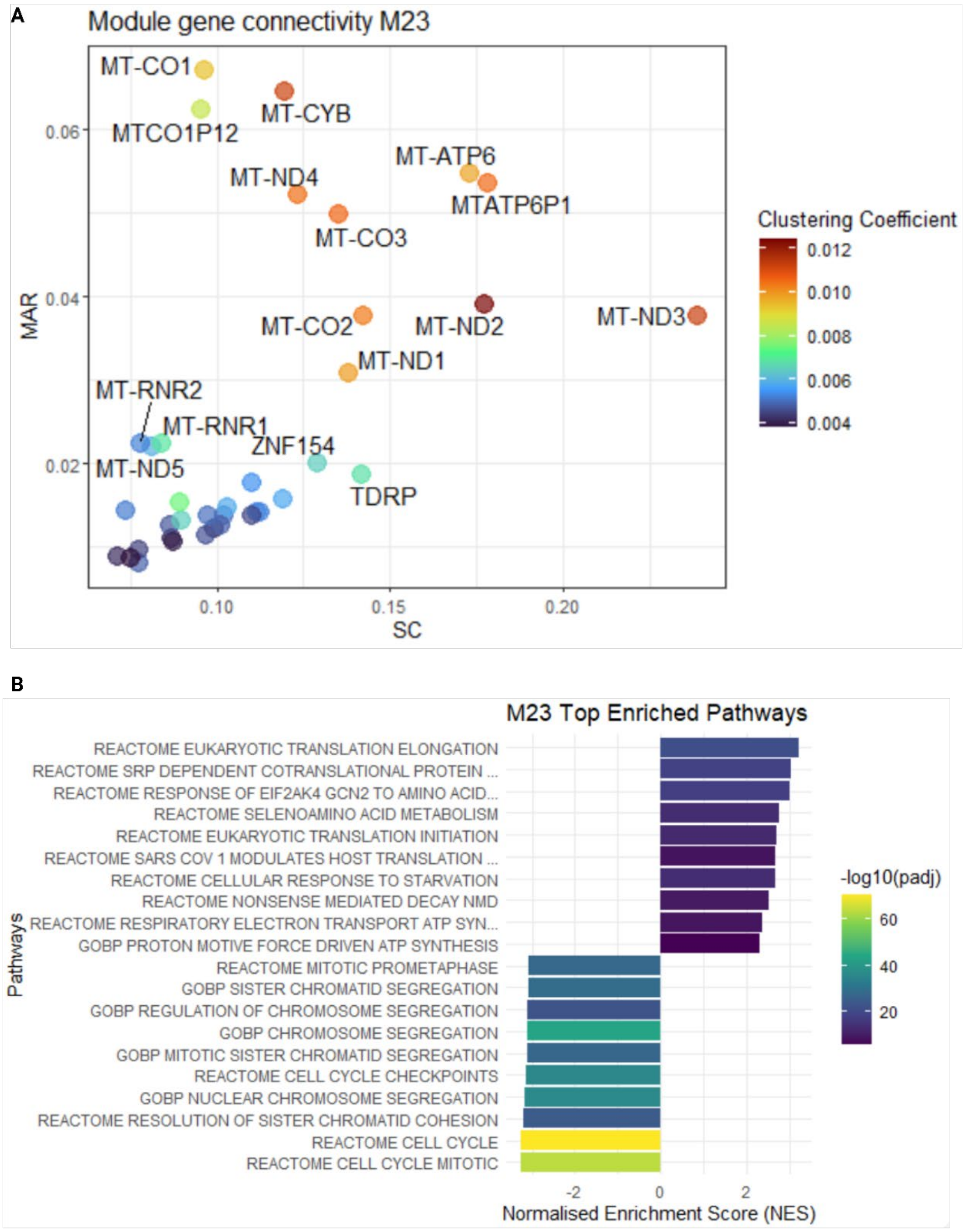


Figure 6.4: A: The module gene connectivity for M23. The SC – scaled connectivity measures how well-connected genes are in respect to others (gene connectivity scaled to the maximally connected gene in that module) and the MAR- maximum adjacency ratio, is the highest value of adjacency between two genes indicating a connection strength. Colour indicates the clustering coefficient, details how well genes cluster together. **B:** M23 enrichment analysis results detailing the ten top and bottom pathways by normalised enrichment score (NES) and passing a -log10(BH adjusted p-value) > 1.3 threshold

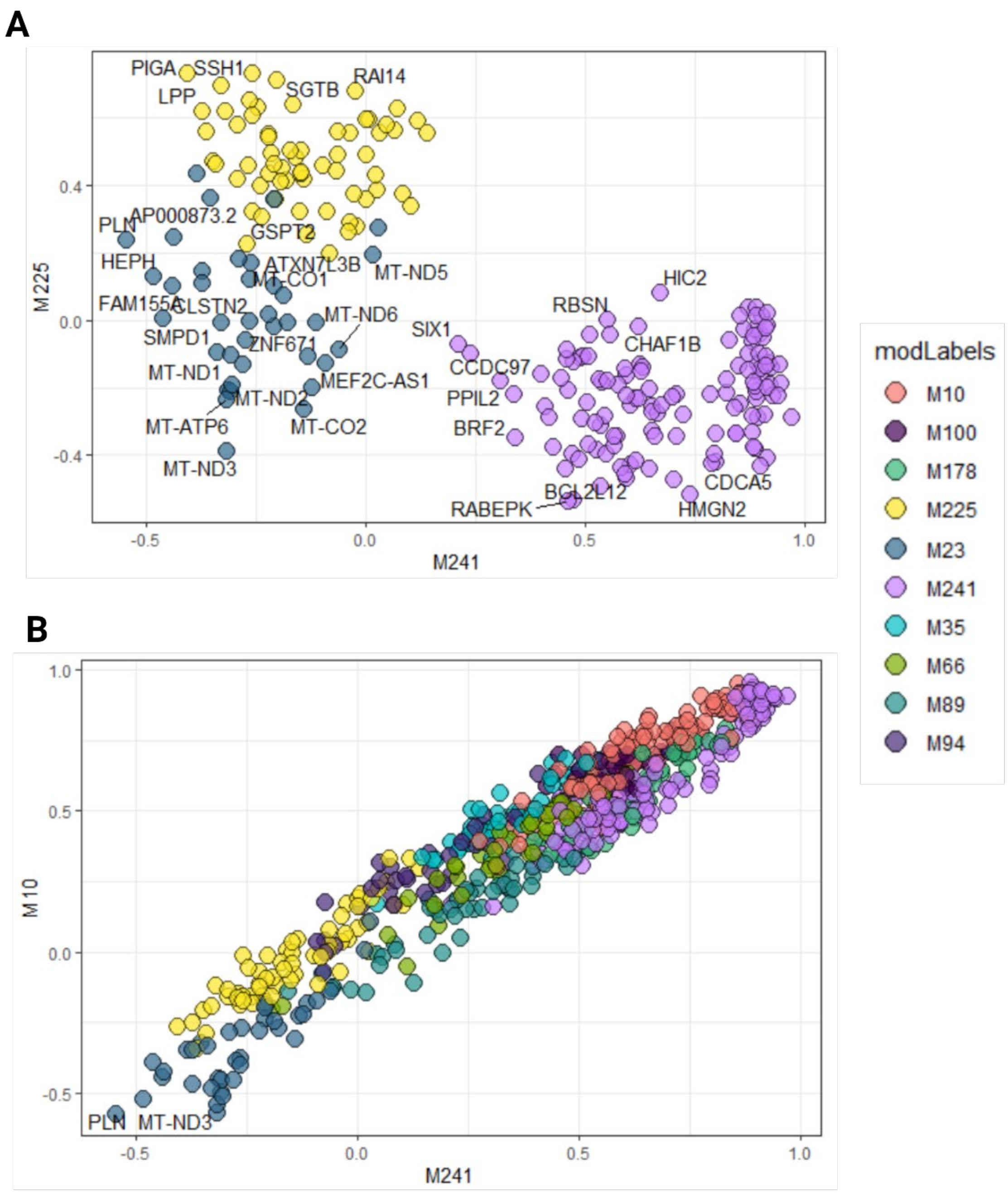


Figure 6.5: Gene module membership (kME – Spearman correlation between gene expression and ME expression) For **A:** M241, M225 and M23 module gene assignments. **B** The top ten modules. Colour indicates the module to which genes were clustered.

6.4.2 Neighbours of top modules are cell cycle related

Aside from M23 and M225, the top ten modules were enriched for cell cycle functions, where the top five modules (M241, M10, M66, M35 and M100) represent a tightly connected community (***Figure 6.1A; Figure 6.5B***). To explore other module relationships with cell cycle modules, a RWR was conducted on the EGN revealing five neighbouring modules; M140, M196, M156, M124 and M174 (***Figure 6.6A***). Not surprisingly, enrichment analysis indicated that these modules are enriched for cell cycle (***Figure 6.6B***), with M174 and M94 showing significant overlap for cycling cell markers in the scRNA-seq data (***Figure 6.6C***). The overlap of genes in M241 and the single-cell data was closer inspected revealing 22 overlapping genes that show high expression in cycling cell clusters including *UBE2C, UBE2T, CDK1,* and *TPX2* (***Figure 6.7A***). Notably, *SIX1* and *CRABP2* show high expression levels in tumour cells (***Figure 6.7A***). M10, the second highest ranked module, showed 11 overlapping genes highly expressed in cycling cells (***Figure 6.7B)*** including *TOP2A, CENPK, CENPW,* and *TUBB***.** Notably, *FLNC, PBK* and *HAS2* showed a high expression in tumour cells (***Figure 6.7B***).

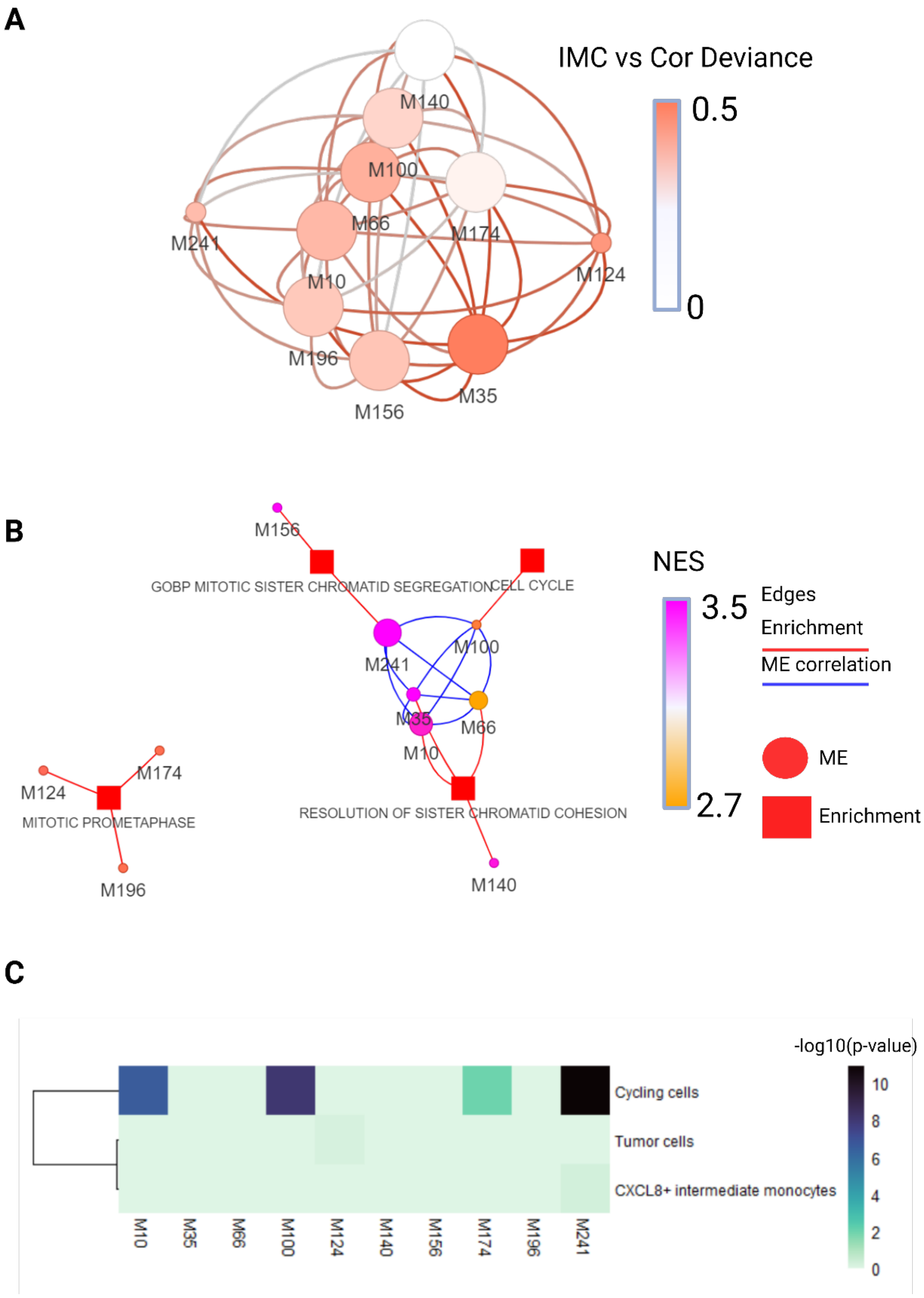


Figure 6.6: **A:** The top five negative modules and five nearest neighbouring seeds by RWR. Colour indicates the Spearman correlation between the intramodular connectivity (IMC) and the CorDeviance (Clinical GS measure). **B**: Enrichment-module network showing the module-eigengenes (MEs - circles) and their gene enrichment (both REACTOME and GOBP terms - squares). Blue edges indicate a correlation between MEs, and red edges indicate the top enrichment term for that module (as represented by its ME. **C:** Heatmap detailing top overlap of module genes with top cell type marker genes via an overlap test (overrepresentation, Fishers exact).

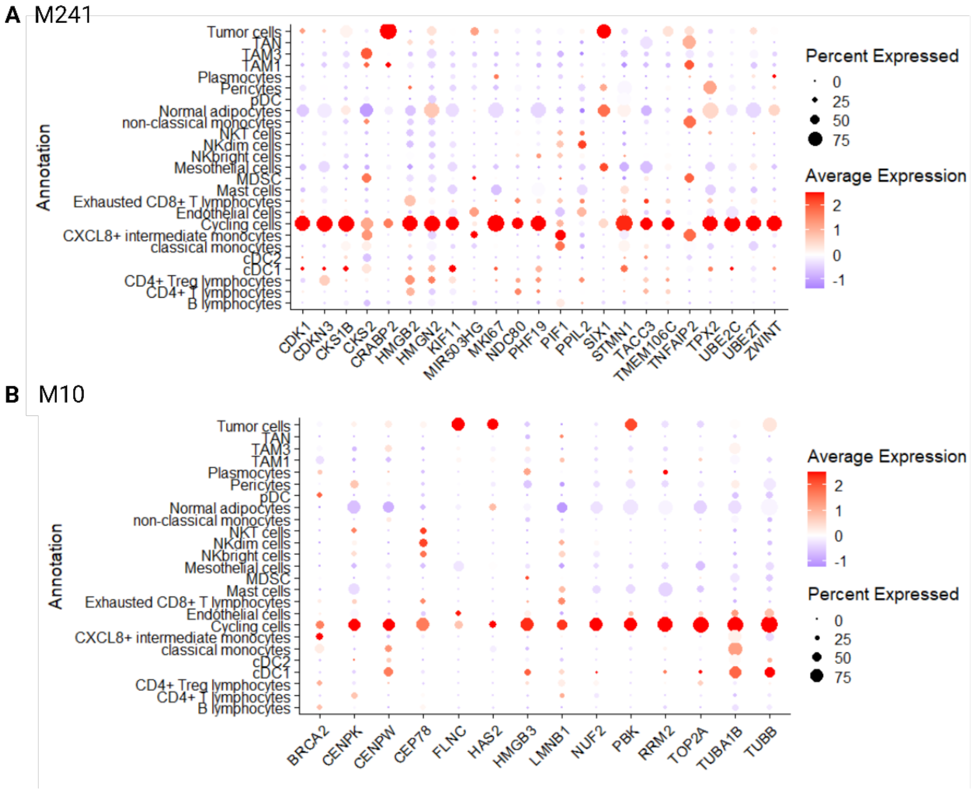


Figure 6.7: Expression of module genes in scRNA-seq data for **A** M241 **B** M10. Colour indicates the average expression among cells, and the size of the dot indicates the percentage of cells the genes are expressed in.

6.4.3 APC/C and Chromosome segregation

The presence of notable anaphase onset and mitotic-exit related genes in M241 (*CDK1, UBE2C, CDC20, PLK1, AURKA*), M178 (*PTTG1* - securin), and M10 (*TOP2A, RACGAP1, KIF23*) infers the role of the Anaphase Promoting Complex/Cyclosome (APC/C – UBE2C, CDC20) in anaphase-onset, and correct cytokinesis functions via the centralspindlin complex and its associated proteins (RACGAP1, KIF23, PLK1), among other genes shown to be crucial in metaphase progression and anaphase onset (e.g., *TOP2A*).[503,504] Notably, several of these genes show increased expression relative to normal adipose tissue (***Figure 6.8A***) and a known inhibitor *PPP2CA* (PP2A) shows downregulation. This led to an assessment of the E1 and E2 ligase environment within DDLPS vs Adipose, revealing differential E1 and E2 expression (***Figure 6.8B***). Of note, *UBE2C, PTTG1* (Securin), *UBE2T* all show a logFC > 1, and -log10(p) >2. Furthermore, many of the APC/C-related genes show a gene effect score of >0.5 suggesting that these genes have a moderate dependency in DDLPS cell lines (LPS141, LPS510, and LPS853) (***Figure 6.8C***).

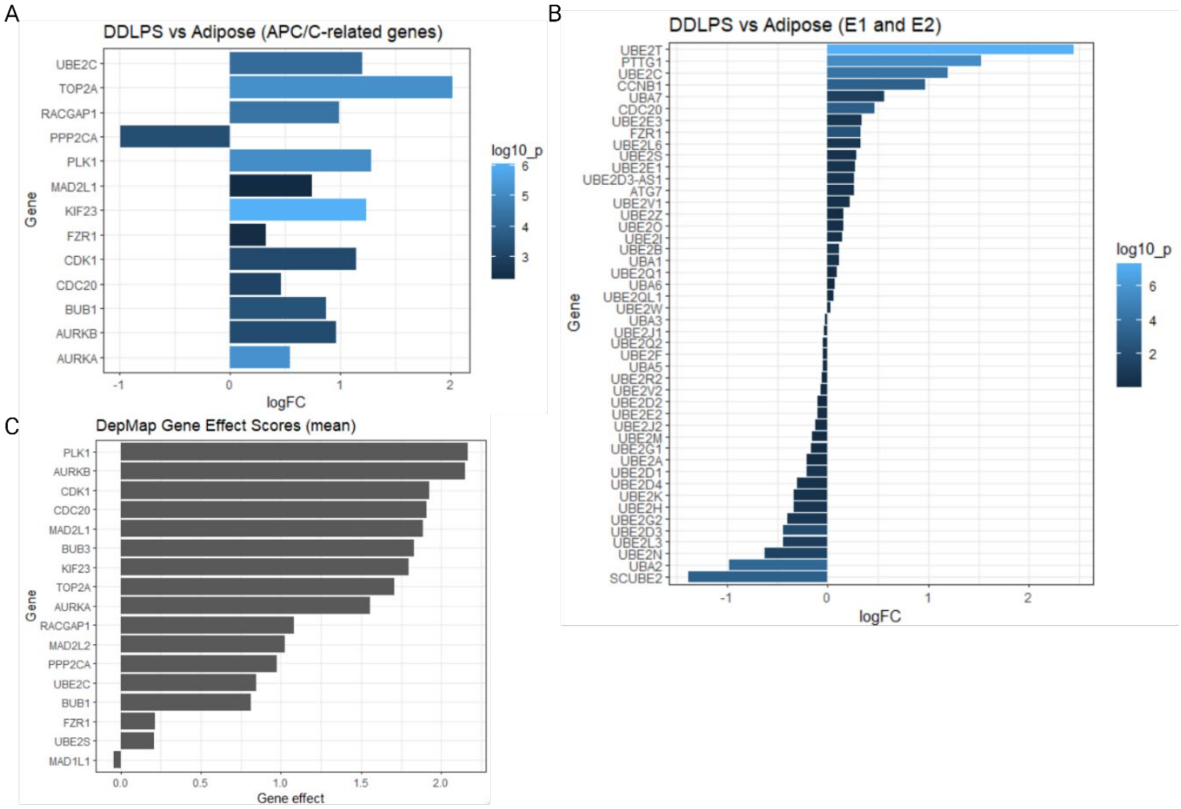


Figure 6.8: Exploration of APC/C- and E1/E2 ubiquitin enzymes. **A** Differential expression of APC/C-related genes in DDLPS versus adipose tissue (GSE159659). **B** Differential expression expanded to E1/E2 enzymes. Colour indicates the -log10(adjusted p-value) – filtered for those >1.3- from a differential expression test. logFC – log-fold change. **C** The mean gene effect scores from CRISPR-Cas9 sgRNA-abundance assays for three DDLPS cell lines – LPS141, LPS853, and LPS510. A gene effect of 0,5 indicates an anti-growth effect.

6.4.4 Sub-graph analysis

The next goal was to explore module at the subnetwork network level and conduct graph-based analysis to identify hub genes. A major aspect of this was to ensure the concordance in specific gene co-expression patterns between TCGA and NCC GCNs and define a concordant GCN subnetwork for each of the top ranked modules.

6.4.5 GCN stacking

The next goal was to explore top ranked modules at the subnetwork network level and conduct graph-based analysis to identify hub genes. A major aspect of this was to ensure the concordance in specific gene co-expression patterns between TCGA and NCC GCNs and define a "concordant" GCN subnetwork for each of the top ranked modules.

6.4.5.1 Top modules

Overall, as first evidenced in the strong preservation detected in module preservation analysis (***Table 6.1***), the TCGA and NCC GCNs show a high degree of concordance, using M241 as an example (***Figure 6.9A – TCGA right, NCC left***), but also indicates distinct co-expression profilers between these two data (***Figure 6.9B***). The stacked M241 network (***Figure 6.10A***) contained 204 nodes (genes) and 4980 edges with a mean degree connectivity of 83, a maximum degree of 212 (*CDK1*), a minimum of 2 (*NDC1, CENPK, TCF19, KIF20B, INCENP, SPC24, UHRF1*), and a density of 0.24. The top five connected genes were found to be *CDK1* (degree = 212), *UBE2C* (degree = 206), *AURKA* (degree = 192), *TTK* (degree = 192), and *KIF4A* (degree = 190) (***Figure 6.10B*** *– A reduced number of nodes for visualisation purposes*). The eigencentrality is highest for *UBE2C (*1.00) in the M241 sub-graph. This suggests that whilst *UBE2C* is not the most connected node, it does have the highest influence among network connectivity. The subsequent top ranked modules were then also assessed. It should be noted that at TOM > 0.05, M241 shows a very dense network with the 204 nodes and 19656 edges for a maximum possible of 20706, giving a density of 0.95. The increase to a TOM > 0.10 did not alter the hub structure but was done to provide easier visualisation.

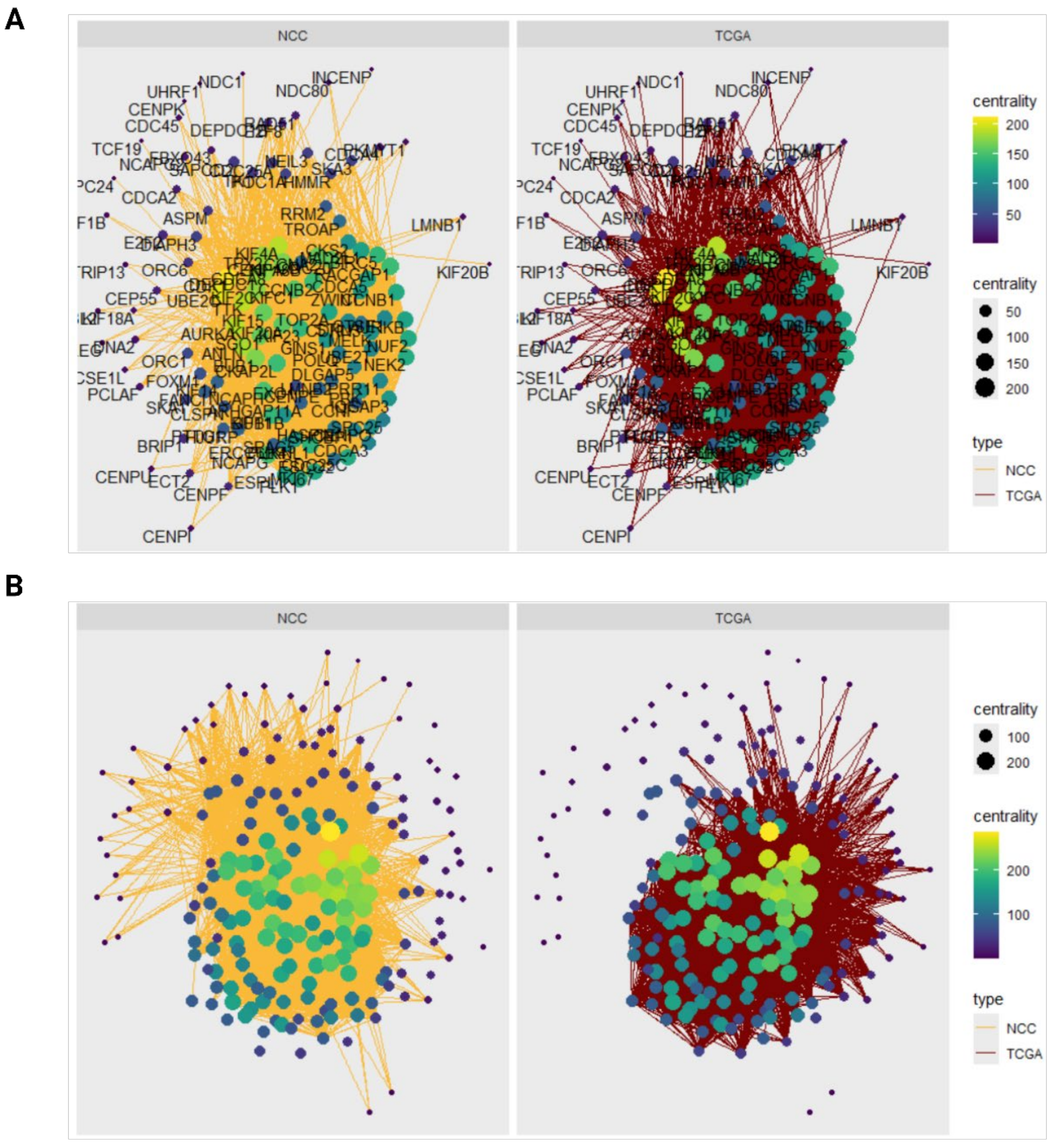


Figure 6.9: Multiplex networks for M241. **A:** Concordant edges between TCGA and NCC data. **B** No filtering for concordant edges revealing unique co-expression profiles between them.

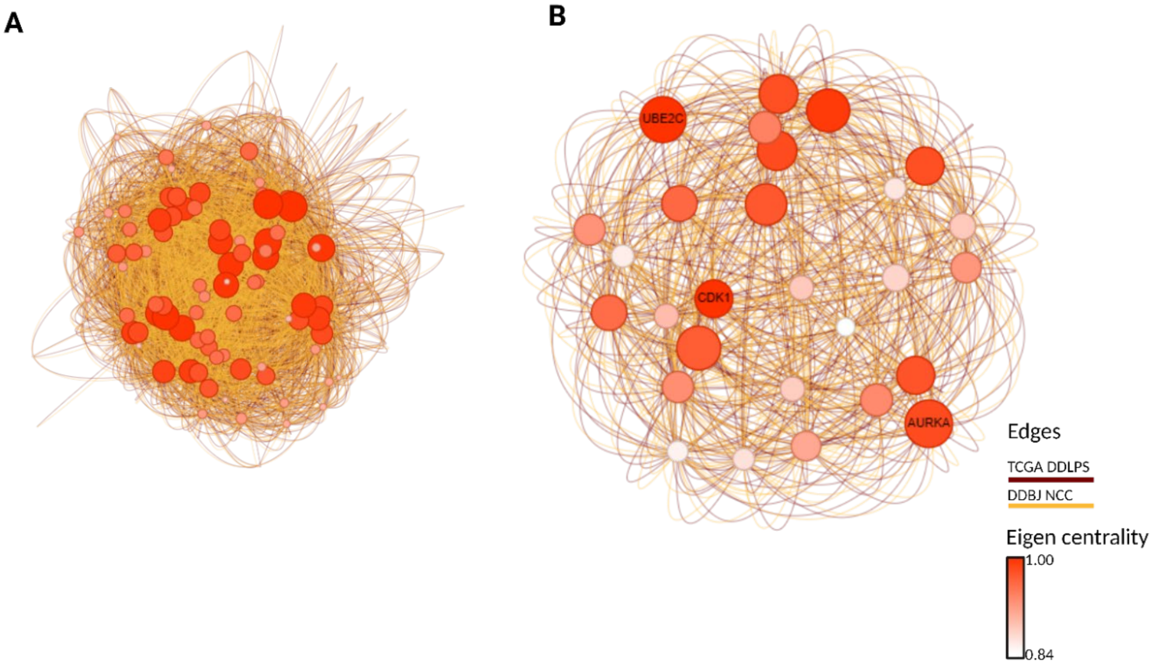


Figure 6.10: VisNetwork visualisation of the M241 stacked GCN. **A** The full 205 nodes (from RWR) using a TOM threshold of 0.10. **B** A node reduced network for degree centrality > 130 to better visualise eigencentrality hubs. Node colour represents the eigen centrality, node size is proportional to the degree centrality. Edges are separated by source dataset.

M10 and M241 show a tight relationship (***Figure 6.5B***) where there is strong kME overlap between these modules. A result of this is that when exploring the M10 subgraph using a kME > 0.5 threshold, the top hubs remain the same. In this context this could be a perceived benefit as M241 sub-graph analysis resembles more of a “meta” module encapsulating valuable information across modules, providing further strength to the importance of the hub genes (inferred). To identify patterns of connectivity associated with M10, genes were extracted based on module gene assignment. The top five hub-genes of M10 GCN (partition subset – ***Figure 6.11A***) were *CCNA2* (Degree = 102), *TOP2A* (degree = 100), *GTSE1* (degree = 98), *RACGAP1* (degree = 98), and *MELK* (degree = 96). *TOP2A* showed the highest eigencentrality. An edge threshold of TOM > 0.05 was used. The resulting graph had 99 nodes with 1, 558 edges, with a mean connectivity of 31.474. It is promising that M10 achieves a moderately dense network (density = 0.32).

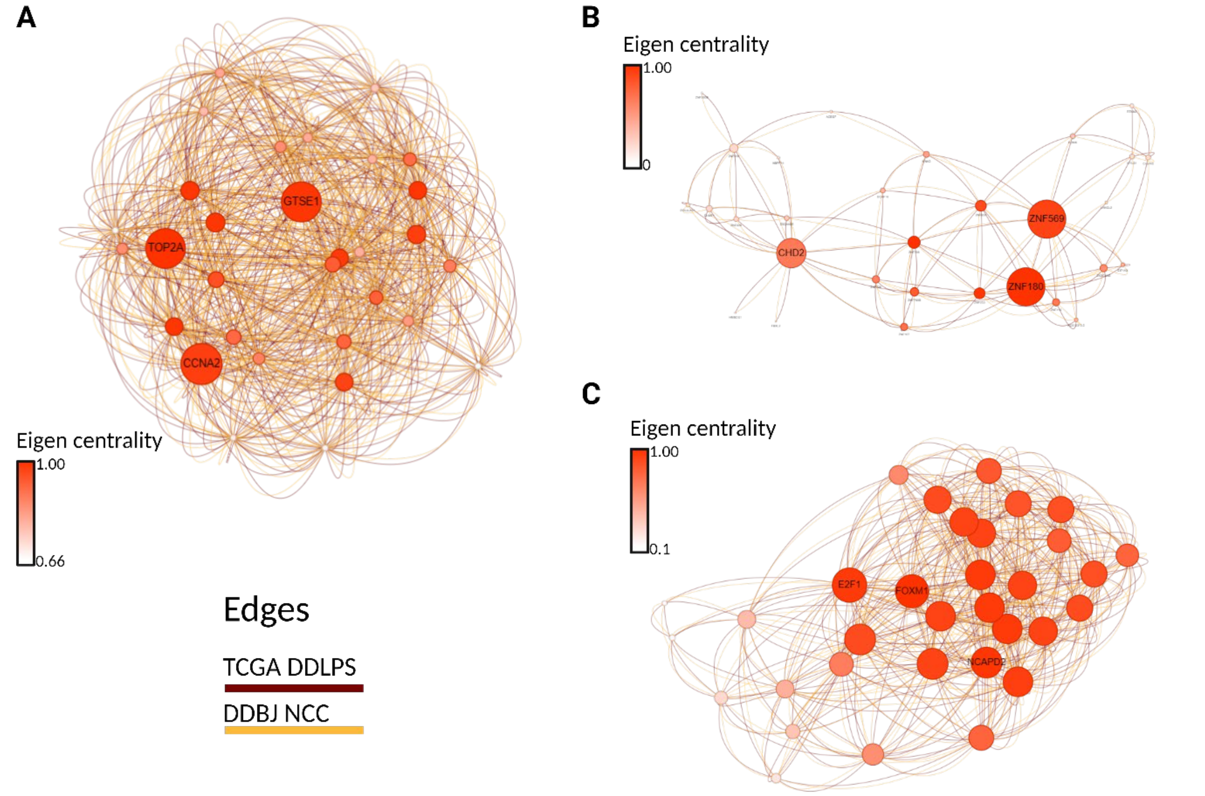


Figure 6.11: GCN subnetworks for **A** M10 – reduced subnetwork by degree centrality > 50, **B** M66 and **C** M100. Node colour represents the eigen centrality, node size is proportional to the degree centrality. Edges are separated by source dataset.

A similar module relationship was observed between M241 with M100 (*rho* = 0.79, *p* < 0.001) and M66 (*rho* = 0.55, *p* < 0.001). The M66 network (***Figure 6.11B***) was found to be node sparse with 37 nodes and 156 edges but a retain a density of 0.23, with a mean connectivity of 8.43, for a low TOM > 0.01 threshold. The top connected nodes were *CHD2* (degree = 26), *ZNF180* (degree = 22), *ZNF569* (degree = 22), *ZNF708* (degree = 20), and *ZNF253* (degree = 18). For the M100 network (***Figure 6.11C***) using a TOM edge threshold of 0.01, there were 33 nodes and 672 edges, giving a mean connectivity of 40.727 and a network density of 0.64. The top connected genes were found to be *E2F1* (degree = 58), *FOXM1* (degree = 56), *NCAP2* (degree = 52), *EZH2* (degree = 52) and *RECQL4* (degree = 52).

M35 and M241 also showed a significant correlation (*rho* = 0.58, *p < 0.001*) and hence module gene selection was used resulting in a network (***Figure 6.12A***) with 44 nodes, 656 edges, a mean connectivity of 29.82, and a network density of 0.69. The five top connected nodes were found to be *WDHD1* (degree = 68), *TOPBP1* (degree = 68), *CCDC138* (degree = 58), *MSH2* (degree = 56), and *XPO1* (degree = 54) with *WDHD1* showing the highest eigencentrality.

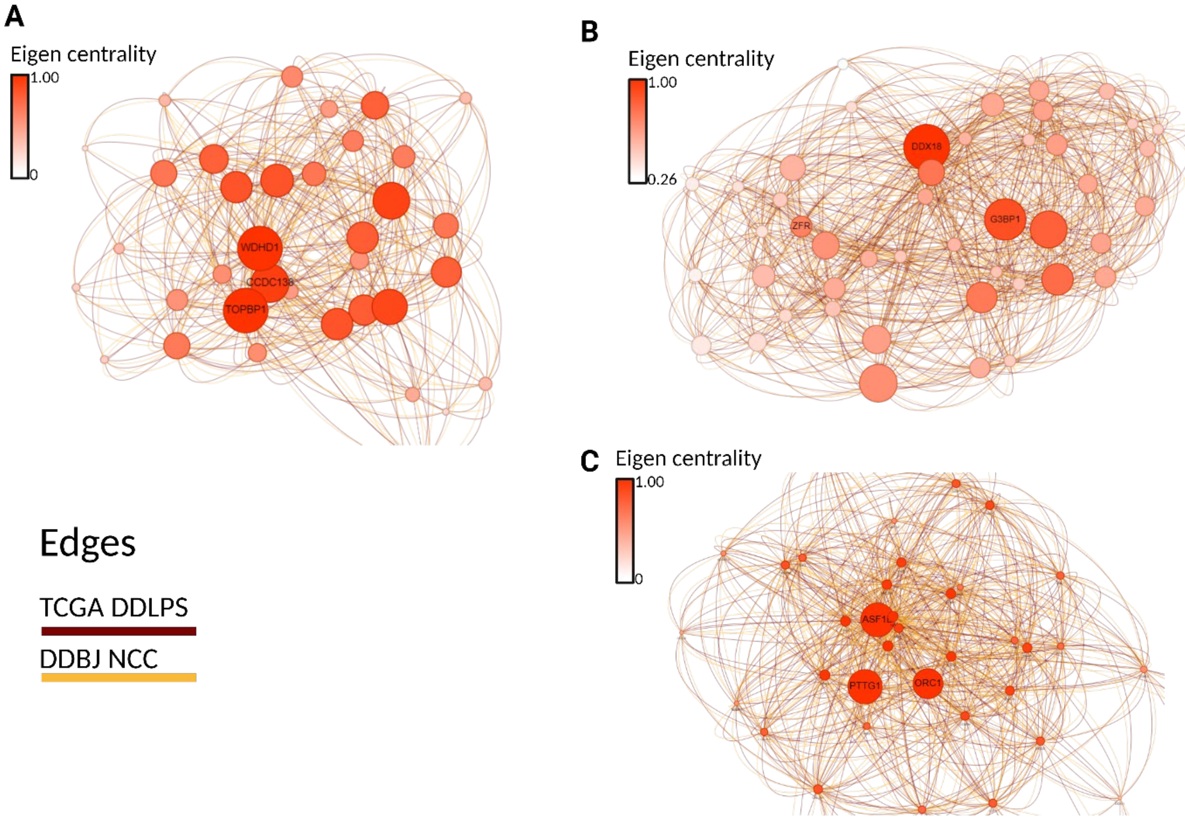


Figure 6.12: GCN subnetworks for **A** M35, **B** M94– reduced subnetwork by degree centrality > 50 and **C** M178. Node colour represents the eigen centrality, node size is proportional to the degree centrality. Edges are separated by source dataset.

M94 and M241 were found to have a low correlation that was not significant (*rho* = 0.22, *p* = 0.18). M94 genes were selected based on the kME >0.5 criteria and using an edge threshold of TOM > 0.03. The M94 GCN subnetwork (***Figure 6.12*B**) was found to contain 204 nodes across 5220 edges, with a mean connectivity of 51.18 and a network density of 0.25. The top five nodes were found to be *ZFR* (degree =254)*, DDX18* (degree =250)*, G3BP1* (degree = 228)*, KIF5B* (degree = 210)*, ACTR2* (degree = 210).

M178 showed a high and significant correlation to M241 (*rho* = 0.90, *p* < 0.001) overlap with M241 kME > 0.5. The M178 graph was found to have 51 nodes, 992 edges with a threshold of TOM > 0.01, a mean degree of 38.90 and a network density of 0.78. the top five nodes of this *PTTG1* (degree = 70)*, ORC1* (degree = 68), *ASF1B* (degree = 68), *CDK2* (degree = 68) and *TCF19* (degree = 68). TOM > 0.01. PTTG1 had the highest eigencentrality.

M89 as previously shown was not found to be correlated with M241, using the kME to select for genes a network (***Figure 6.12C***) with 204 nodes, 6036 edges at TOM > 0.01 with a mean connectivity of 59.18, and a network density of 0.29. The hubs were found to be *RGS12* (degree = 300), *HIF3A* (degree = 246), *CNTFR* (degree = 242), *CACNA2D2* (degree = 222), and *EMLIN3* (degree = 208). *RGS12* showed the highest eigen centrality (1).

M23 sub-network for the kME partition showed nodes for MT-CYB (degree = 74), MTATP6P1 (degree = 72), MT-ND4 (degree = 70), MT-ATP6 (degree = 62) and CCDC17 (degree = 52). The mean connectivity was found to be 17.756 across 82 nodes and 728 edges at a TOM > 0.01 threshold with a density of 0.2 (***Figure 6.13A***).

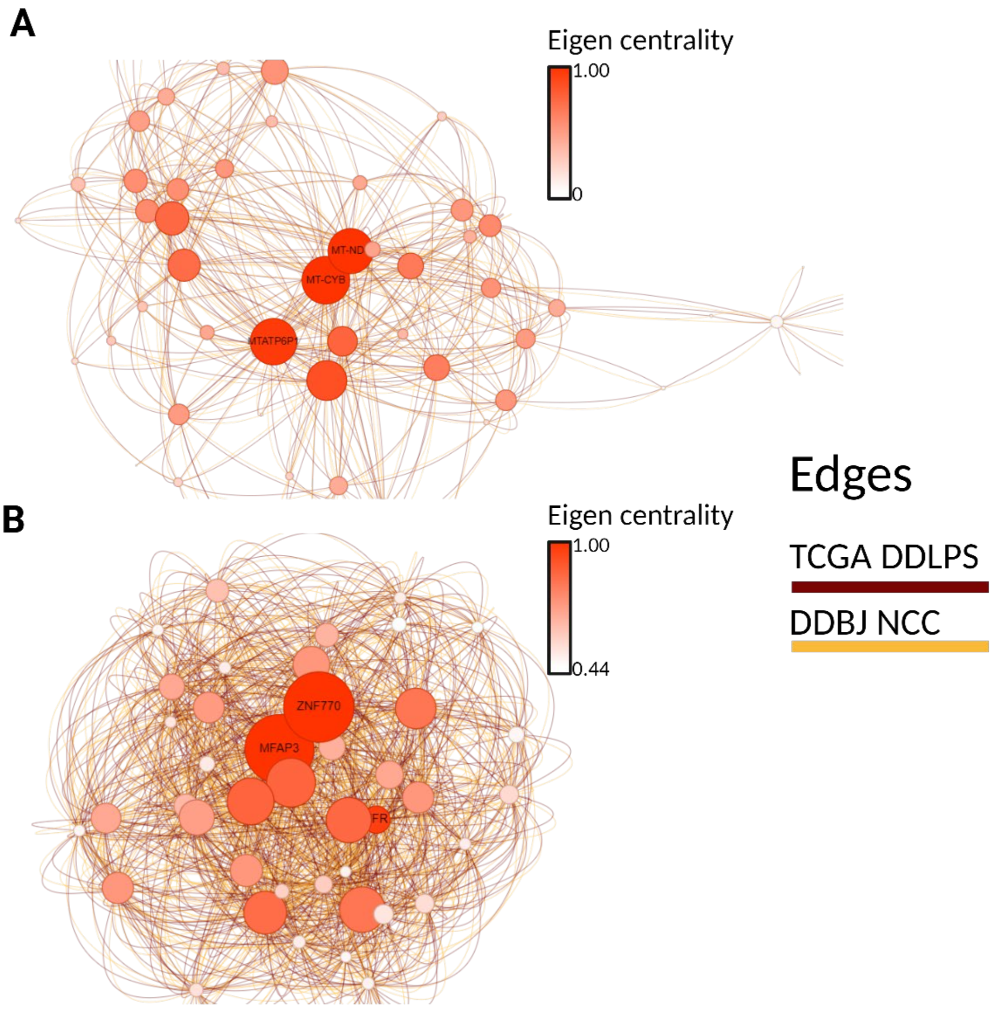


Figure 6.13: GCN subnetworks for **A** M23, **B** M225– reduced subnetwork by degree centrality > 100. Node colour represents the eigen centrality, node size is proportional to the degree centrality. Edges are separated by source dataset.

M225 (***Figure 6.13B***) had a kME network of 204 nodes across 7026 edges at a threshold of TOM > 0.03 with a mean connectivity of 68.88, and a network density of 0.34. The top five connected nodes are *ZNF770* (degree = 370), *MFAP3* (degree = 364), *ZFR* (degree = 330), *ERBIN* (degree = 274), *G3BP2* (degree = 264). ZFN770 showed the highest eigen centrality.

As nodes with the highest eigen centrality represent those with the highest influence over the network, these were selected for further analysis as the module “hub” (***Table 6.2***). This leaves the following nine candidates: *UBE2C, TOP2A, WDHD1, FOXM1, DDX18, PTTG1, RGS12, MTCYB*, *ZNF770* and *ZNF180*.

***Table 6.2*:** Summary of top module subnetwork analysis.

Module	Subset	TOM	Top nodes (degree)	Eigencentrality hub	Gene name	Number of Nodes (edges)	Mean degree	Density
M241	kME > 0,5	0.10	*CDK1* (212), *UBE2C* (206), *AURKA* (192), *TTK* (192), *KIF4A* (190)	UBE2C	Ubiquitin Conjugating Enzyme E2 C	204 (4980)	48.82	0.24
M10	Partition	0.05	*CCNA2* (102), *TOP2A* (100), *GSE1* (98), *RACGAP1* (98), and *MELK* (96)	TOP2A	Topoisomerase II Alpha	99 (558)	31.47	0.32
M66	Partition	0.01	*CHD2* (26), *ZNF180* (22), *ZNF569* (22), *ZNF708* (20), and *ZNF253* (18)	ZNF180	Zinc Finger Protein 180	37 (156)	8.43	0.24
M35	Partition	0.01	*WDHD1 (68), TOPBP1 (68), CCDC138 (58), MSH2 (56), and XPO1 (54)*	*WDHD1*	WD Repeat And HMG-Box DNA Binding Protein 1	44 (656)	29.82	0.69
M100	Partition	0.01	*E2F1* (58), *FOXM1* (56), *NCAP2* (52), *EZH2* (52) and *RECQL4* (52)	*FOXM1*	Forkhead Box M1	33 (627)	40.73	0.64
M94	kME > 0.5	0.03	*ZFR* (254)*, DDX18* (250), *G3BP1* (d228)*, KIF5B* (210)*, ACTR2* (210)	*DDX18*	DEAD-Box Helicase 18	204 (5220)	51.18	0.25
M178	Partition	0.01	*PTTG1* (70), *ORC1* (68), *ASF1B* (68), *CDK2* (68) and *TCF19* (68)	*PTTG1*	Pituitary Tumor-Transforming Gene 1	51 (992)	38.90	0.78
M89	kME	0.02	*RGS12* (300), *HIF3A* (246), *CNTFR* (242), *CACNA2D2* (222)	*RGS12*	Regulator Of G Protein Signaling 12	204 (6036)	59.18	0.29
M23	kME	0.01	for MT-CYB (74), MTATP6P1 (72), MT-ND4 (70), MT-ATP6 (62) and CCDC17 (52)	*MTCYB*	Mitochondrially Encoded Cytochrome B	82 (728)	17.76	0.21
M225	kME	0.03	*ZNF770* (370), *MFAP3* (364), *ZFR* (330), *ERBIN* (274), *G3BP2* (264).	*ZNF770*	Zinc Finger Protein 770	204 (7026)	68.88	0.34

Subset: The selection criteria by which the TOMs were subset to identify subnetworks. **kME –** gene module membership. **TOM –** Topological Overlap Matrix is the weighted GCN. **Density** – Proportion of edges observed to the maximum number of potential edges according to a signed network – density = E/(N-(N-1)/2), where E is the observed number of edges, and N is the observed number of nodes.

6.4.5.2 Hub-genes as drug targets

These ten hubs were taken as candidates for drug-target screening using the TTD and the CPP drug-target databases to obtain a list of inhibitors. The results (Table 6.3) of this analysis revealed TOP2A targeted by NK314 and camsirubicin and FOXM1 targeted by D01FSW. NK314 (TOP2A) and D01FSW (FOXM1) are investigative drugs. Camsirubicin (TOP2A) is in Phase-2 clinical trials in unresectable STS (NCT02267083).[505] At present, of the nine candidates proposed, only TOP2A has an available drug in Phase-2 clinical trials. To highlight more drug-target interactions, the known pathways/interactors of these nine candidates will also be inspected for targeting. This could be conducted by expanding the number of top ranked genes (by eigencentrality/degree) although a central hypothesis of this project is that hub genes are key molecules.

Table 6.3: Hub gene target screening results

Target	TTD Drug ID	Drugname	Status
FOXM1	D01FSW		Investigative
TOP2A	DGZ38L	NK314	Investigative
TOP2A	DES53JQ	Camsirubicin	Clinical Trial – phase 2

Status – Highest clinical investigation/approval status. TOP2A – Topoisomerase 2A, FOXM1 – Forkhead box M1. TTD – Therapeutic Target Database drug ID.

6.4.5.3 Protein-protein interaction networks and drug-target integration

Next protein-protein interaction data was extracted from STRING.db v12.0 2024 release. This was data was filtered according to the nine candidates (individually) to identify PPIs and build a PPI network with drug-target information.

6.4.5.4 Drug screening of hubs and their protein interactors

UBE2C PPI network revealed four proteins in the network that were targeted by eight drugs (Table 6.4). Of these drugs, one (TAK-243) is currently in a phase 1 clinical trial targeting Ubiquitin-activating enzyme E1 (UBA1) (Figure 6.14A). The TOP2A PPI network revealed just one interaction between TOP2A and Bromodomain-containing protein 4 (BRD4). In addition to the two drugs targeting TOP2A (section 6.4.4.2) 55 further drugs targeting BRD4 were identified (Table 6.5). Six of these drugs are in phased clinical trials. In addition to camsirubicin (TOP2A) five targeted BRD4: D07GHA, DOP7FM, AZD5153, ABBV-744, and (+)-JQ1 (Figure 6.14B).

Inspection of PTTG1 PPIN revealed six proteins with seven edges. Similar to the UBE2C PPIN, CDC20 was present and targeted by DSU7K2, a pre-clinical drug, and was the only drug-target interaction retrieved. The MT-CYB PPIN was found to be very dense and retrieved four drug compounds targeting two proteins in the network (Table 6.6). D0GV9Q: N-Formylmethionine, D07FRX: 6-Thiophen-3-yl-imidazo[2,1-b]thiazole, D0J3TS: 6-Thiophen-2-yl-imidazo[2,1-b]thiazole, are investigative drugs targeting MT-ND3. Flurpiridaz F 18 is a phase 3 clinical trial candidate targeting NDUFA13 (Figure 6.14C).

Table 6.4: Targets identified within the UBE2C protein-protein interaction network.

Targets – UBE2C	TTD Drug ID	Drugname	Stage
UBA1	D0T4PA	TAK-243	Clinical Trial – phase 1
UBA1	D0YE0V		Patented
UBA1	D0QQ3B	SCHEMBL15198146	Patented
UBA1	D0P3PT	SCHEMBL15198145	Patented
UBA1	D0A0YZ	PYZD-4409	Investigative
UBC	D0B2AE		Patented
UBC	DOJA8J		Patented
CDC20	DSU7K2	Tosyl-l-arginine methyl ester (TOME)	Pre-clinical

Status – Highest clinical investigation/approval status. UBA1 – Ubiquitin activating enzyme E1, UBC – Ubiquitin C, CDC20 – cell division cycle 20. Bold indicates a drug with disease indications being explored in a clinical trial.

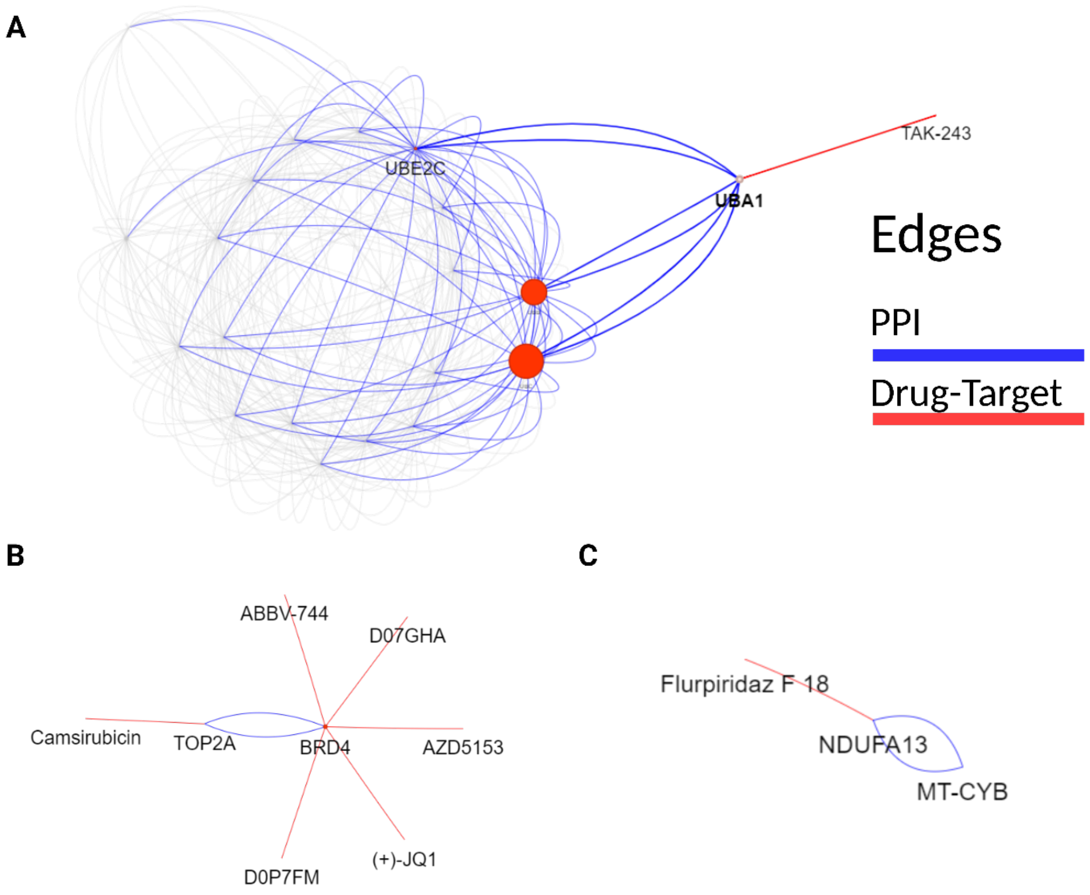


Figure 6.14: Drug-target networks for A: The UBE2C/UBA1 protein-protein interaction network, with UBA1 the target of TAK-243. B: TOP2A/BRD4 and C: MT-CYB/NDUFA. Nearest (k = 1) neighbours highlighted. Protein-protein interaction (PPI) edges are blue, drug-target interaction edges are red.

Table 6.5: Targets identified within the TOP2A protein-protein interaction network.

Target – TOP2A	TTD drug ID	Drugname	Status
BRD4	D03LIP	GW841819X	Investigative
BRD4	D0WU1S	I-BET151	Investigative
BRD4	D0T3YY	MS417	Investigative
BRD4	D09HBR	MS436	Investigative
BRD4	D01YSN	isoxazole azepine compound 3	Investigative
BRD4	D03LNF	PFI-1	Investigative
BRD4	D01DMN	XD1	Investigative
BRD4	D07FGY	XD14	Investigative
BRD4	D09ZPM	PMID25408830C1	Investigative
BRD4	D01EVE	PMID25703523C7d	Investigative
BRD4	D0O2RX	PMID23517011C9	Investigative
BRD4	D0C2JS	CPI-203	Investigative
BRD4	D0J6XI	PMID25408830C2	Investigative
BRD4	D03BTJ	PMID25408830C3	Investigative
BRD4	D0M5DB	BzT-7	Investigative
BRD4	D0C2XT	PMID24000170C36	Investigative
BRD4	D0Q2GK	PMID24000170C38	Investigative
BRD4	D0H3TC	PMID21851057C4d	Investigative
TOP2A	DGZ38L	NK314	Investigative
BRD4	D0ZW4W	(+)-JQ1	Phase 1
BRD4	D05ICT	ABBV-744	Phase 1
BRD4	D0PH9I	AZD5153	Phase 1
BRD4	D07GHA	NA	Phase 1/2
BRD4	D0P7FM	NA	Phase 1/2
TOP2A	DE53JQ	Camsirubicin	Phase 2
BRD4	D04HTM	Aminocyclopentenone compound 1	Patented
BRD4	D08JHQ	Aminocyclopentenone compound 2	Patented
BRD4	D09QSH	Aminocyclopentenone compound 5	Patented
BRD4	D0AU3O	NA	Patented
BRD4	D0KF8C	Aminocyclopentenone compound 3	Patented
BRD4	D0N6JZ	Aminocyclopentenone compound 4	Patented
BRD4	D09FVR	Pyrrolo-pyrrolone derivative 3	Patented
BRD4	D0HO8V	Pyrrolo-pyrrolone derivative 4	Patented
BRD4	D0PE1R	Pyrrolo-pyrrolone derivative 2	Patented
BRD4	D0XR9Y	Pyrrolo-pyrrolone derivative 5	Patented
BRD4	D0YM3G	Pyrrolo-pyrrolone derivative 1	Patented
BRD4	D00MXR	PMID26924192-Compound-23	Patented
BRD4	D03NWT	PMID26924192-Compound-104	Patented
BRD4	D05LTB	Pyrazole and thiophene derivative 4	Patented
BRD4	D05XXS	PMID26924192-Compound-103	Patented
BRD4	D07IYX	PMID26924192-Compound-20	Patented
BRD4	D08BCK	PMID26924192-Compound-24	Patented
BRD4	D0C9RQ	PMID26924192-Compound-22	Patented
BRD4	D0CB8D	PMID26924192-Compound-31	Patented

Target – TOP2A	TTD drug ID	Drugname	Status
BRD4	D0EN4B	Pyrazole and thiophene derivative 2	Patented
BRD4	D0FN6F	PMID26924192-Compound-25	Patented
BRD4	D0J4QG	Pyrazole and thiophene derivative 3	Patented
BRD4	D0L7SK	PMID26924192-Compound-30	Patented
BRD4	D0NN4U	PMID26924192-Compound-105	Patented
BRD4	D0SB3H	PMID26924192-Compound-32	Patented
BRD4	D0SQ0F	Pyrazole and thiophene derivative 1	Patented
BRD4	D0VB3P	PMID26924192-Compound-33	Patented
BRD4	D0XM8B	PMID26924192-Compound-21	Patented
BRD4	D03RPJ	PMID26924192-Compound-102	Patented
BRD4	D07CSE	Benzothiazepine analog 12	Patented
BRD4	D0S5ID	Benzothiazepine analog 11	Patented
BRD4	D0Y7YC	Benzothiazepine analog 10	Patented

Status – Highest clinical investigation/approval status. TOP2A – Topoisomerase 2A, BRD4 – Bromodomain-containing protein 4. Bold indicates a drug with disease indications being explored in a clinical trial.

Table 6.6: Targets identified within the MT-CYB protein-protein interaction network.

Targets – MT-CYB	TTD Drug ID	Drugname	Status
MT-ND3	D0GV9Q	N-Formylmethionine	Investigative
MT-ND3	D07FRX	6-Thiophen-3-yl-imidazo[2,1-b]thiazole	Investigative
MT-ND3	D0J3TS	6-Thiophen-2-yl-imidazo[2,1-b]thiazole	Investigative
NDUFA13	D0U1VZ	Flurpiridaz F 18	Phase 3

Status – Highest clinical investigation/approval status. NDUFA13 – NADH:ubiquinone oxidoreductase subunit A13, MT-ND3 – NADH dehydrogenase subunit 3. Bold indicates a drug with disease indications being explored in a clinical trial.

In summary, 63 drug-target interactions were identified. Three of which (5%) were targeted towards the candidate gene. The remainder 59 were targeted against BRD4 (55 drugs), CDC20 (one drug), NDUFA13 (one drug), and MT-ND3 (three drugs). Eight of these drugs (7.88%) were found to be currently in phased clinical trials. Of these, Flurpiridaz F 18 (NDUFA13) is in phase 3, TAK-243 (UBA1), (+)-JQ1, ABBV-744, AZD5153 are in phase 1 clinical trials, D07GHA, D0P7FM each targeting BRD4 are in phase 1/2 clinical trials, and Camsirubicin (TOP2A) in phase 2. No drug targets were identified for KMT2D, RGS12 and no PPIs were identified for FOXM1, DDX18, ZNF770. **Table 6.7** summarises the drug compounds currently in phased clinical trials.

Table 6.7: Summary of drug compounds currently in phased clinical trials.

Targets	TTD Drug ID	Drugname	Status
NDUFA13	D0U1VZ	Flurpiridaz F 18	Phase 3
BRD4	D0ZW4W	(+)-JQ1	Phase 1
BRD4	D05ICT	ABBV-744	Phase 1
BRD4	D0PH9I	AZD5153	Phase 1
BRD4	D07GHA	NA	Phase 1/2
BRD4	D0P7FM	NA	Phase 1/2
TOP2A	DE53JQ	Camsirubicin	Phase 2
UBA1	D0T4PA	TAK-243	Phase 1

Status – Highest clinical investigation/approval status. NDUFA13 – NADH:ubiquinone oxidoreductase subunit A13, TOP2A – Topoisomerase 2A, UBA1 – Ubiquitin Activating Enzyme E1, BRD4 – Bromodomain-containing protein 4.

6.4.5.5 Further candidate exploration

Targeted (directly or through interactors) candidates were further assessed on differential expression, dependency and survival association. As previously detailed (Figure 6.8) TOP2A and UBE2C show significant differential expression, BRD4, MT-CYB and NDUFA13 do not (Figure 6.15A). BRD4, UBE2C show gene effect scores of >0.5 suggesting cell depletion upon knockdown (Figure 6.15B). TOP2A and UBA1 show a strong depletion effect >1.0. Only BRD4 was found to significantly predict overall survival in the Multivariate Cox Ph model (as described in section 3.4.3.1.1). Furthermore, this association was reproducible in the NCC data using the progression free-survival measure (Figure 6.16A-B).

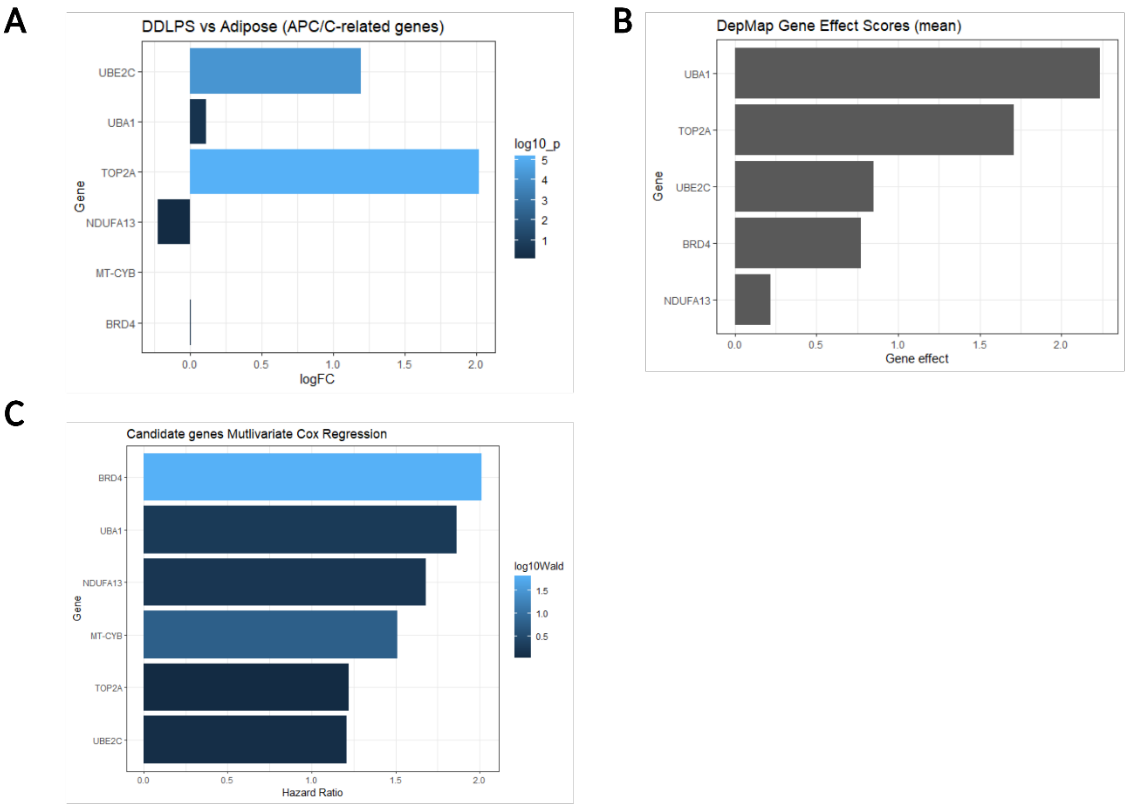


Figure 6.15: A Differential gene expression result (GSE159659) for DDLPS vs adipose. Colour indicates the -log10(BH adjusted p value). B The mean gene effect sizes from the DepMap CRISPR-cas9 sgRNA abundance assays across DDLPS cell lines (LPS141, LPS510, LPS853). C The hazard ratio from a multivariate Cox Ph (TCGA DDLPS) model on overall survival. Colour indicates the -log10(Wald test p value).

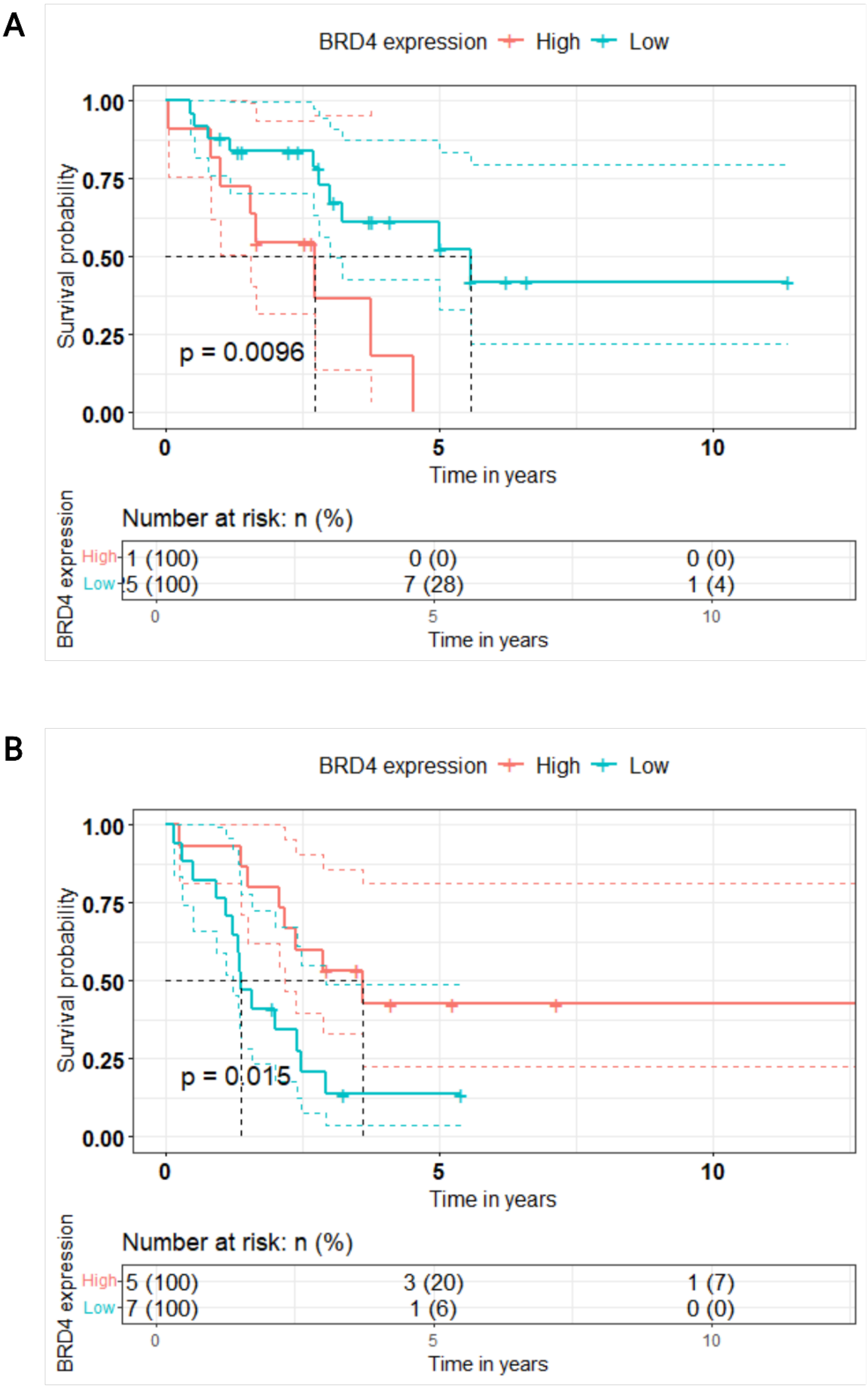


Figure 6.16: Univariate survival analysis for BRD4 expression in A TCGA and B NCC

6.4.6 Gene expression of key genes are typically not correlated with tumour ploidy

Tumour ploidy is a measure of chromosomal instability. Ploidy variable was taken from the clinical metadata available from the TCGA SARC publication.[64] The analysis was not conducted on NCC samples as ploidy count was not available in the clinical metadata. During the assessment of outliers in **section 4.3.2** it was found that ploidy was not a significant variable between included and excluded samples. It was found in **section 3.4.3.1.1** (***Table 3.4***) the mean tumour ploidy was 2.7 which denoted moderate aneuploidy and classified here as near triploid (NT) suggesting genomic complexity within the cohort, although most tumours were estimated as being closer to diploid in both the original data (n = 50, 64.5% ; ***Table 6.8***) and the filtered dataset (n = 36, 56.8% ; ***Table 6.9***). It was noted that tumour ploidy estimates were available for 31/36 (n = 36) and 44/50 (n = 50) of tumour samples.

Table 6.8: Tumour ploidy classes in the original TCGA DDLPS data (n = 50).

Ploidy class	Number in class	Percentage of total (%)
Near diploid	25	56.8
Moderate aneuploid	2	4.5
Near Triploid	7	15.9
Near Whole Genome Duplication	5	11.4
Whole Genome Duplication	5	11.4

Percentage (%) was calculated as a percentage of the number in class compared to the number of tumour samples for which tumour ploidy was available.

Table 6.9. Tumour ploidy classes in the filtered TCGA DDLPS data (n = 36).

Ploidy class	Number in class	Percentage of total (%)
Near diploid	20	64.5
Moderate aneuploid	1	3.2
Near Triploid	4	12.9
Near Whole Genome Duplication	3	9.7
Whole Genome Duplication	3	9.7

Percentage (%) was calculated as a percentage of the number in class compared to the number of tumour samples for which tumour ploidy was available.

Key genes identified in this study (network hubs and drug targets) which included *YEATS4*, *UBE2C*, *UBA1*, *TOP2A*, *BRD4*, *NDUFA13* along with those that are known drivers for DDLPS including, MDM2, CDK4, CPM, FRS2, HMGA2 and YEATS4. CPM was found to not be present in the gene expression data post-filtering. Tumour ploidy is incorporated to distinguish whether there is a correlation between increasing tumour ploidy and the variability in gene expression (and thus co-expression) patterns observed in these key genes. If variability is driven by transcriptional regulation modifications that are more focal in nature and not the result of global and passive increased genomic content.

For most genes there was no significant correlation identified in both the filtered (n = 36; ***Figure 6.17***) and original (n =50; ***Figure 6.18***) data. The exception was ***BRD4*** in the filtered data (***Pearson r = 0.42, p = 0.019***) showing a moderate positive correlation in the filtered data (n = 36). However, this could not be reproduced when assessing significance with tumour ploidy categorised (***Figure 6.19***) into discreet groups.

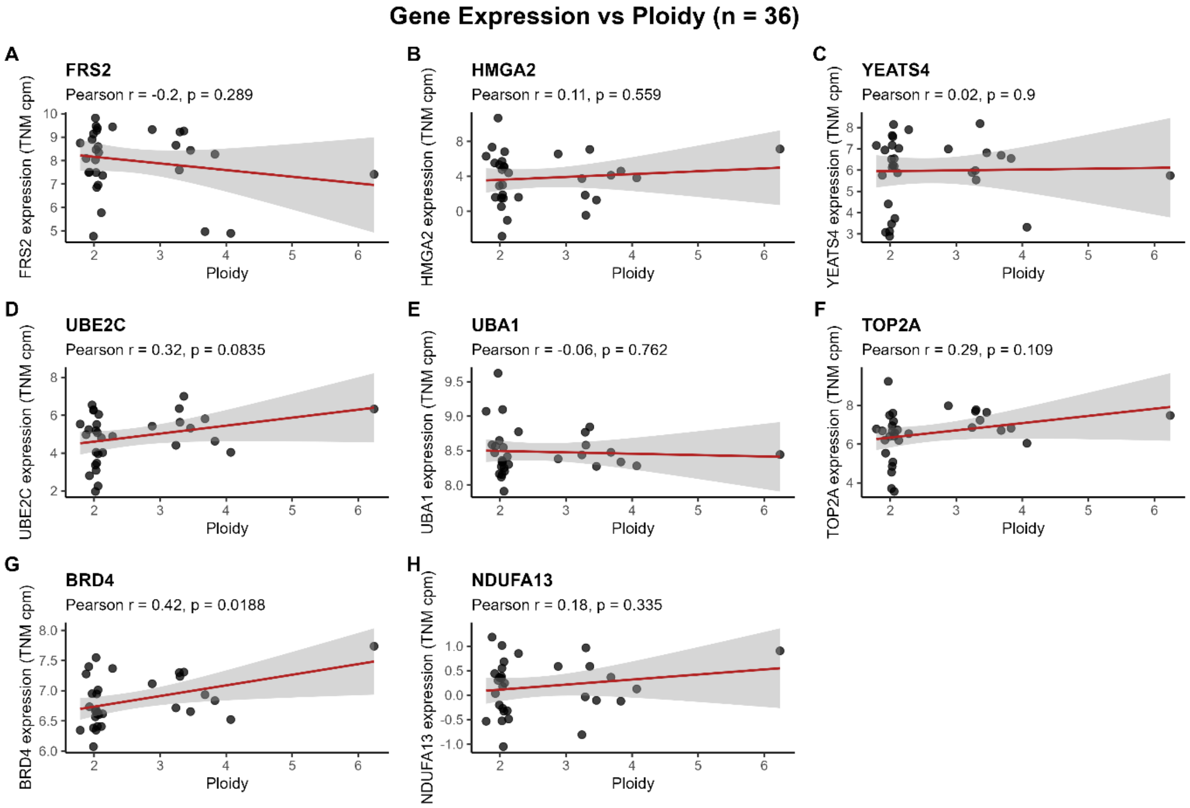


Figure 6.17: Pearson correlation of tumour ploidy and gene expression for key genes in the filtered TCGA DDLPS data (n = 36). ***A***- FRS2 (Fibroblast Growth Factor Receptor Substrate 2), ***B*** - HMGA2 (High Mobility Group AT-Hook 2), ***C*** - *YEATS4* (YEATS Domain Containing 4), ***D*** - *UBE2C* (Ubiquitin Conjugating Enzyme E2 C), ***E*** - *UBA1* (Ubiquitin Like Modifier Activating Enzyme 1), ***F***- *TOP2A* (DNA Topoisomerase II Alpha), ***G*** - *BRD4* (Bromodomain Containing 4), and ***H*** - *NDUFA13* (NADH: Ubiquinone Oxidoreductase Subunit A13).

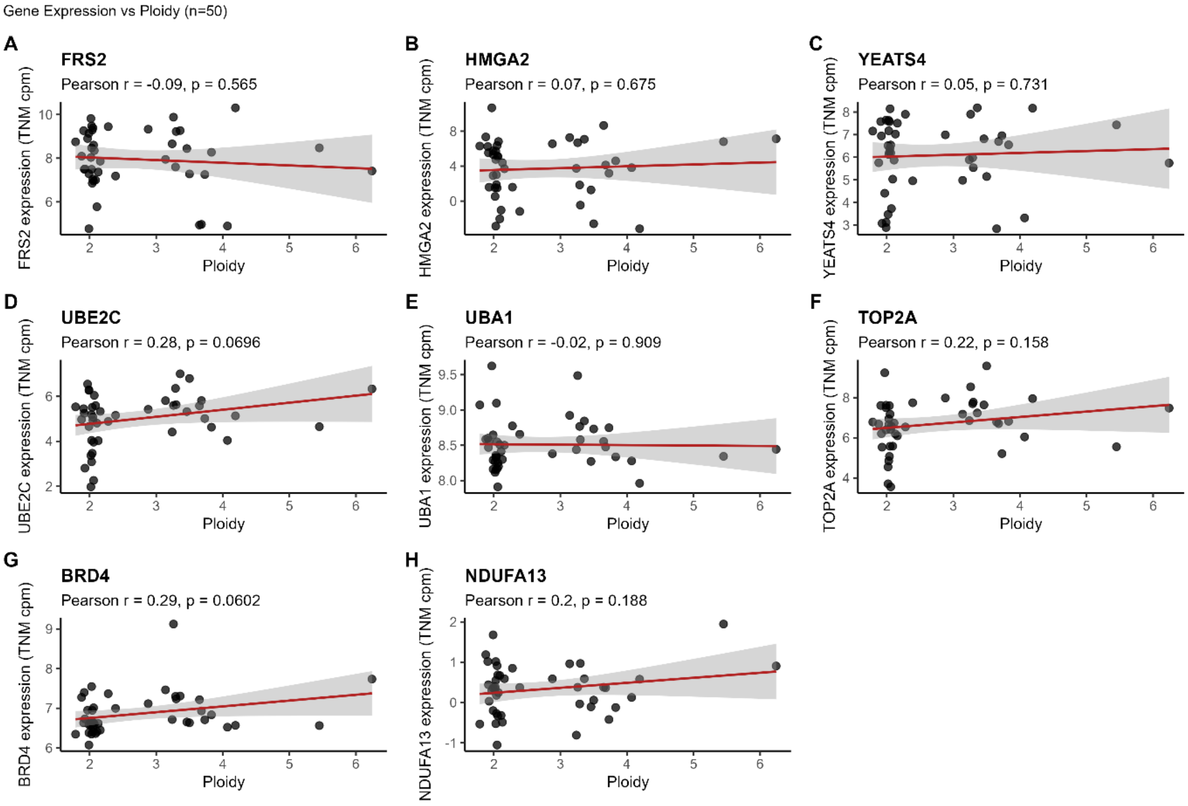


Figure 6.18: Pearson correlation of tumour ploidy and gene expression for key genes in the original TCGA DDLPS data (n = 50). ***A***- FRS2 (Fibroblast Growth Factor Receptor Substrate 2), ***B*** - HMGA2 (High Mobility Group AT-Hook 2), ***C - YEATS4*** (YEATS Domain Containing 4), ***D - UBE2C*** (Ubiquitin Conjugating Enzyme E2 C), ***E - UBA1*** (Ubiquitin Like Modifier Activating Enzyme 1), ***F- TOP2A*** (DNA Topoisomerase II Alpha), ***G - BRD4*** (Bromodomain Containing 4), and ***H - NDUFA13*** (NADH: Ubiquinone Oxidoreductase Subunit A13).

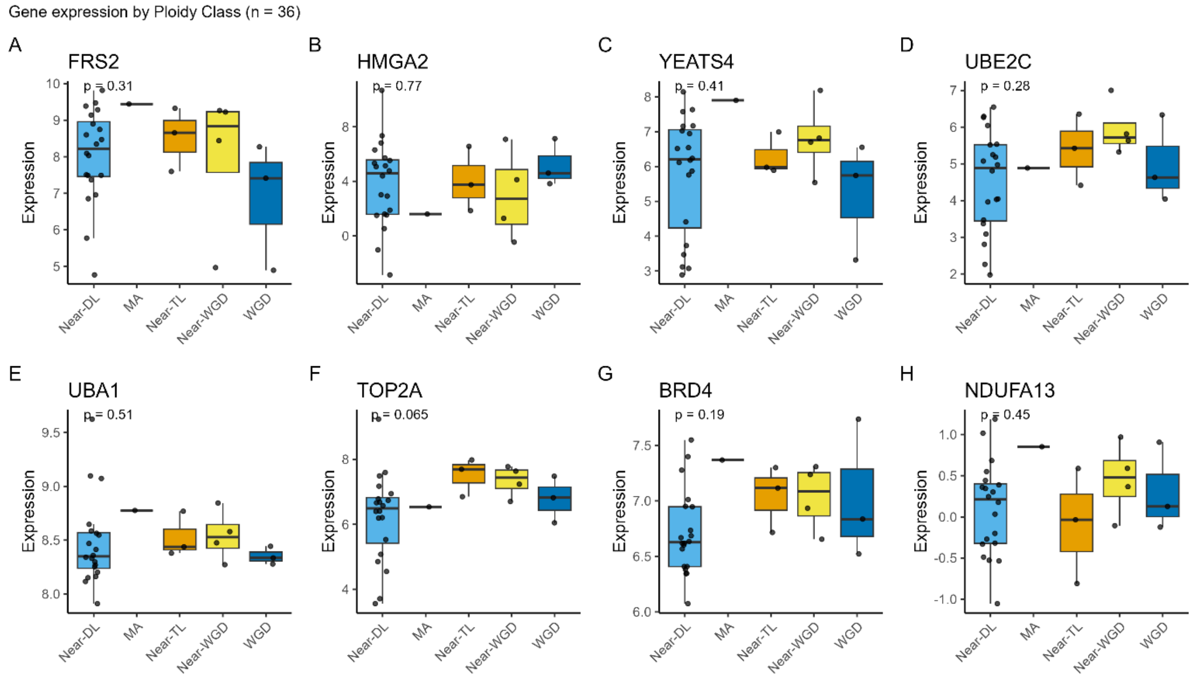


Figure 6.19: Gene expression of key genes separated by tumour ploidy class in the filtered TCGA DDLPS data (n =36). ***A***- FRS2 (Fibroblast Growth Factor Receptor Substrate 2), ***B*** -HMGA2 (High Mobility Group AT-Hook 2), ***C - *** *YEATS4* (YEATS Domain Containing 4), ***D - *** *UBE2C* (Ubiquitin Conjugating Enzyme E2 C), ***E - *** *UBA1* (Ubiquitin Like Modifier Activating Enzyme 1), ***F- *** *TOP2A* (DNA Topoisomerase II Alpha), ***G - *** *BRD4* (Bromodomain Containing 4), and ***H - *** *NDUFA13* (NADH: Ubiquinone Oxidoreductase Subunit A13). Near-DL – near diploid, MA – moderate aneuploid, Near-TL – near triploid, Near-WGD – near whole genome duplication, WGD – whole genome duplication. Statistical significance was tested using a Kruskal-Wallis test.

Upon inspection in the original TCGA DDLPS (n = 50) data it was noted that several genes become significant after categorising the tumour ploidy into classes (***Figure 6.20***). These include *UBA1 (P = 0.033;* ***Figure 6.20E****),* *TOP2A* (*p* = 0.043; ***Figure 6.20F***) and *BRD4* (*p* = 0.012; ***Figure 6.20G***). After further inspection via performing a Wilcoxon rank-sum test and correcting for multiple tests using the Benjamini-Hochberg procedure reveals that only *BRD4* remains significant (*test statistic = 2.95, padj = 0.03*).

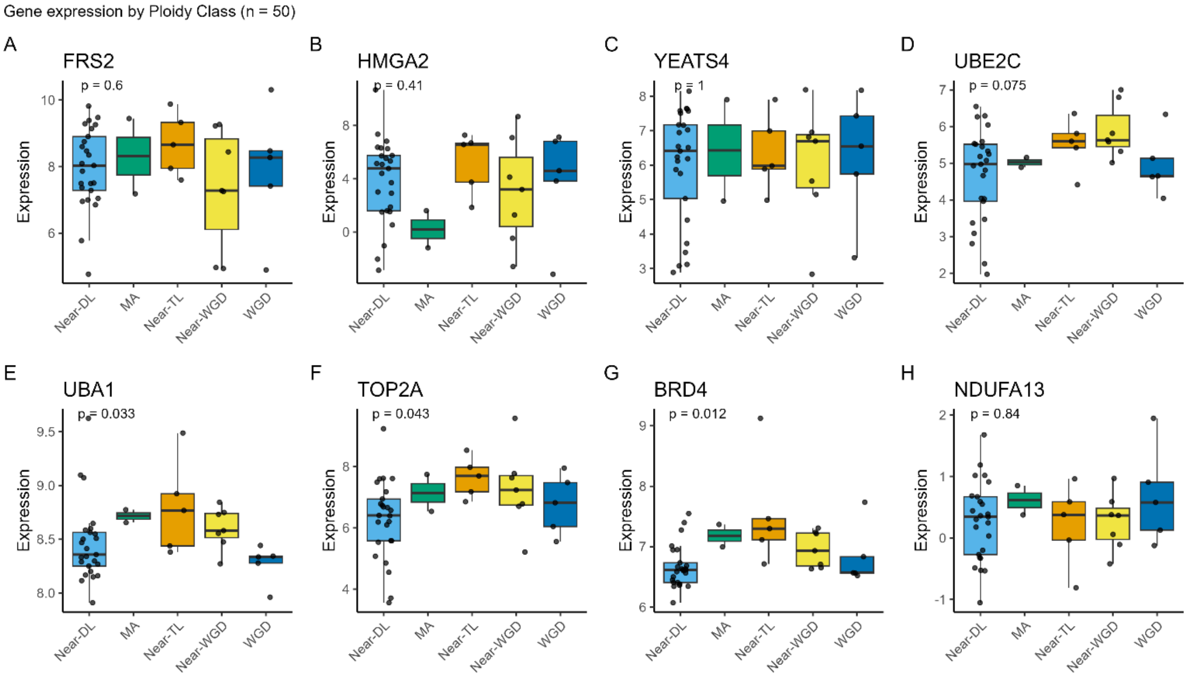


Figure 6.20: Gene expression of key genes separated by tumour ploidy class in the original unfiltered TCGA DDLPS data (n = 50). ***A***- FRS2 (Fibroblast Growth Factor Receptor Substrate 2), ***B*** -HMGA2 (High Mobility Group AT-Hook 2), ***C -*** *YEATS4* (YEATS Domain Containing 4), ***D -*** *UBE2C* (Ubiquitin Conjugating Enzyme E2 C), ***E -*** *UBA1* (Ubiquitin Like Modifier Activating Enzyme 1), ***F-*** *TOP2A* (DNA Topoisomerase II Alpha), ***G -*** *BRD4* (Bromodomain Containing 4), and ***H -*** *NDUFA13* (NADH: Ubiquinone Oxidoreductase Subunit A13). Near-DL – near diploid, MA – moderate aneuploid, Near-TL – near triploid, Near-WGD – near whole genome duplication, WGD – whole genome duplication. Statistical significance was tested using a Kruskal-Wallis test.

6.4.7 CINSARC 67 signatures are enriched in top ranked modules

The CINSARC 67 genes signatures were taken from the full 67 gene panel as denoted in **section 6.3.10**. The module enrichment of the CINSARC 67 signature by pre-ranked GSEA using the kME (modular membership) values of genes concluded that top ranked modules were enriched for CINSARC 67 genes (***Table 6.10***). This was also reproduced when looking at the gene overlap via a hypergeometric test (***Table 6.11***). Eight modules contained at least one CINSARC 67 gene these include M10, M100, M104, M174, M177, M178, M201, and M241. M241 showed the highest number of intersecting genes at 25, comprising nearly 20% of the module and greater than a third of CINSARC 67 genes (***Table 6.11***). No modules ranked via the clinical score GS metric alone contained CINSARC 67 genes.

Table 6.11. Modules with CINSARC 67 gene overlap (overrepresentation).

Module	Intersect (gene symbols)	Number intersecting	Percentage of module (%)	Percentage of CINSARC 67	Log10 (pvalue)
M241	*SPAG5, AURKA, TPX2, BIRC5, KIF4A, CDCA3, TTK, CENPA, CDC20, NEK2, NCAPH, ZWINT, CKS2, CDCA8, ESPL1, KIF11, PLK4, KIF2C, BUB1B, CCNB2, KIF15, BUB1, UBE2C, AURKB, KIFC1*	25	19.69	37.31	**33.118**
M10	*ANLN, FBXO5, KIF20A, TOP2A, KIF23, CEP55, NUF2, CCNA2, MELK, PBK, RRM2*	11	9.91	16.42	**10.095**
M177	*ASPM, TRIP13, ECT2, KIF14, CENPL, KIF18A, CCNB1, CENPE, MAD2L1, CDCA2*	10	22.73	14.93	**12.659**
M100	*MCM2, CDC6, FOXM1, SMC2, MCM7*	5	15.15	7.46	**4.837**
M174	*NDE1, CDC7, RAD51AP1*	4	8.82	4.48	**1.729**
M178	*OIP5, PTTG1*	2	3.92	2.99	0.105
M104	*CHEK1*	1	1.75	1.49	0
M201	*HP1BP3*	1	3.33	1.49	0
Modules are ordered here by the number of overlapping genes (decreasing). Reported p-value is a one-sided Fishers Exact Test (hypergeometric test). **Bold** text indicates a significant result. Pvalues were BH adjusted and transformed to -log10.					

6.5 Discussion

This chapter identified a preserved mitotic cell cycle programme represented in the top ten ranked modules according to the GS measures used. Aside from cell cycle there were also other modules (M225, M23 and M94) enriched for mixed functions with inferred roles in protein trafficking and mitochondrial reactive oxygen species (ROS) production. These top modules were then inspected at the sub-network level revealing hub-genes. After assessing the known PPIs for these hub-genes several inhibitor class drugs and targets were identified. These include most notably, camsirubicin (TOP2A), TAK-243 (UBA1), DSU7K2 (CDC20), NK314 (TOP2A). Of the nine candidate hubs brought forward, two were targeted by inhibitors – TOP2A and FOXM1 giving a hub-target success rate for inhibitors of ~22%. By expanding towards an assessment of the interactors of proteins, two further candidates were identified in TAK-243 targeting UBA1, and DSU7K2 (tosyl-l-arginine methyl ester - TAME) targeting CDC20. A similar success rate was reported in a study in cervical cancer[506] where TOP2A, MERK, KIF11, TTK, PBK, MELK were found to be hubs that have available drugs. Here drugs that were either pre-clinical or already in existing clinical trials/approved were considered.

The most promising hub-gene of those identified with drug-targets was UBA1 through protein-protein interaction with UBE2C the eigenhub of M241 the top-ranked module. It is estimated that UBA1 primes ~99% of cellular ubiquitin for many of the E2 ligases available.[507] It was noted that cancers responded to bortezomib, a first in class proteome inhibitor, which has also had indications for DDLPS along with other proteasomal inhibitors, along with of course the well-covered targeting of MDM2.[508-510] TAK-243 forms TAK-243-ubiquitin adducts that potently inhibit UBA1 activity, leading to defective protein turnover and signalling, cell cycle progression, DNA repair, proteotoxicity due to ER stress and cancer cell death.[511-516] One mechanism by which TAK-243 has been shown to induce cancer cell death in pre-clinical models of solid tumours is through ER stress where an accumulation of misfolded/damaged proteins can occur.[511,513,514]

The main benefit of targeting the E1 enzyme is the functional redundancy observed among the numerous E2 and E3 enzymes.[517,518] Thus the activity of multiple proteins can be directly modified which likely explains the high potency of TAK-243 in reducing the functionality of the ubiquitin-ligase system. As stated, TAK-243 is currently undergoing clinical trials for acute myeloid leukaemia (AML) (ClinicalTrials.gov, ClinicalTrials.gov, NCT03816319) and advanced or metastatic solid tumours (NCT06223542). The results of these are not to be expected until circa

2027. The ubiquitin proteasome system is involved in numerous cancer pathways including cell cycle progression and apoptosis.[519]

UBA1, the hub gene UBE2C and CDC20 are all involved in the ubiquitin-proteasome system, which is a protein degradation system responsible for tagging >80% of cellular proteins for degradation via the attachment of Ubiquitin in a three step system.[520]. The E1 ubiquitin-activating enzymes, including UBA1, via hydrolysis of adenosine tri-phosphate (ATP) attaches Ubiquitin to an active site cysteine residue within the E1 active site, and after structural changes transfers it to the cysteine residue in E2 ubiquitin-conjugating enzyme (e.g., UBE2C) active site.[521] The E3 ubiquitin-ligating enzymes then attaches the ubiquitin to a substrate, this process may repeat until the protein is polyubiquitinated. There are three major classes of E3 ligases, the Really Interesting New Gene (RING), RING-between-RING (RBR), and homologous to E6AP C-terminus (HECT).[518] The anaphase promoting complex/cyclosome (APC/C) is a highly conserved multi-protein RING E3 ligase involved in the degradation of a number of proteins, including cyclins and securing proteins crucial for cell cycle regulation.[522] APC/C co-activating proteins, cell division cycle 20 (CDC20) and CDC20 homolog 1 (CDH1 – encoded by the FZR1 gene) aid in the recruitment of substrates and conformational changes to allow binding of ubiquitin bound UBE2C.[523,524] The substrate is then monoubiquitinated or multi-ubiquitinated at multiple sites. UBE2S is then responsible for chain elongation forming the polyubiquitin.[524]

In late-metaphase and anaphase APC/C-CDC20 targets Cyclin B1 and Securin (PTTG1) to progress the cell cycle into anaphase, and move towards mitotic-exit and completion of the cell cycle.[525,526] Securin sequesters separase, a protease essential for the disassembly of the cohsein complex and sister chromatid segregation. Cyclin B1 partners with CDK1 during the G2/M transition and into metaphase, where levels of cyclin B1 drop into anaphase, and is crucial for G2/M transition.[527] To ensure sufficient levels of cyclin B1 for G2/M transition, it is protected from the APC/C by the mitotic checkpoint complex (MCC).[527] Upon APC/C activation via release from the MCC, the APC/C can tag Cyclin B1 and securin. The APC/C then associates with CDH1 in late anaphase allowing for the degradation of CDC20 and range of mitotic kinases (e.g., PLK1, AUKRA/B) allowing for exit.[528]

The overexpression of UBE2C and CDC20 along with UBE2S observed in a range of cancers associated with a poor prognosis[529-533]. It is thought that this leads to chromosomal mis-segregation leading to genomic instability, aneuploidy, decreased apoptosis and increase cell proliferation.[534-536] There has been several attempts to target the human proteasome in cancer.[537] Bortezomib (velcade) is an FDA approved first-generation proteasome inhibitor targeting the function of the 26S proteasome with indications in myeloma.[519] It has been shown to promote cellular arrest which may be through the accumulation of cyclin B1 levels and a stalling at the spindle-assembly checkpoint.[525] Bortezomib and similar FDA approved proteasomal inhibitors (e.g., carlfilzomib) has shown promise in DDLPS in vitro patient-derived and mouse models in inhibiting tumour growth, and decreasing MDM2 expression levels.[538-540] Furthermore, bortezomib, ixazomib (a 20S proteasome inhibitor) and carlfilzomib potentiated the effect of nutilin, an MDM2 inhibitor in WDLPS cell lines.[541] Other inhibitors, including bendamustine, have sought to target the E3 ligase machinery.[519]

Upcoming proteasome inhibitors in clinical trials have sought to target the E1 ligases including the first-in-class TAK-243 (MLN7243) which targets UBA1. TAK-243 works by binding free ubiquitin, blocking UBA1 activation, leading both to protein accumulation (and proteotoxic effects due to the unfolded protein response) and dysregulation of downstream E2 and E3 functions, including the APC/C.[507] TAK-243 is currently in phase I clinical trial for leukemia (NCT03816319), although was previously the subject of a terminated trial for advanced solid tumours (NCT02045095). TAK-243 has shown repeated anti-cancer effects in cancer cell lines and patient-derived xenograft mouse models in lung, leukaemia, pancreatic cancer and adrenocortical carcinoma.[512-515] TAK-243 may represent a promising inhibitor for use in DDLPS.

A notable challenge was the decision between selecting genes for sub-networking analysis from modules. Typically, genes are selected using the kME at a given threshold for that module, this is generally kME > 0.7.[234,251] However, for highly correlated modules, kME shows a high level of overlap, where at times a given gene can show high module membership for both the module it was partitioned into and modules where there was high co-expression with another modules ME. Whilst this is certainly a beneficial aspect of WGCNA, as it allows the retrieval of additional co-expression relationships outside of module gene assignment, it also can complicate subnetwork analysis aiming to identify hub genes, as often similar modules will contain genes that are high in kME for other MEs. To circumvent this, instead of using a kME cut-off to select genes, gene module assignment was used to filter out genes. The contextual benefit here was to identify additional hubs (and candidates) that were exclusive for a given module. The caveats are that for modules with a low intramodular connectivity, the network generated will be sparse, furthermore, here it was typically found that in order to retain co-expression edges, the TOM had to be set low, which decreases the confidence in co-expression edges.[196,235,347]

To select for hubs both the degree and the eigencentrality were inspected. Ultimately, eigencentrality was used as the measure to select a candidate as outlined in section 6.1. However, there are many other centrality indices that could be used. For instance, a repeat of these analysis using the betweenness, PageRank or average degree density indices may retrieve different results. Although, this is highly contextual and dependent on the topological structure of the network. For example, a network where there are dissimilarities in clustering coefficients among nodes and distinct clusters (i.e., it can be observed that there are two separate/bridged network structures) may yield drastically different results if using the betweenness centrality as there are likely conduits between clusters.

The CINSARC 67 is a well-known and utilised prognostic gene signature and measure of chromosomal instability for sarcomas including DDLPS although, it has been applied to other neoplasms.[499-502] CINSARC 67 genes were mostly enriched in top ranked modules notably M241 and M10 which were both potently enriched for mitotic and chromosomal segregation genes. CINSARC 67 genes are known to represent regulatory genes that are key in mitotic processes that when overexpressed can lead to dysregulated chromosomal segregation and associated with poorer outcome.[499-501] The concordance observed here provides further credence that hub genes (and targets) are important molecules in DDLPS cancer biology.

Despite the strong CINSARC 67 score, it was found that the gene expression of key genes identified in this chapter (*YEATS4*, *UBE2C*, *UBA1*, *TOP2A*, *BRD4*, *NDUFA13*) and known drivers (*MDM2, CDK4, YEATS4,* and *FRS2*) did not show a significant relationship with increasing ploidy. Many of these genes, notably *UBE2C* and *TOP2A* are included in the CINSARC 67.[500,502] This suggests that for the majority of tumour samples tested here, gene expression variation is likely driven by focal genomic (e.g., the characteristic amplifications of 12q13-15) or regulatory changes. Work on the CINSARC 67 signatures would suggest that increasing genomic instability leads to a more aggressive tumour and vulnerabilities due to increased stress on the tumour cel.[502] Indeed, chromosomal instability can differentially impact sensitivity to targeted therapy, where WGD have been shown to make colon cancer cells more sensitive to proteasome inhibitors.[542,543]

To summarise, in this chapter a network-based approach is leveraged to identify putative drug targets. Multiple hub-genes were identified across the top ten modules, although not all were targetable. Including PPIN data allowed for interactors to be identified as targets. Notably, UBA1 and the drug TAK-243 should undergo further investigation as novel target in DDLPS. This acts as a demonstrable proof-of-concept for network-based analysis in the identification of novel drug targets.

Chapter 7 Final Discussion

Dedifferentiated liposarcoma (DDLPS) is a poor prognostic subtype of liposarcoma (LPS), which are part of a diverse range of rare mesenchymal malignancies called soft tissue sarcoma (STS).[1] In general, STS malignancies have limited treatment options beyond surgical resection.[60,71,544] Chemotherapeutic drugs, typically employed for advanced/metastatic disease, are largely ineffective in all treatment settings and show high levels of toxicity.[545] Doxorubicin, the most common first-line anthracycline based agent used across STS cancers, has been employed since 1973 (adriamycin) after notable tumour regression in STS.[545,546] Treatment options available to patients have not drastically changed, and there is a desperate need for additional therapeutic interventions.

In the past decade several drug indications have been made targeting molecular alterations in LPS, most prominently *MDM2* and *CDK4* amplifications.[508,547] Inhibitors targeting these have shown mixed results in clinical trials (see **section 1.2.6**). The MDM2 inhibitor milademetan was the most promising but failed to meet clinical trial endpoints.[71,160,548] A new favourite, Brigimadlin, is currently undergoing a phase 2/3 trial (NCT05218499).[167] However, previous results of clinical trials targeting somatic copy number alterations (SCNA)) and immune checkpoint blockade (ICB) indicate that this will only work in a subset of patients.

Targeting prevalent somatic mutations under the assumption that such alterations are oncogenic drivers and are distinguished from unaltered genes in healthy tissue.[549] DDLPS has a low mutational burden and hence molecular targets based on somatic mutations, outside of SCNAs, are few.[64] Cancer vulnerabilities can go beyond oncogenic driver mutations where non-mutated genes can also be vulnerabilities to be targeted.[313,550]

Network analysis is a powerful tool, that fundamentally recapitulates molecular circuitry via topological patterns of connectivity between entities as a network graph.[192,551,552] Gene co-expression is the similarity between gene expression patterns, usually at the scale of hundreds or thousands of gene transcripts.[228] Gene co-expression networks are an application of graph theory to analyse gene co-expression data.[234] GCNs can then be used for numerous applications including biomarker discovery and as an effective gene screening method for analysing whole transcriptomic (global) data to select interesting groups of genes for further analysis out of the thousands of genes used as input. WGCNA is by far the most common tool used for such purposes with well documented applicability and has simple yet effective data integration tools.[234,251,256]

Indeed, varying applications of WGCNA, with adjunct bioinformatic/biostatistical analysis exist across multiple cancer types for the purpose of highlighting new and exciting genes.[186,243,248,259,265,308,553,554] Hence it was identified here that this could be a tool to screen out interesting genes in DDLPS for further investigation as putative drug targets.

The main aim of this project was to identify and implicate putative targets through use of an integrated network approach using DDLPS gene expression data. **Section 3** saw the identification of gene expression datasets and the construction of a multivariate Cox Proportional Hazard model ready for WGCNA clinical trait data analysis.[64] Subsequently in **section 4** WGCNA parameters and in-depth quality control was conducted to build a robust weighted GCN. In **Section 5** the three GS measures were used to rank the modules in the weighted GCN and subsequently a comprehensive characterisation of the modules followed along with the integration of single-cell data, identifying several novel findings. Then in **section 6** sub-graph analysis was applied to the top-ranked modules, and a hub gene *UBE2C* was identified, where protein-protein interactions implicated TAK-243 as a drug targeting *UBA1* the enzyme upstream of *UBE2C***.** The goal of this project was met.

This chapter will summarise the pivotal findings, provide an overall interpretation of results, and their wider implications. Furthermore, the limitations of this project will be discussed, and future directions will be highlighted for further research to be conducted.

7.1 Project overview

7.1.1 Building a robust DDLPS GCN that reconstructs DDLPS biological pathways

The first aim of this project was to construct a robust GCN using WGCNA.[234] This was achieved in **sections 3** and **4** which involved the identification of secondary data sources for WGCNA and the construction of a robust weighted GCN.[64,114,234,332,333] The GCN derived is not only a resource that can be later mined for candidate targets, but also better understand disease mechanisms by highlighting functional pathways. To ensure a robust network was built, several tests were conducted to choose optimal parameters. This included an assessment of the distribution of gene connectivity, the correlation of intramodular connectivity (IMC) to GS measures, and comparisons to randomly generated graphs. Each of these assessments were successful and indicated a robust GCN. Subsequently the GCN also showed strong co-expression preservation in a validation RNA-seq dataset.[114]

Testing multiple parameters marks a shift from using default settings which are commonly adopted. Other studies have also found that WGCNA has favourable performance against other methods with only minor modifications required to parameters.[555] Typically studies have relied on using clinical variables as a measure of GS to identify important modules.[238-240,306] Whilst this is highly beneficial for identifying modules that are correlated with clinical variables (e.g., outcome) and is directly interpretable and contextual to clinical behaviour of the disease, it does not detail mechanistic insights. Hence, in addition two other data sources were integrated which was differential gene expression data (GSE159659 – DDLPS versus WDLPS) and gene dependency data (DepMap – LPS141, LPS853 and LPS510 cell lines).[313,333] The former described dysregulated genes and the latter gene essentiality. Modules correlating highly to these GS features may represent pathways that are clinically relevant, dysregulated and essential were taken together may represent robust targetable cancer vulnerabilities.

The GS measures were also used to inform on the most optimal GCN partition. According to WGCNA, modules with highest relevance should have modules where intramodular connectivity correlates with measures of GS.[196,234,251] Hence, the partition that best emphasised these correlations was chosen. Modules with high IMC vs GS are interesting, and according to assumptions of WGCNA graph theory, are important to underlying biological processes.[234]

It is common in cancer studies to pair WGCNA with differential gene expression analysis (DGEA).[244,246,248,259,309,312,556] This is done either in a pre-or-post WGCNA manner, although identifying DEGs after WGCNA module identification, and using some of the summary metrics available in WGCNA is regarded as good practice.[259] Pre-filtering the number of transcripts too strictly can invalidate the scale-free assumptions of a GCN.[234] In this project, differential gene expression was integrated through correlation of the IMC to the signed significance of differential expression (signed -log10 FDR adjusted p-value) which follows guidelines for clinical trait data analysis.[234] This is also similar to applications of the gene essentiality data, although this is much less common in WGCNA studies.[248] Finally, the GCN constructed here also showed strong module preservation which is a robustness test not yet conducted for WGCNA in DDLPS.[252]

The relative performance of WGCNA against other methods has been noted.[555] Whilst the benefits of adopting a WGCNA approach have been discussed in depth (see **sections 1.3.3, 3.1, 4.1, 4.5, and 5.1**) along with studies detailing the robustness of WGCNA across datasets (with correct tuning as was made evident in the NCC data).[555] There is a wide suite of module detection methods available, and whilst it is noted that some outperform WGCNA for module detection, they lack the contexture of graph-based theory applied to co-expression data.[555]

7.1.2 Identifying and characterising modules of interest

To identify modules that show high biological relevance to the disease, the modules were ranked according to the summed correlations of IMC to all three GS measures. The same measures used to identify optimal parameters. This novel approach in DDLPS incorporates information from several sources of disease data and streamlines downstream analysis. Modules with the highest correlation, the top-ranking modules, may describe essential and dysregulated cancer pathways that correspond to tumour progression and therefore may contain targets. This is why studies primarily focus on modules with the highest positive correlation to GS measures as they are likely most relevant to the research question.[557,558] However, modules with negative correlations to GS may also provide important information, likely representing tumour-protective modules and immune enrichments.[313] Furthermore the context of the DEG GS score (WDLPS vs DDLPS for matched tumours) used implies that modules lowest ranking also correspond to WDLPS-like portions of the DDLPS tumour. Both the top and bottom rankings were inspected individually where negative correlations were assessed in **section 5.4.3** and positive correlations in **section 6.4**.

The DDLPS GCN was found to be enriched for cancer hallmarks and immune enrichments. The GCN approach was able to reconstruct transcriptomic programmes through unsupervised clustering of gene co-expression data. Perhaps the best demonstration of this was the correspondence between unsupervised ME cluster analysis (based on gene connectivity and modularity gain), GSEA, single cell enrichment, and a clear extraction of known transcriptional programmes (e.g., lipid metabolism, cell cycle and interferon signalling) in DDLPS. Top ranked modules were strongly enriched for cell cycle processes. Abrogated cell cycle is a well-known cancer mechanism and commonly targeted in cancer with multiple drug agents available.[559]

The results from **Section 5.4.1** paired with the favourable quality of the GCN in **section 4.4.2.2** gives a strong indication that transcriptional patterns correspond to DDLPS biology as opposed to random connections or noise. Furthermore, many of these transcriptional patterns correspond to cytological features of DDLPS and WDLPS as discussed in **section 5.5**.[44]

7.1.3 Complex interactions in the TME may produce a chronic tumour promoting environment for DDLPS

Inspecting low-ranking modules revealed four main observations:

1. Gene co-expression signatures for Interferon responses, antigen presentation, tissue fibrosis, and vasodilation suggest a chronically inflamed environment.
2. Several tumour-associated processes were evidenced in vascular cells inferred by gene expression.
3. Immune signalling through interferon and interleukins corresponds to epigenetic heterogeneity in DDLPS the GCN and by inference the TME.
4. Modules enriched for lipid metabolism correlates to modules with genes expressed in single-cell tumour clusters expressing mesenchymal stem cell like markers.

Chronic inflammation is known to be associated with cancer and immune profiling in DDLPS has revealed inflamed (or immune hot) and non-inflamed groups.[560-562] Inflammation can lead to tissue fibrosis which further amplifies inflammation through the release of inflammatory cytokines including TGF-β1 and IL-6.[563] Inflammation within the TME can induce ECM remodelling through the activation of fibroblasts. Notably, subsets of tumour cells display a fibroblast-like phenotype. Whether this represents a population of cancer-associated fibroblasts (CAFs) in DDLPS or a characteristic of DDLPS tumour cells is not yet confirmed. Undifferentiated mesenchymal malignancies display spindle-cell morphology and are phenotypically similar to CAFs.[564,565] Perhaps due to related cellular lineage and notable similarities between fibroblasts and mesenchymal stem cells.[566-569]

Vascular mimicry is a process that has been noted to occur in WDLPS (pericyte mimicry), and was noted in a case study for DDLPS.[430,570] Other sarcomas also show potential for vascular mimicry and the ability to line and construct new vasculature with abnormal vascular cells.[571] Perhaps vascular mimicry would explain the observed correlation between tumour cell profile enriched modules with angiogenic modules. The signature observed here could also be endothelial-to-mesenchymal transition (EndoMT) where ECs loose classical markers and take on a more mesenchymal like profile, being also able to mimic fibroblasts in the presence of TGF-β.[572] It may be both processes occurring in a cooperative and coordinated manner which has been termed mutual mimicry.[573] This source of angiogenesis in the tumour may explain the low levels of correlation between angiogenesis and hypoxia. This may have implications for treatment as mutual mimicry is associated with targeted drug resistance across multiple mechanisms and pathways.[572,573] vascular mimicry has also been able to predict a poor outcome in osteosarcoma.[574]

Lipid metabolism impairs T cell function against the tumour through excessive uptake of lipids and cholesterol and is associated with the resistance to therapy and supporting tumour growth.[575,576] With cholesterol export signatures identified, it is possible that adipogenic regions of the DDLPS tumour export cholesterol which accumulates and sustains dedifferentiated regions. Although, this required additional research to confirm and would be a research question with direct therapeutic implications such as cholesterol biosynthesis/efflux can be targeted. It has already been postulated that statins could be useful for sarcomas and has been associated with enabling differentiation in osteosarcomas and EWS.[577] Furthermore, statins have been shown to promote cell cycle arrest.[578] The role of lipid metabolism in sarcomas has been underreported, although a recent study has explored this. Knockdown of SQLE has been noted to suppress cellular proliferation and induce cell apoptosis in sarcoma A-673 and U2OS cell lines from EWS and osteosarcoma, respectively.[579]

7.1.4 Identification of TAK-243 as a novel drug for DDLPS

Finally in **Section 6.4.5** an ML RWR sub-graph approach using both TCGA DDLPS and NCC DDLPS gene co-expression data was used to find hub genes.[282] Following the assumptions and theory of network biology, these hub genes were hypothesised to represent important genes in DDLPS cancer biology, and according to the module screening strategy adopted (**section 5.4.2**) pertinent to DDLPS proliferation and survival.[200,202,228,234,317,580] Where the primary result was the identification of *UBE2C* as an eigen-hub in the top ranked module M241 with integrated PPI and DTI analysis revealing *UBA1* as a target of *TAK-243*. The effectiveness of targeting the proteasome has already been demonstrated in DDLPS.[510] The identification of TAK-243 and UBA1 has wider implications and hence is discussed further in the following section.

7.2 Combined interpretation and wider implications of results

The results emphasise the power of a systems approach in mining biological data and act as a framework for future work to be integrated and gain new insights. Several biological process and pathways that may correspond to DDLPS disease mechanisms have been identified in the GCN corresponding to multiple cell types. These include lipid metabolism, chronic inflammation, fibrosis, stem-cell like signatures and cell cycle. The GCN represents a reconstruction of the transcriptome of not just tumour cells but also the various cells across the tumour stroma that is captured during sampling and sequencing.

The underlying gene connectivity within a GCN being purely correlation based with no inference on causality. This is true for both the gene connectivity and the correlation of MEs to GS measures. However, it is assumed that observed patterns of co-expression are the measurable consequence (be it directly or indirectly) of the shared induction of gene expression as part of a biological pathway where genes are influenced by each other or a shared mechanism.[581] Here, despite the lack of causal relationships, several gene pathways were evident at the gene level.

Similar studies have applied gene dependency, differential gene expression, and prognostic filtering to identify interesting co-expression modules. However, these mostly conducted as filtering steps deployed prior to or adjacent to WGCNA, rather than being used as a GS measure.[248,582-584] It is generally accepted that pre-filtering based on differential expression can hinder WGCNA results and performance.[259] The approach taken here differs slightly as module screening is performed through an aggregated GS metric rather than a singular feature-based selection.

Subsequent subnetwork analysis then revealed relationships among genes that recapitulate known biological processes – perhaps most notable in cell cycle modules where the APC/C was modelled (**Section 6.4.3-6.4.5**). Subnetwork analysis is commonly incorporated in network analysis, most frequently through the construction of module PPINs using interactome data.[582,583]

The primary result of the project that satisfied the projects final aim was the identification of UBA1 ubiquitin pathway as a targetable pathway through TAK-243.[516] This was implicated through the observed importance of the APC/C in cell cycle modules through *UBE2C*. Cell cycle has become a commonly targeted pathway pan-cancer.[559] This is partly due to the relatively good understanding of cell cycle machinery and its dysregulation, and has been modulated by drugs for decades (e.g., etoposide, doxorubicin, etc), and still remains an area of much interest.[585,586] A challenge presented for cell cycle targets is the heterogenous dysregulation of cell cycle proteins in the tumour, and off-target effects in normal tissues that may also show rapid proliferation, necessitating new delivery mechanisms.[587]

The result achieved here was data driven. However cell cycle was predicted to be enriched in top ranking modules as it was understood that differential expression and DepMap data would likely highlight cell cycle as a dysregulated and dependent pathway.[313] However, the drug interaction identified is not strictly a discovery pertaining to the cell cycle but to the ubiquitin-proteasome system (UbiPS) and seeks to more precisely target the increased demand on cancer cells to combat a build-up of unfolded proteins.[519] This result taken with other results indicates some relationships among the various biological processes extracted from the DDLPS GCN.

Among the pathways that ubiquitination regulates are lipid metabolism, immune processes and cell cycle.[576,588] Our results identified *UBE2C* and through PPI data UBA1, the former was clustered in a cell cycle annotated top-ranking module. As mentioned previously, UBA1 is the master protein for priming the majority of cellular ubiquitin subsequently used to degrade proteins involved in a range of biological processes.[589]

UBA1 expression levels has been correlated to low levels of CD8+ T cell infiltration and interferon signalling and could predict a poor response to ICB.[590] It was also noted that UBA1 depletion could inhibit tumour growth which could be rescued by depletion of CD4+ and CD8+ T cells.[590] Notably, TAK-243 or other proteasomal inhibitors can synergise with ICB therapy increasing response rates.[590-593] Furthermore, some E3 ligases (including MDM2) are known to be involved in chronic inflammation, along with ECM and vascular remodelling.[594] It is possible that DDLPS leverages UbiPS to create an immune suppressive environment in DDLPS. DDLPS is known to show diminished response to ICBs but may show improved response rates in combinatorial therapies using selective UBA1 inhibitors such as TAK-243.[595,596]

Ubiquitination has also been found to be important in regulation and reprogramming lipid metabolism in cancer.[576,597] This may correspond to a suppression of lipogenic programmes, as E3 ligases are responsible for degrading key lipogenic enzymes and transcription factors including HMGCR, SREBP1 and PPARγ.[576,597] Additionally, where there is excess cholesterol, the ubiquitination of efflux proteins is reduced to maintain cholesterol transport out of the intracellular environment.[588] It is hypothesised here that cholesterol/lipid efflux from lipogenic components may contribute to promoting a tumour permissive TME for DDLPS cells. Strategies combining proteasomal inhibition with lipid lowering drugs to counteract the secretion of unfolded proteins has been proposed in multiple myeloma.[598] Additionally use of proteasomal inhibitors has been shown to improve cholesterol homeostasis through promoting reverse cholesterol transport.[599,600]

It has been found that TAK-243 is a substrate of ABCB1 which can decrease the cytotoxic effect of TAK-243 through transportation out of the cell.[601] Together with the hypothesis that the WD-like components may be supplying the DDLPS cells with lipids may implicate a potential combinatorial intervention. Statins inhibit ABC transporters, which may not only serve to eradicate ABC-mediated drug resistance to TAK-243 but also starve the tumour of lipid resources.[602] Statins have been tested on DDLPS cell models but are only effective in cells with lower (relative to typical levels in DDLPS) levels of MDM2 and was not effective in higher levels of MDM2, likely due to composition changes within lipidomic profiles.[603] However, these cell models do not model the TME and its role in DDLPS which also does not take into account the biphasic nature of DDLPS.

In this project an *apriori* bioinformatic approach was taken with no filtering based on known driver mutations in DDLPS tumorigenesis. The result of this was evident where classical DDLPS markers were not prioritised in the module screening. Furthermore, *MDM2* was partitioned into M82 which was the grey module due to the variance of *MDM2* across samples being low. The implication of this is that these classical markers play a less important role in transcriptional networks and regulation than may have been assumed.[234] Here genes with somatic driver mutations were not identified through the use of network analysis. Instead, the inference is that coordinated dysregulation causes a rewiring of modular gene programmes that can also associate significantly with poor clinical course and cancer cell survival. This is likely the single most beneficial advantage of the systems approach used here as not only does it highlight novel findings, but it is also inclusive of the TME.

7.3 Limitations

There are several limitations pertaining to this project. The most obvious are to do with the sample number, the reliance on correlations and following the assumptions of network biology. Starting with sample number, not only is this a key limitation for the WGCNA technique used here and would likely prevent good performance for more advanced machine-learning approaches.[264,283,290,317,604] Furthermore, larger sample pools (approaching 200 for correlations to stabilise) would allow for exploration of patient specific networks, to overcome the limitation that WGCNA is an aggregate network of co-expression patterns where sample-specific co-expression networks would be able to delineate heterogenous gene interactions.[359,605]

This project also hinges on the assumptions that biological networks are indeed scale-free, that hubs are important, and network topology reflects hierarchical molecule interactions. Fundamentally, there is still much debate on the scale-free nature of biological networks.[206,389,606,607] Large studies of networks find that only 4% strongly fit a scale-free distribution.[389] Much of the early studies that supported scale-free networks in biology were based on PPINs from yeast being eventually applied to human systems.[208,218,219,608,609] PPINs were also the basis for central lethality, which is a compelling argument for the importance of hub-genes.[208,209,610] Whilst the hub gene hypothesis is well established and valid, it is, in the context of scale-free networks still a theory. Scale-free networks hinges on preferential attachment describing the propensity of newly added nodes to, according to probability distribution, attach to already highly connected nodes.[205] This tendency becomes less apparent at smaller scales despite power-law degree distributions being fairly scalable across network sizes.[205] Although it is noted that biological networks do, more often than others, show a stronger scale-free degree distribution.[389] In **section 4.4** it was ensured that the DDLPS GCN approximated a scale-free topology, removing outlier samples to optimise the scale-free distribution of gene connectivity.

Furthermore, gene co-expression networks are founded on measures of correlation, along with much of the post-network analysis (e.g., module correlations).[234] Providing that patterns in values are similarly adjusted then there will be a correlation which has varying interpretations.[611] Furthermore, correlations does not equate causation. Another limitation with co-expression networks is that it is, using co-expression data alone, not possible to distinguish regulators from regulated elements.[270] This then links back to the argument that hub-genes in co-expression networks are key regulators, or are regulated by many genes. Using co-expression data alone and without direct biological context or other information, it is not possible to provide a definitive answer. Despite this, the WGCNA approach utilised here was able to identify patterns of co-expression that align with biological context.

A debate in the study of DDLPS, and other dedifferentiated sarcomas, is the existence of a so-called low-grade dedifferentiated subtype (low-grade DDLPS).[4,43] The dedifferentiated histological elements exhibited lower cellularity compared to typical DDLPS and bore a resemblance to well-differentiated fibrosarcoma.[4] It was noted to have markedly higher cellularity versus sclerosing WDLPS, but a lower mitotic rate that is typically observed in DDLPS (≥5 mitosis per 10 High Power Fields - HPF).[43,612] However, later studies demonstrated that low-grade DDLPS showed survival trends that were more in-line with WDLPS and that the mitotic criteria for DDLPS were crucial.[612,613] These studies recommended the assignment of DDLPS cases with ≤5 mitosis/10 HPF as cellular WDLPS. As pointed out by both Kilpatrick et al[614] and Dry et al[43] this has led to a lack of consensus in defining DDLPS in studies.

The data collected by the TCGA SARC on DDLPS, which underwent pathologist review, did not use the criteria and nearly half of the DDLPS samples have a mitotic rate of <5. From some perspectives, these would be classified as cellular WDLPS and not DDLPS, yet from a transcriptomic perspective, these samples are similar. Therefore, underlying molecular programmes are DDLPS-like. Stark differences in the transcriptomic profiles would have shown discrete clustering, correspond to the categorisation of mitotic rate during the sample quality checks in **section 4.4**. This was not the case.[43]

7.4 Future directions

The future directions of the project are discussed in this follow sub-section. First and foremost, the most immediate future direction is to be the *in vitro* testing of the top candidate drug, TAK-243, on DDLPS cell lines. Laboratory validation is an important step in giving evidence of the drugs ability to inhibit and kill cancer cells.[248,615,616] In addition to UBA1/UBE2C and TAK-243, there were also other hub-genes identified in **section 6.4.5**. Further exploration of these is warranted.

There is also the potential to apply data from available drug screen data to the DDLPS GCN. The DepMap project have conducted and made available drug screens against a single DDLPS cell line (LPS141).[550] This could highlight further indications with the caveat that the results would be based on a single cell line. Use of the Connectivity MAP (CMAP) chemical gene perturbation data could also supplement this, although DDLPS data is not currently available in CMAP.[617] Drug screens have already been conducted for patient-derived organoid models for DDLPS revealing responses for drugs targeting VEGF, histone deacetylase, mTOR, and proteasomal pathways.[510] However, these models do not account for the DDLPS TME and the substantial importance this may play in DDLPS disease biology. Testing drugs, acting against hubs, on robust DDLPS co-cultures that better model the TME could highlight new disease mechanisms and vulnerabilities.[618]

It would be beneficial to explore and demonstrate the effectiveness of targeting hub genes over traditional drivers in DDLPS. This project identified many hub genes, where there was the potential to identify more using different centrality indices (e.g., betweenness), of which only a few were highlighted in this project. Hubs are regarded as central to a network and by inference disease biology.[208] In the context of the DDLPS GCN there are some questions: Are hubs really essential, and do they make better drug targets than targeting oncogenic drivers? What is the level of redundancy in hubs? Is a given hub crucial to module networks and global GCNs?[619]

Future analysis would benefit from the input of additional omics data. Most notably this includes the integration of proteomic data to confirm molecular pathways present within the proteome. It is known that protein levels can vary substantially compared to mRNA levels with Spearman *rho* values ranging from 0.3-0.7.[620-623] This is largely due to translational regulation, modification and protein stability. Proteomic data has recently been made available for analysis. Most recently A large cross STS study including WD/DDLPS (n, 36 and 35 respectively) and MLPS (n, 11) with 5593 mean number of proteins per sample.[624] Other studies have also been published with available data.[625,626] There is sufficient data to now perform in-depth and integrative analysis in future projects with the aim of both confirming molecular signatures at the proteomic level of this projects results, and building new novel resources based on the input data. However, it should be noted that to truly benefit from additional omics layers, results from experiments should be sample matched. Unmatched data presents addition challenges and noise that is difficult to account for.[627-630]

Furthermore, since the conception of the project, additional WGCNA techniques have been developed including multiWGCNA, an approach which seeks to create sample-trait or timeseries specific networks, and compare the changes between them.[631] Although it is noted that 12 samples per trait/group are required as minimum, which would be challenging for DDLPS. Likewise, single-cell GCN analysis methods have also become available, as either modified WGCNA approaches (e.g., hdWGCNA) or similar methodologies.[632] Along with those specifically tailored for single-cell RNA sequencing datasets which are suited to account for data sparsity, where some cells express genes at low or zero levels.[633,634] The method used in this project to assess gene co-expression (COTAN) in scRNA-seq data does not take into account the benefits of applying network-based methodology in its analysis.[428] The main benefit of network-based single-cell GCN approaches would have been cell-type specific networks (perhaps as a method of further delineating EndoMT signatures identified in **section 5.4.3**), although this would be contingent on a sufficient number of cells (some of the test datasets used for hdWGCNA contained over 1 million cells).[632]

Targeted GCN is a recent methodology that deviates from WGCNA to provide trait-specific GCNs that are substantially smaller, and contain less noise.[635] This leverages the use of a pre-processing step where a LASSO regression is used to screen out genes most predictive of a given trait. Input transcript pools used as input are magnitudes smaller resulting in much smaller modules. However, it can be argued that there is a substantial loss in the graph structures observed in WGCNA modules due to this filtering. The authors do highlight that these TGCN modules highlight WGCNA module subsets not too dissimilar from the approach adopted in the project, both for increasing the granularity of the WGCNA approach through parameter optimisation but also using sub-network analysis. The difference being that the approach here is post-WGCNA where as TGCN is pre-clustering where it can be argued that despite the benefits of this approach, the assumptions of a scale-free network have been breached. The benefit of TGCN over WGCNA is in streamlining the process when it is known on what traits are available. For this project the hypothesis free nature of WGCNA is preferred as it was more likely to uncover novel findings.

A potential solution for overcoming sample size limitations could be pooling high-grade undifferentiated/dedifferentiated sarcoma (similar to the TCGA SARC study[64]) and well-differentiated sarcoma samples, to reduce the sample number limitation, with the objective of identifying key molecules applicable across high-grade sarcomas. This is a contradiction to the ethos as set out in projects research statement, although still sits within the scope of addressing treatment needs or unresolved biological understanding of high-grade, poorly performing tumours.

In **section 6** when drug-target data from drug target databases (TTD and CPP) focused on small molecule inhibitors (SMIs) with known drug-target indications that had already achieved FDA approval or were investigational.[319,320] There were three main reasons why SMIs were prioritised. The first being that they are the most common type of drug, and still to date the most frequently approved novel drug type by the FDA, where FDA approval would also decrease the possibility of early drug failure, as is common in cancer.[636] SMIs are mechanistically better understood, both in terms of their chemical and biological activity, easier to synthesise and are lower cost.[636,637] Furthermore, small molecule drugs can, due to their small size, easily target intracellular targets as they can pass permeable membranes.[637,638] Although their small size can limit their coverage of protein-protein interaction surfaces, especially large ones.[638] Finally, the goal of this bioinformatic approach was to screen for drugs that could be used in cancer cell cultures – where there is a lack of TME information from additional cell types of presents.

As outlined in **section 1,** traditional *de novo* drug development is a multi-step process from molecule discovery to FDA review and safety monitoring.[639] It is a costly (millions of dollars) process across years of work and is marred by failure rates of up to 90%.[639] Drug repurposing is a resource saving method to overcome these issues seeking to find new uses for “ready-to-go” FDA approved drugs in other cancers or other indications entirely.[639] Between *de novo* and repurposing the time resource reduction is substantial.[639] The benefits of using already known drugs with FDA approval is clear. However, several of the hub genes identified in **section 6.4.4** were not currently targetable by such drugs.

Peptide drugs, specifically anti-cancer peptides (ACPs) for the treatment of cancer are bioactive peptides of up to 60 amino acids in length.[640] They work by binding cell surface receptors and modulating intracellular signalling in a way that promotes various anti-cancer pathways ultimately leading to apoptosis, membrane disruption or oncogenic pathway inhibition.[640] Compared to SMIs, peptides due to their larger size can effectively cover larger interactions surfaces, leading to more potent activity and specificity leading to reduced cytotoxicity.[636,638] Furthermore, peptides also typically exhibit much lower immunogenicity and higher membrane permeability compared to biologic treatment including monoclonal antibodies.[638,641]

Successful anti-cancer peptide drugs include Sandostatin, Zoladex, Decapeptyl, and Lutathera, although the number of FDA-approved anti-cancer peptide drugs is still limited.[636,640] This is likely due to their disadvantages which include short half-life, challenges in delivery and stability as they are targeted for degradation.[636,642] However, progress in peptide drugs for oncology has been slower compared to targeted therapy, even with enormous interest.[640] There are however a vast range of peptide drugs currently in pre-clinical testing and clinical trials for various cancers.[640]

7.5 Conclusion

To conclude, this project aimed to identify and present candidate drug targets for therapeutic development using biological networks. The motivation for this was to increase the number of drugs available to DDLPS patients. The WGCNA approach here was successful in proposing a putative drug target UBA1 and the drug TAK-243 which is currently undergoing clinical trials for acute myeloid leukaemia (AML) (ClinicalTrials.gov, ClinicalTrials.gov, NCT03816319) and advanced or metastatic solid tumours (NCT06223542). TAK-243, at the time of writing, has not been indicated for use in DDLPS. Whilst key limitations remain, these are not easily circumvented without higher sample numbers, and typically the methodology here has been robust. Furthermore, future directions of research were highlighted, including validating novel findings of the project in DDLPS, expanding the omics for multi-omics integration particularly proteomics focusing on patient matched experiment data, and expanding the search to other drug types.

Appendix A Supplementary material

A.1 Transcriptional/gene regulatory networks (GRNs)

Gene regulatory networks (GRNs) attempt to infer causal relationships between regulators (e.g., TFs or microRNA (miRNA) or long non-coding (lnc)RNAs) and target genes. Whilst in the literature, GCNs have also been labelled as gene regulatory network, there is the distinction that GCNs do not model causal relationships between entities.[643] To gain a better understanding of the regulatory interactions that are at work in systems (particularly for studies investigating the regulatory environment) it is possible to apply pre-defined or predicted regulator to target interactions in a GRN-based approach.[117,118] There are several methods used to infer GRNs. ARACNE, PANDA (Passing attributes between networks for data assimilation), CLR (Context likelihood relatedness), semi-supervised SIRENE (Supervised Inference of Regulatory Networks) and gene network inference and ensemble of trees (GENIE3) are some examples.[229,644-648] These methods use different approaches to infer regulatory interactions. For example, ARACNE infers direct relationships by filtering out indirect interactions from gene triplets by removing the edge with the lowest smallest MI value.[645,649] Whereas, GENIE3, uses a tree-based method (Random Forest) to infer a variable importance measure to rank input genes as potential regulators of other genes.[648] Each of these methods require as input a list specifying regulator (transcription factor) to target edges.[229,644-647] Transcription factor (TF), transcription factor binding sites (for validation) to target interactions and miR binding sites can be obtained from databases such PAZAR, TF2DNA, miRGen, and miRTarBase[650].[651-653] miRNA-target and TF-target information can be obtained from prediction tools such as TargetScan.[654]

A.2 Cox proportional hazards assumptions testing

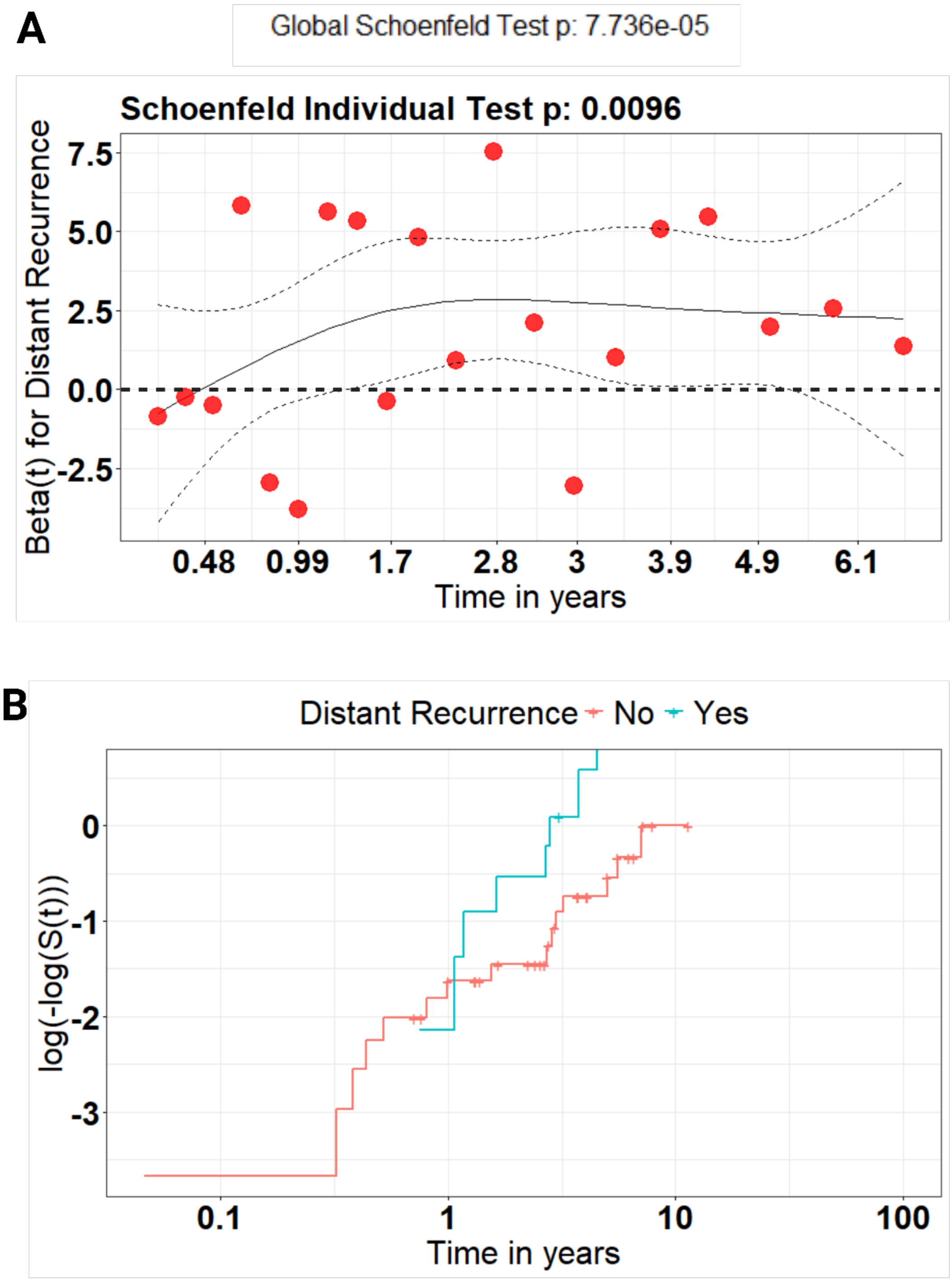


Figure A 1: Diagnostics of proportional hazards for distant recurrence. A: Schoenfeld residuals representing the differences in observed vs expected hazard overtime. **B:** A log-log plot of survival times against times stratified by distant recurrence.

A.3 NCC topology assessment

The NCC DDLPS cohort contains 32 DDLPS patient tumour samples (see **Section 3.4.1**) for which RNA-seq experiments were conducted. In assessing additional clinical information that was sent through correspondence with the data generators it was identified that 2 patients (sample ID: NCC_05 and NCC_NCC_07) did not go through surgery for tumour removal, of which one (sample ID: NCC_07) was from recurrent tissue. These samples were noted when conducting diagnostics for WGCNA.

The gene connectivity in the NCC data is approximately scale-free (***Figure A 2***). Outlier analysis of NCC data reveals a linear relationship between clustering coefficient and scaled connectivity values for sample gene expression profiles (***Figure A 3***). The distribution is visually similar to the TCGA SARC data after applying a 2 x MAD threshold with no obvious cliques. The sample located most distally from the main clique could be an outlier which is supported when drawing 2x MAD threshold lines. Samples that did not have surgery, NCC_05 is proximal to the maximally connected sample, whereas NCC_07 is located distally, also being a potential outlier upon visual inspection of the sample network metrics. Most of the samples that were recurrent (6/8) cluster within the majority samples in the main clique, suggesting that whether a tumour is recurrent, or primary, does not distinguish samples by clusters according to the sample network metrics.

The goodness-of-fit to a power law distribution across MAD thresholds (2 x MAD was used as the upper threshold) shows that R^2 values are higher in the original data (no threshold) for powers $9 \leq \beta \leq 16$ but use of stringent thresholds (median) maximises the R^2 value (0.90) for $\beta = 20$ (**Figure A 4**). Formally assessing the goodness-of-fit of the NCC data to a power law distribution reveal that for no threshold or power, an alternative exponential model best describes the degree distribution within the data (**Figure A 5**).

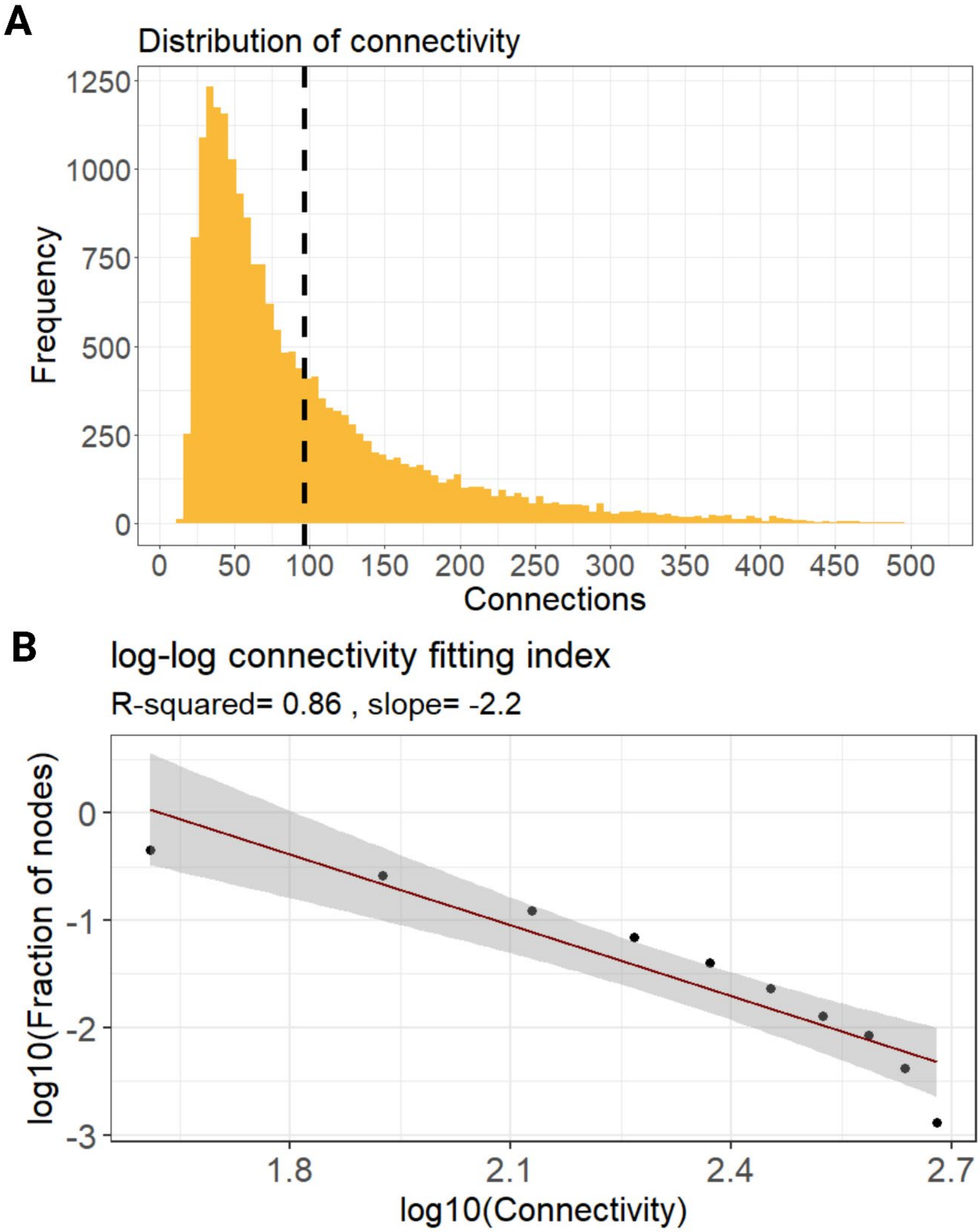


Figure A 2: A. Frequency of gene connectivity bins. **B.** Log-log plot of distribution of connectivity values. The red line indicates the line of best fit from a linear model and the grey shaded area is the 95% confidence interval.

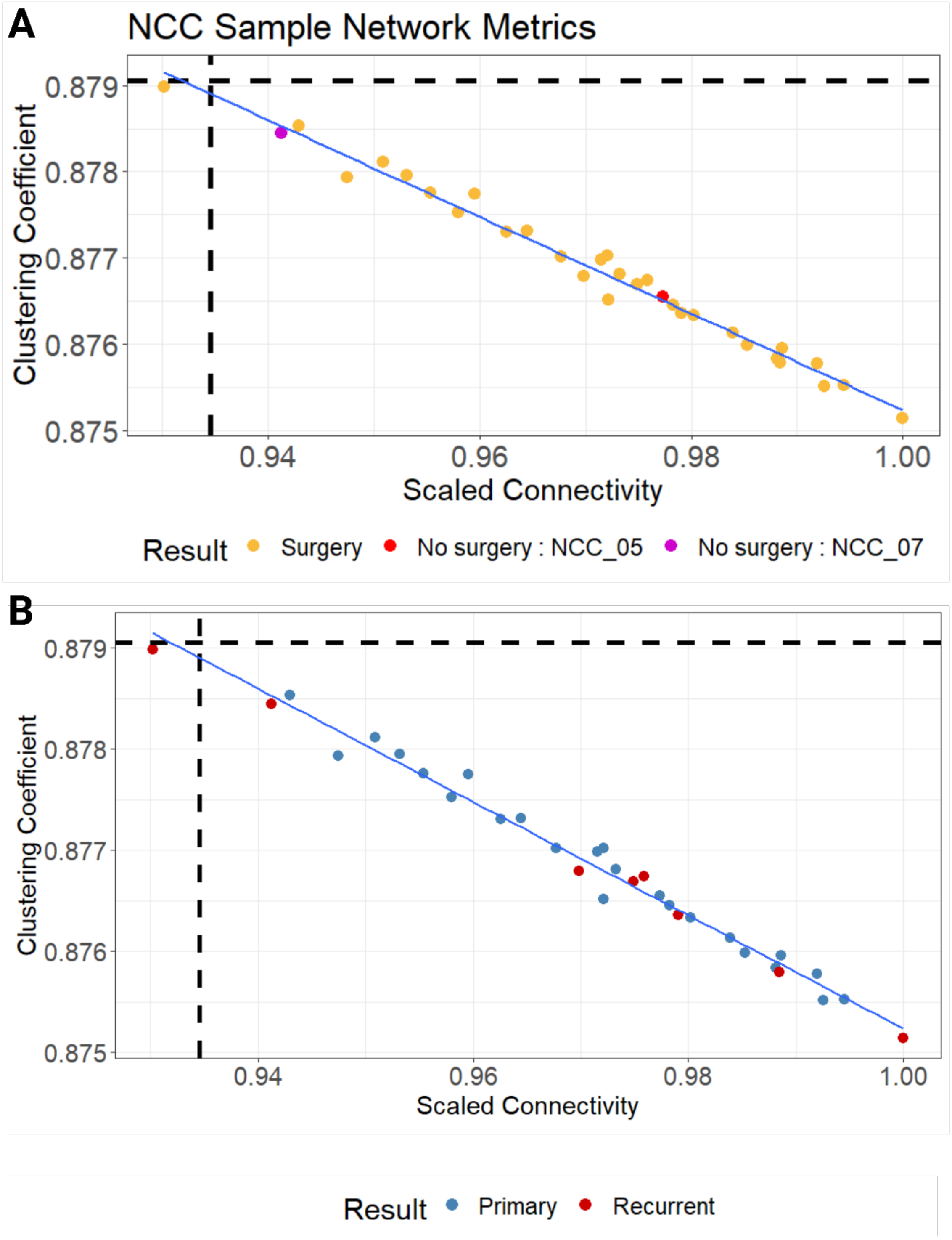


Figure A 3: Colour indicates samples from patients that underwent surgery (blue), patients that did not go through surgery but were from primary tissue (red) and those that did not go through surgery and were from recurrent tissue (magenta).

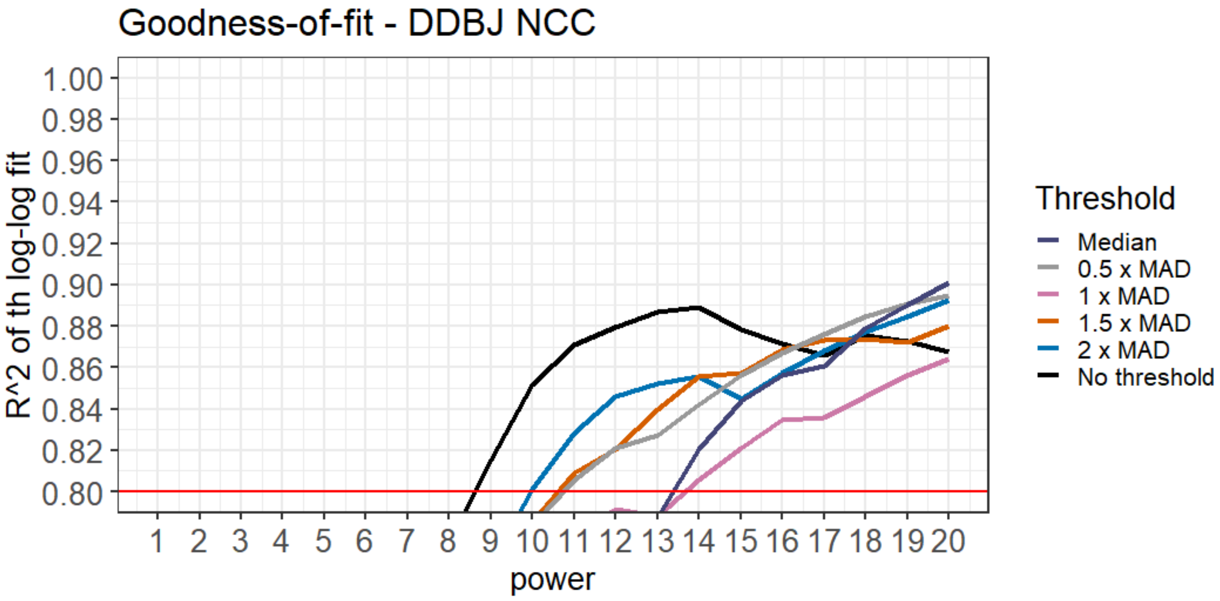


Figure A 4: R^2 of the linear model fit for connectivity across values of β for the tested thresholds. The red dotted line indicates the value of R^2 that suggests an approximately scale-free network which is set out by the authors of WGCNA as being 0.8.

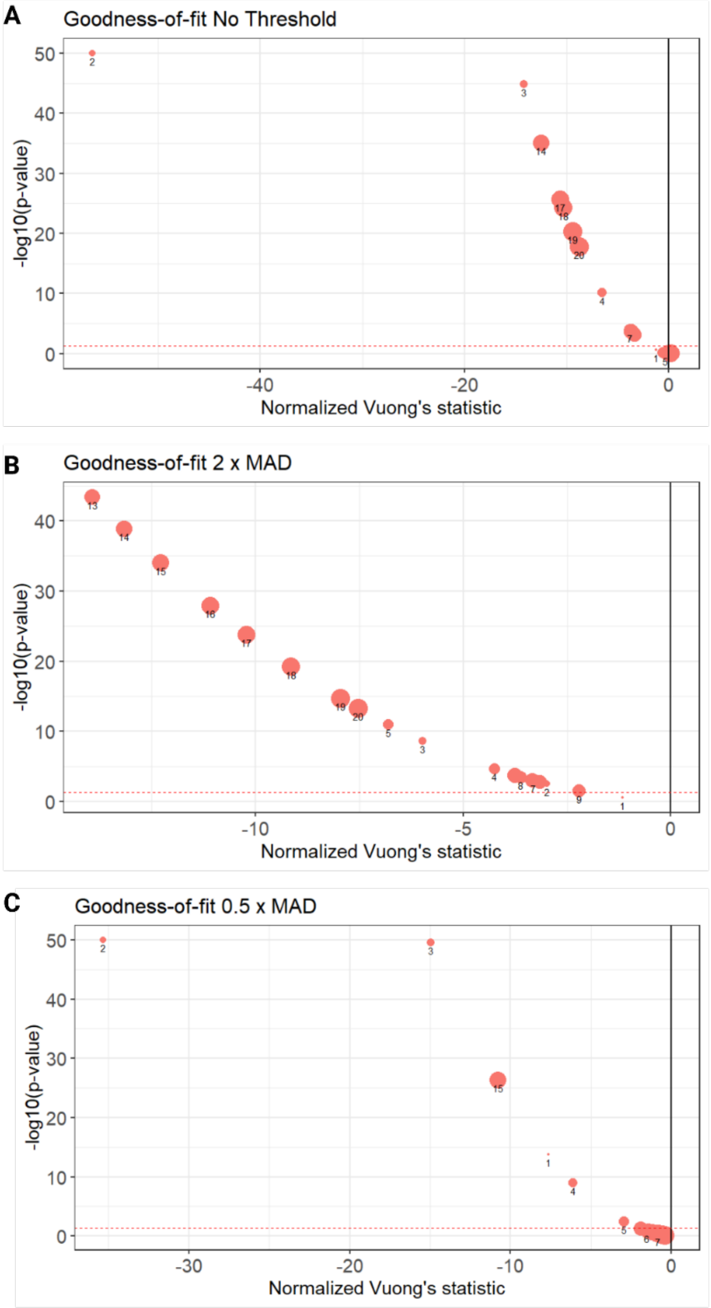


Figure A 5: Goodness-of-fit of connectivity distribution for power law. **A.** Connectivity distribution with no sample network metric filtering **B.** 2 x MAD threshold on sample network metrics. **C:** 0.5 x MAD threshold. Red dotted line indicates a significant result in the Vuong's test (-log10(p-value) ≥ 1.3) where any point above the line is significantly described by one of the distributions tested. The normalised Vuong's test statistic details which model is best fitted where any point right of the solid black line is best described by a power law distribution and any point left of the black line is best described with the alternative exponential distribution.

A.4 Heterogenous and repeated elements within the GCN

Filtering the ten most enriched gene sets for each module using a -log10 adjusted p-value > 1.3 and decreasing NES revealed 66, 94, and 21 unique enrichments across Reactome, GOBP, and Hallmark processes, respectively, highlighting a diverse range of biological functions within the GCN. Repeated enrichments across modules occur at varying frequencies (***Table A 1***). Cell cycle annotations were not observed in the top five most frequent GOBP enrichments due to the larger number of gene sets in GOBP.

Table A 1: Most frequent enrichments across Reactome, GOBP biological processes and Hallmark gene sets.

Rank according to frequency	Reactome term	Reactome ID	Number of modules	GOBP term	GOBP ID	Number of modules	Hallmark pathway	ID	Number of modules
1	Eukaryotic translation elongation	R-HSA-156842	30	Adaptive immune response	GO:0002250	34	Allograft rejection	M5950	38
2	Immunoregulatory interactions between a lymphoid and a non-lymphoid cell	R-HSA-198933	24	Cytoplasmic translation	GO:0002181	16	Oxidative phosphorylation	M5936	32
3	Collagen chain trimerization	R-HSA-8948216	15	Ribosome biogenesis	GO:0042254	10	E2F target	M5925	27
4	Mitotic prometaphase	R-HSA-68877	15	ATP synthesis coupled electron transport	GO:0042773	8	Myogenesis	M5909	27
5	Resolution of sister chromatid adhesion	R-HSA-2500257	15	Proton motive force driven ATP synthesis	GO:0042776	8	Mitotic spindle	M5893	20

GOBP – Gene Ontology Biological Processes.

A.5 DDLPS frequent amplifications throughout the DDLPS GCN

MDM2 (M82)*,* CDK4 (M149), *FRS2* (M115), and *HMGA2* (M78) are the most frequently amplified genes (70-100% of all samples with strong co-occurrence) in relation to DDLPS (its variants and WDLPS). Therefore, it was of interest to assess modular relationships to the ME expression of modules in which these genes were sorted (***Figure A 6A-D***).[37] *CDK4* (***Figure A 6A***) and *FRS2* (***Figure A 6B***) highlight relationships with modules associated with cell cycle and developmental processes. They show a stronger modular distribution of correlation compared to *HMGA2* (***Figure A 6C***) and *MDM2* (***Figure A 6D***), which localise to similar modules annotated for oxidative phosphorylation and developmental processes (***Section 5.4.1, Figure 5.2***).

MDM2 was clustered into M82, which is the largest module (n genes) in the GCN. This "grey" module is characterised by genes that exhibit independent patterns of co-expression. It could be that the variance of *MDM2* expression between samples was low (variance = 0.4, 9025th ranked gene by variance), which would impact downstream calculations of correlation coefficients (cf. *CDK4* had a variance of 1.5 and was ranked 3188). Using the kME ranked gene lists to assess enrichment for Reactome pathways: M78 was found to be most strongly enriched for Reactome Activation of the pre-replicative complex (NES = 2.619, -log10p = 7.191). M115 was most enriched for Reactome resolution of sister chromatid cohesion (NES = 2.81, -log10p = 14.64). M149 was most enriched for Reactome collagen chain trimerization (NES = 2.717, -log10p = 7.848).

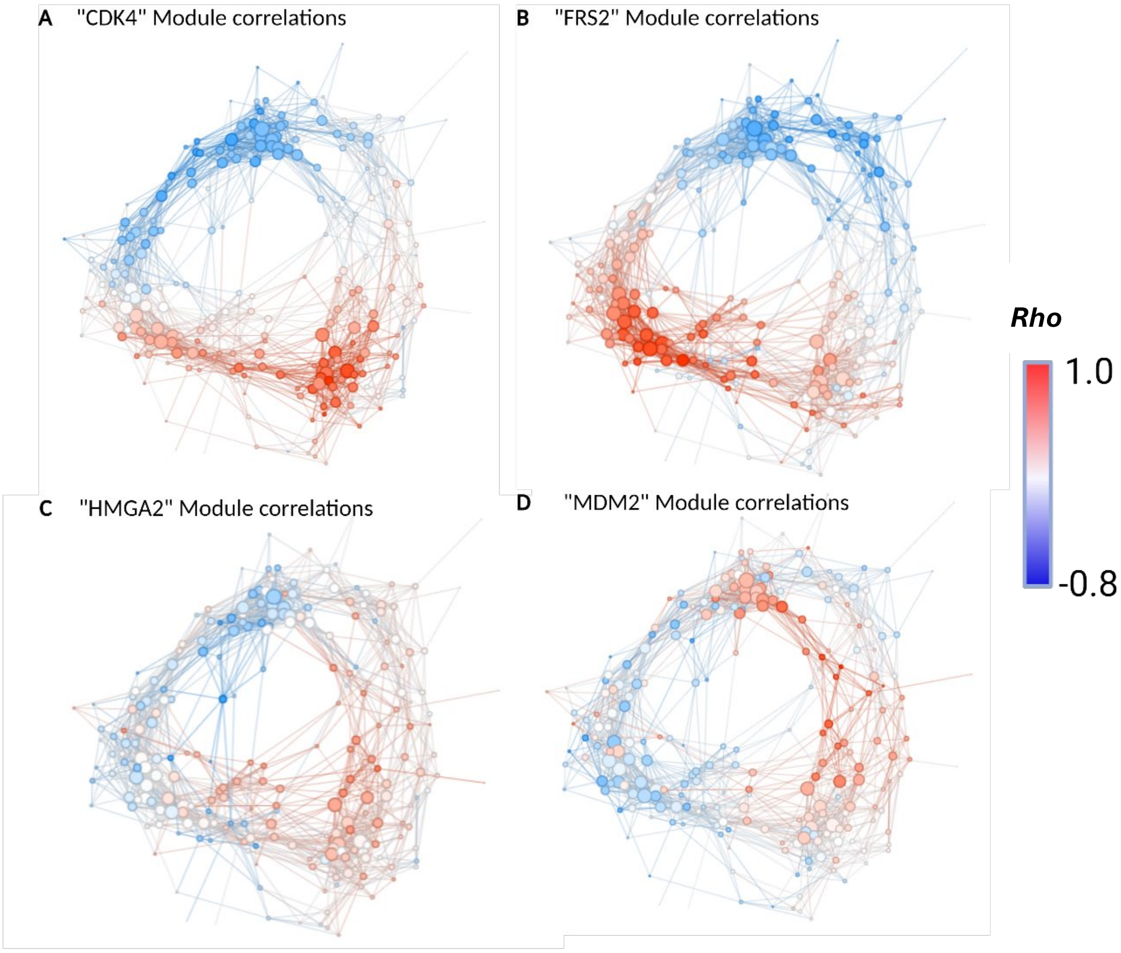

A "CDK4" Module correlations
B "FRS2" Module correlations
C "HMGA2" Module correlations
D "MDM2" Module correlations
Rho
1.0
-0.8

Figure A 6: EGN with ME correlation to modules containing **A** CDK4, **B** RFS2, **C** HMGA2, and **D** MDM2. Colour indicates the spearman correlation. Node size is proportional to the degree.

A.6 Pathway enrichments in top ranking modules by GS

The top module associated with the clinical score was M77 (***Figure A 7A***) (*rho* = 0.64), enriched for eukaryotic translation and immunoregulatory interactions (***Figure A 7B***). Translation gene set does not include any intersecting M77 genes suggesting the enrichment is driven by high kME values of genes assigned to other modules. Enrichments for B-cell receptor signalling pathway (NES = 2.51, -*log10adj-p = 7.1*) showed *BMX Non-Receptor Tyrosine Kinase* (*BMX*) as the overlapping gene. In addition, complement cascade annotations were retrieved with the Complement-4 (C4) genes *C4A* and *C4B* present in the module.

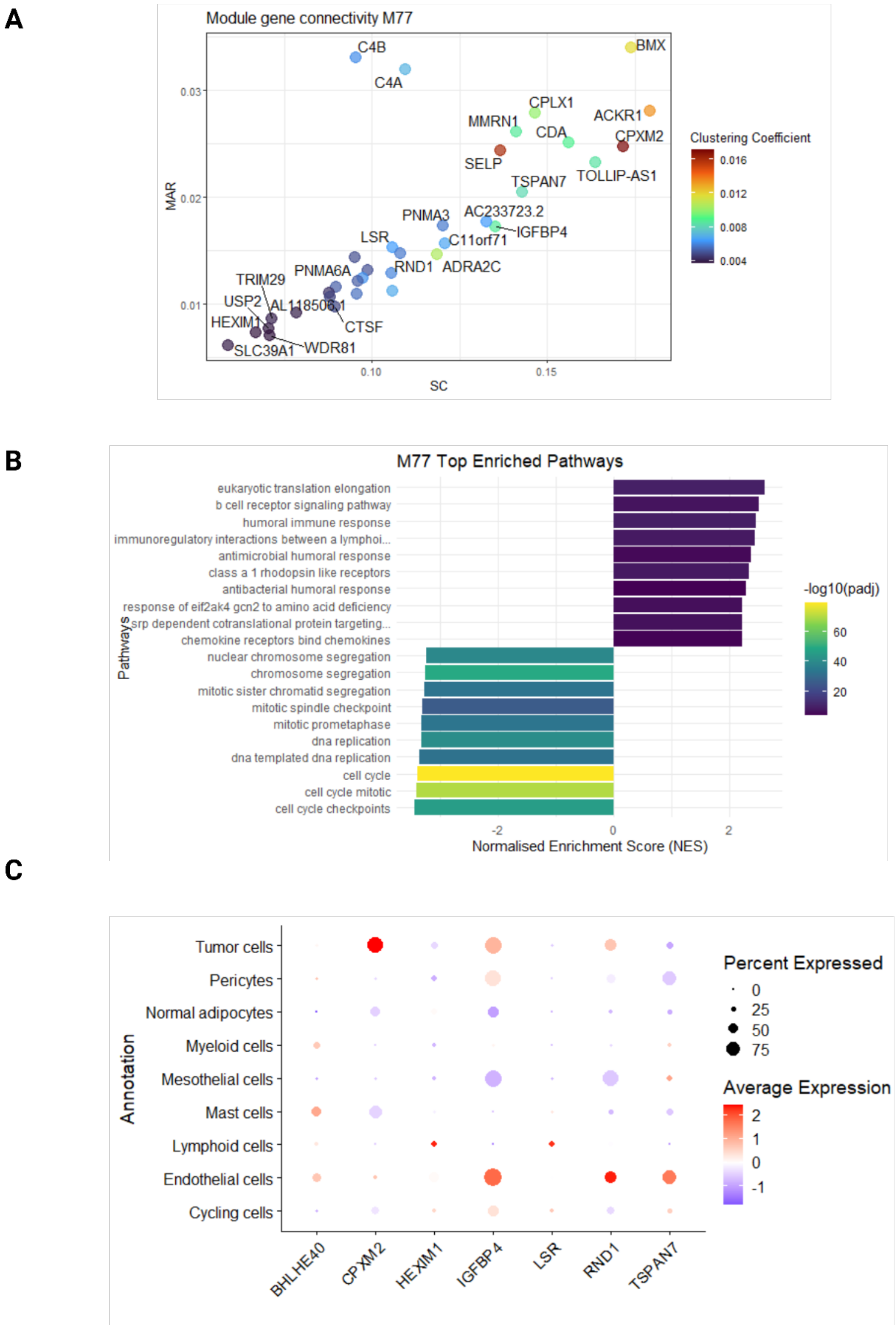


Figure A 7: M77 module analysis. **A** Module connectivity graph showing the SC – scaled connectivity, and MAR – Maximum adjacency ratio, coloured by the clustering coefficient. **B** Reactome and GOBP enrichment analysis results passing a -log10 adj p > 1.3, displayed are the top ten positive and negative results ordered by NES – normalised enrichment score, coloured by -log10adjp value. **C** Intersect of M77 genes with single cell RNA-seq cell clusters, labelled according to "global" annotations. Figure created in BioRender.com.

Assessing the overlap of module genes with single-cell data reveals that *Carboxypeptidase X M14 Family Member 2 (CPXM2*) is expressed in tumour cells (***Figure A 7C***), aligning with the ECMO enrichments retrieved from GSEA (***Table A 2***). Although the overlap details small proportion of genes, there is suggestion of an association between the expression of genes in endothelial cells (EC) (e.g., *insulin like growth factor binding protein 4* (*IGFBP4*) and *Rho family GTPase 1* (*RND1*)) and tumour cells. These associations may depict transcriptional programmes relating to tumour-associated vasculature.

The top module associated with the gene effect score was M106 (***Figure A 8A***) (*rho* = 0.63). This module was associated with translation processes (***Figure A 8B***) and contained multiple ribosomal protein genes such as those of the 40S small ribosomal subunit, including *Ribosomal protein S* (*RPS*) *RPS5, RPS19* and *RPS20,* and the 60S large ribosomal subunit, including *Ribosomal protein L (RPL) RPL6, RPL18, RPL31,* among others such as *Ribosomal protein lateral stalk unit P0* (*RPLP0*). These proteins are highly essential to cellular functions, hence the correlation of this module with high gene essentiality.

The observed higher relative expression of these genes in adipocytic and lymphocytic cells may be due to their increased demand for protein synthesis (***Figure A 8C***). The lower expression of such genes in tumour cells suggests that the annotations from translation processes in the GCN partition represent signatures from across the TME.

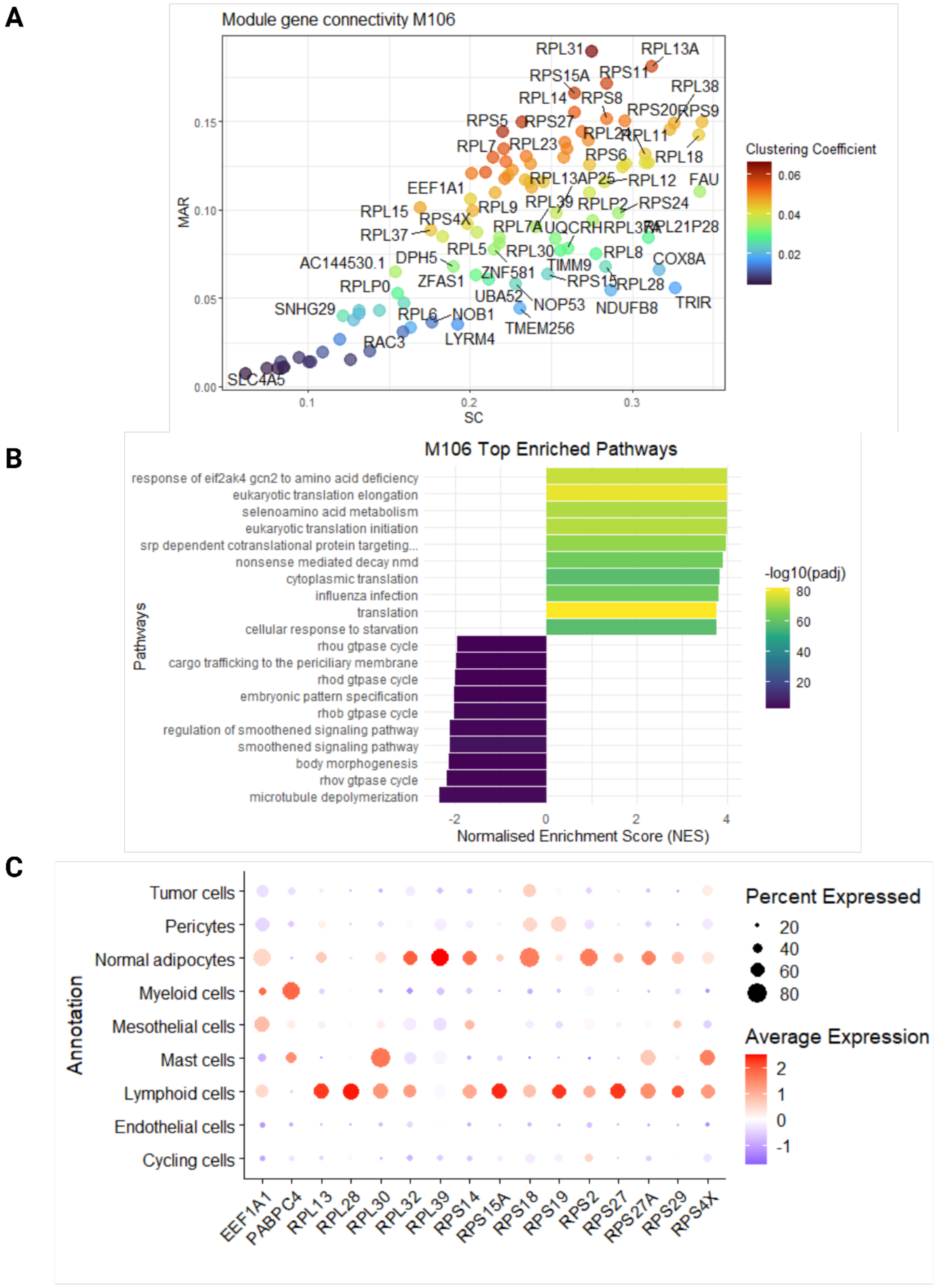


Figure A 8: M106 module analysis. **A** Module connectivity graph showing the SC – scaled connectivity, and MAR – Maximum adjacency ratio, coloured by the clustering coefficient. **B** Reactome and GOBP enrichment analysis results passing a -log10 adj p > 1.3, displayed are the top ten positive and negative results ordered by NES – normalised enrichment score, coloured by -log10 adj p value. **C** Intersect of M106 genes with single cell RNA-seq cell clusters, labelled according to “global” annotations. Figure created in BioRender.com.

A.7 Extended GSEA results for low-ranking modules

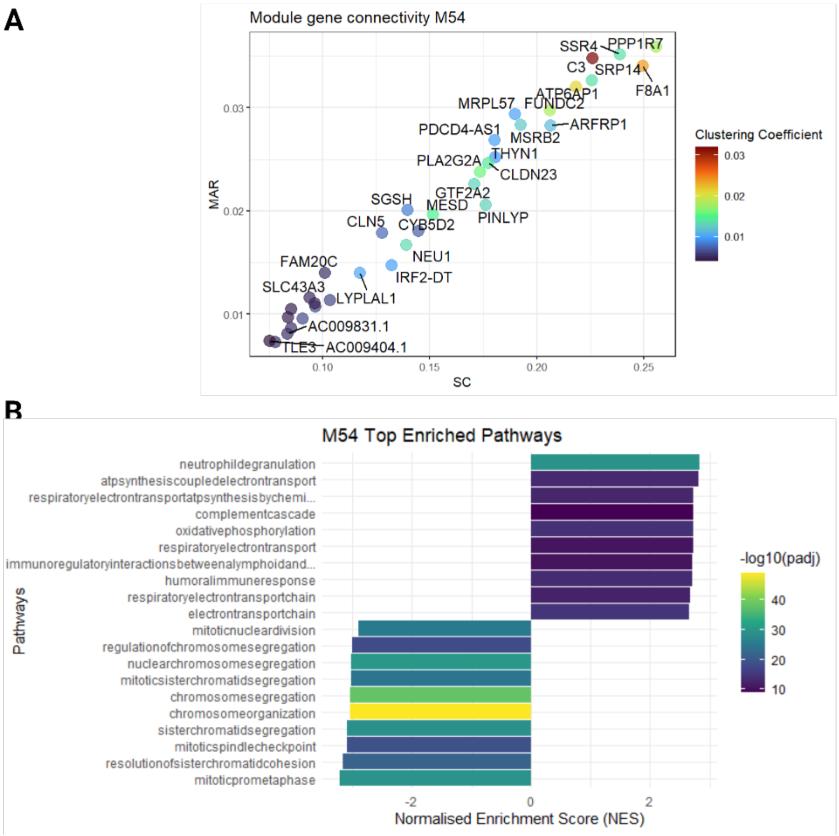


Figure A 9: **A** Module connectivity for M54 detailing the SC – scaled connectivity and the MAR – maximum adjacency ratio within the module coloured by the clustering coefficient. Combined Reactome and GOBP (Gene Ontology Biological Processes) gene set enrichment results for **B** M54. Bars are coloured by the -log10 adjusted p-value (filtered for -log10 adjusted p > 1.3) and sorted according to the top and bottom ten processes by NES – normalised enrichment score.

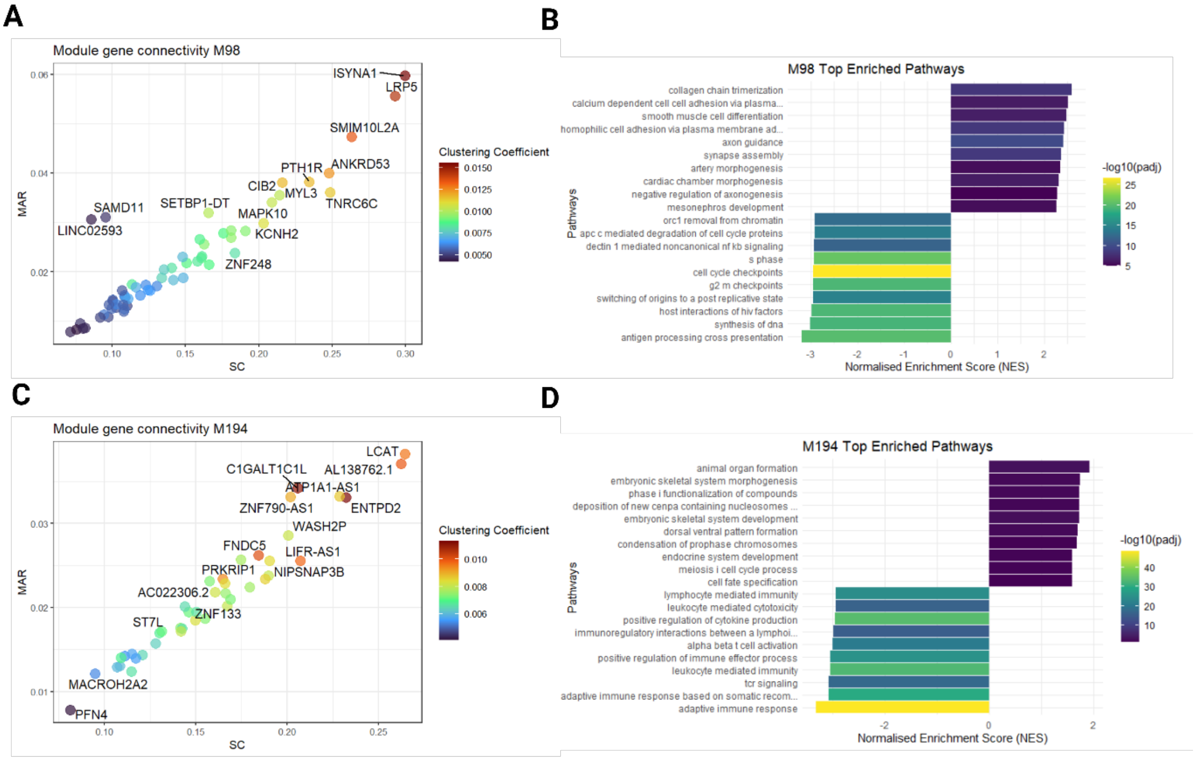


Figure A 10: Module analysis results for M98 (**A-B**) and M194 (**C-D**). Module connectivity for **A** M98 and **C** M194 detailing the SC – scaled connectivity and the MAR – maximum adjacency ratio within the module coloured by the clustering coefficient. Combined Reactome and GOBP (Gene Ontology Biological Processes) gene set enrichment results for **B** M98 and **D** M194. Bars are coloured by the -log10 adjusted p-value (filtered for -log10 adjusted p > 1.3) and sorted according to the top and bottom ten processes by NES – normalised enrichment score.

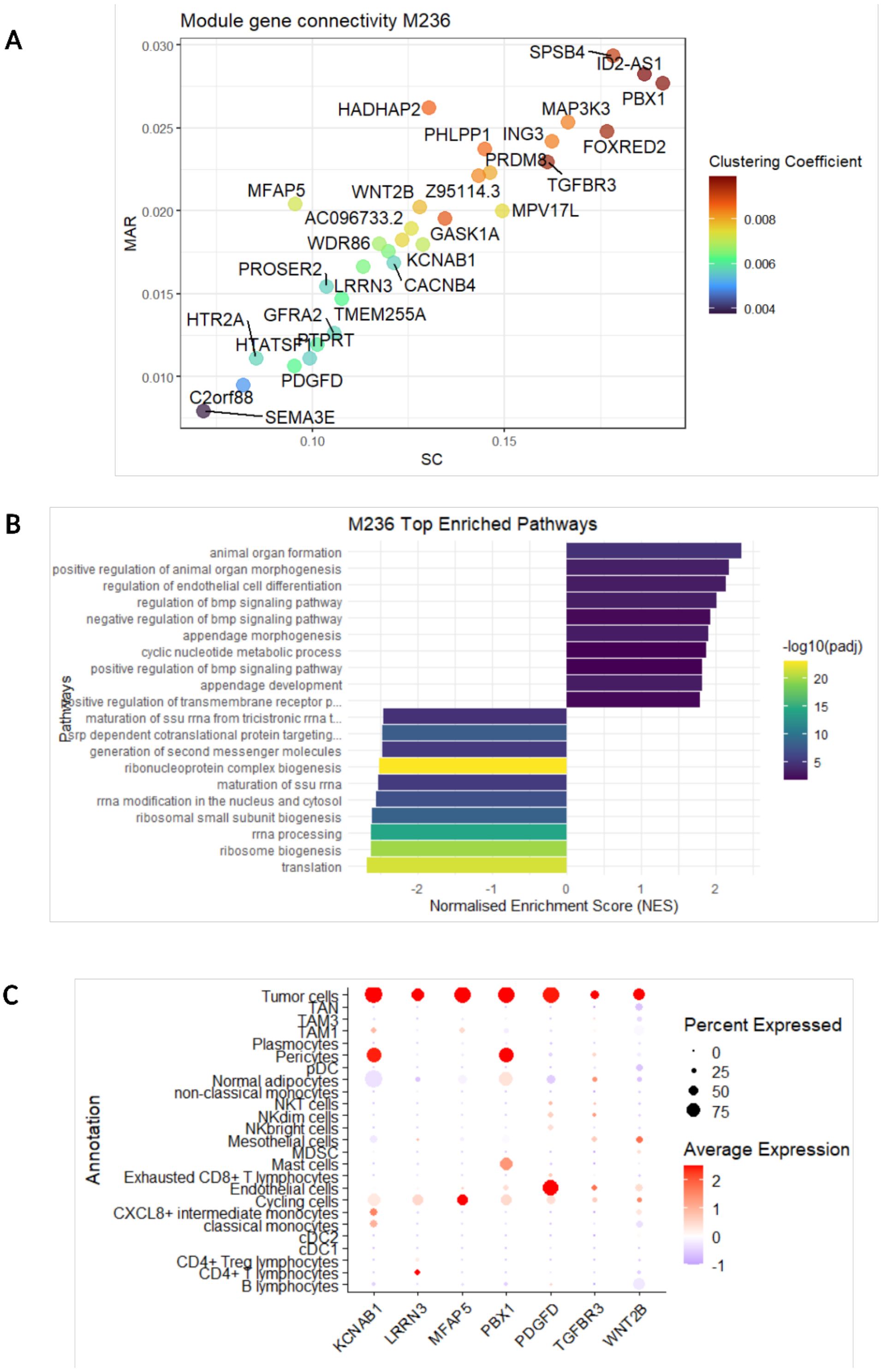


Figure A 11: M236 module analysis. **A** Module connectivity graph showing the SC – scaled connectivity, and MAR – Maximum adjacency ratio, coloured by the clustering coefficient. **B** Reactome and GOBP enrichment analysis results passing a -log10 adj p > 1.3, displayed are the top ten positive and negative results ordered by NES – normalised enrichment score, coloured by -log10 adj p value. **C** Intersect of M236 genes with single cell RNA-seq cell clusters, labelled according to “global” annotations. Figure created in BioRender.com.

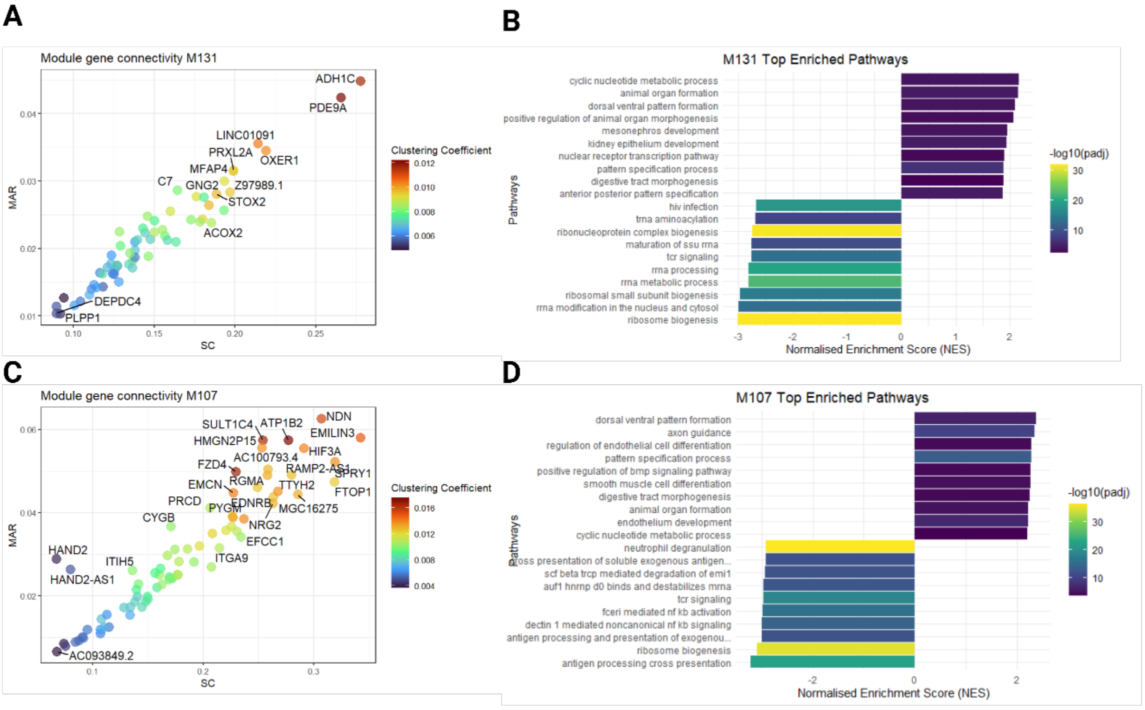


Figure A 12: Module analysis results for M131 (**A-B**) and M107 (**C-D**). Module connectivity for **A** M131 and **C** M107 detailing the SC – scaled connectivity and the MAR – maximum adjacency ratio within the module coloured by the clustering coefficient. Combined Reactome and GOBP (Gene Ontology Biological Processes) gene set enrichment results for **B** M131 and **D** M107. Bars are coloured by the -log10 adjusted p-value (filtered for -log10 adjusted p > 1.3) and sorted according to the top and bottom ten processes by NES – normalised enrichment score.

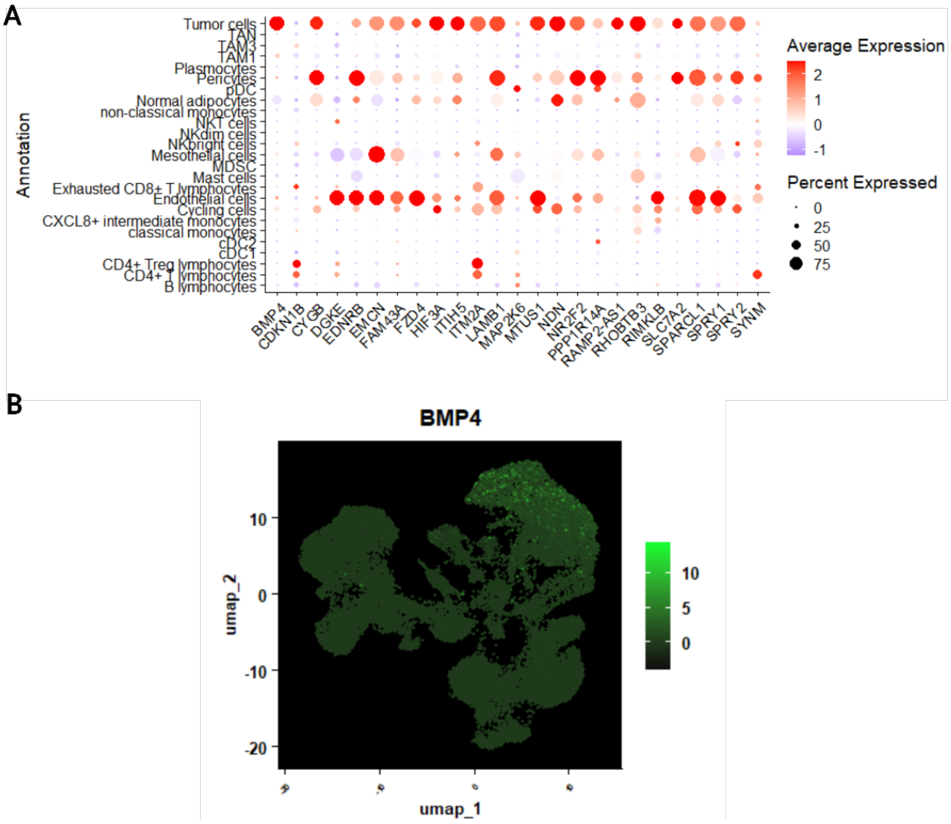


Figure A 13: A Overlap between genes within M107 and single-cell data coloured by average expression among cells. The size of the circle represents the percentage of cells expressing the annotated genes. **B** Expression of BMP4 across the TME (left) showing highest in the tumour.

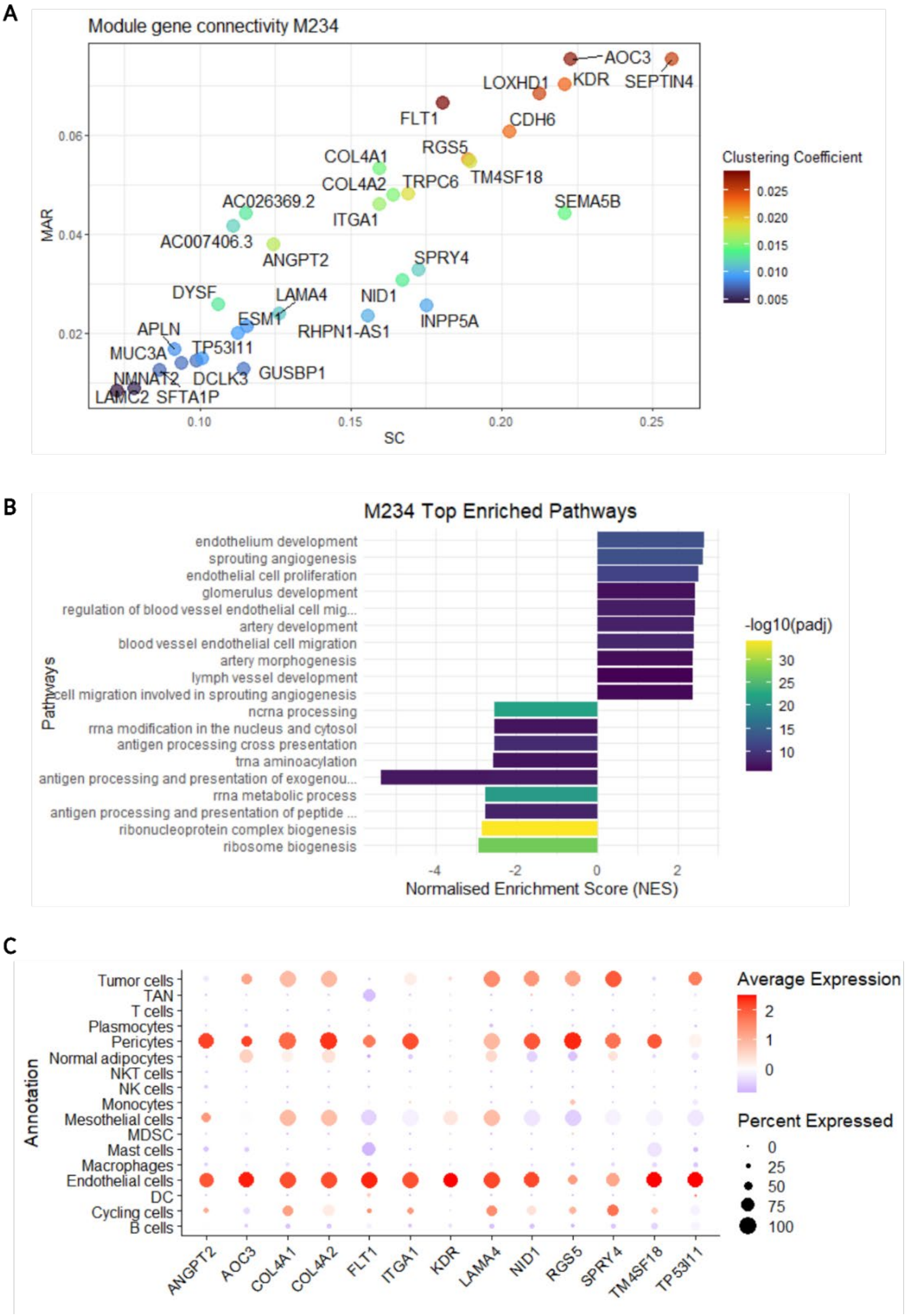


Figure A 14: M234 module analysis. **A** Module connectivity graph showing the SC – scaled connectivity, and MAR – Maximum adjacency ratio, coloured by the clustering coefficient. **B** Reactome and GOBP enrichment analysis results passing a -log10 adj p > 1.3, displayed are the top ten positive and negative results ordered by NES – normalised enrichment score, coloured by -log10 adj p value. **C** Intersect of M234 genes with single cell RNA-seq cell clusters, labelled according to “global” annotations. Figure created in BioRender.com.

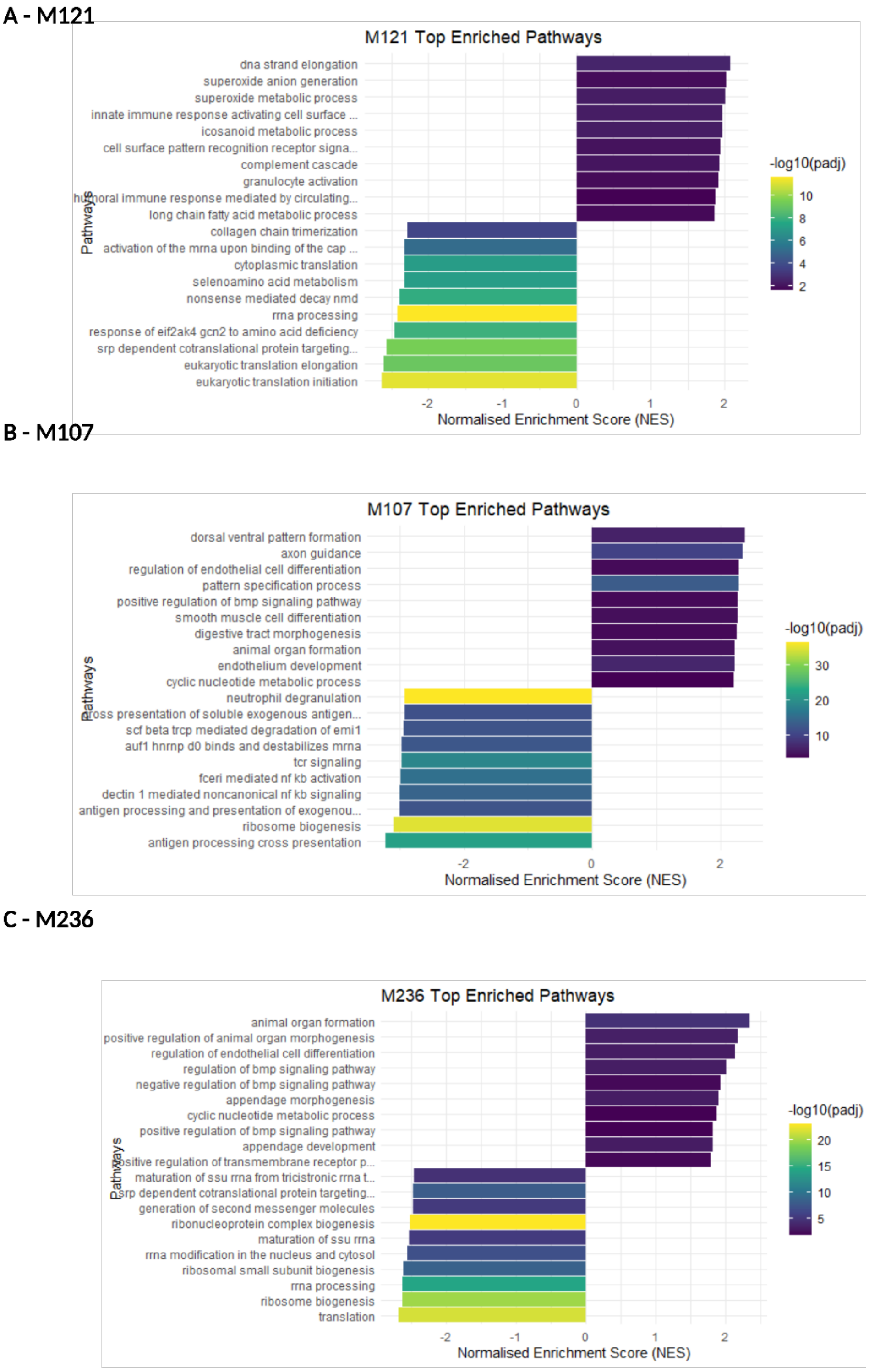


Figure A 15: Top enriched pathways using Reactome and Gene Ontology Biological Processes gene sets for **A** M121, **B** M107 and **C** M236 colour indicates the -log10(adjusted p) enrichment.

Bibliography

1. Choi JH, Ro JY. The 2020 WHO Classification of Tumors of Bone: An Updated Review. *Adv Anat Pathol* 2021;28(3):119-38.

2. Sbaraglia M, Bellan E, Dei Tos AP. The 2020 WHO Classification of Soft Tissue Tumours: news and perspectives. *Pathologica* 2021;113(2):70-84.

3. Hayes AJ, Nixon IF, Strauss DC, et al. UK guidelines for the management of soft tissue sarcomas. *British Journal of Cancer* 2024.

4. Henricks WH, Chu YC, Goldblum JR, Weiss SW. Dedifferentiated Liposarcoma: A Clinicopathological Analysis of 155 Cases with a Proposal for an Expanded Definition of Dedifferentiation. *The American Journal of Surgical Pathology* 1997;31(3):271-81.

5. Ishibe T, Nakayama T, Aoyama T, et al. Neuronal differentiation of synovial sarcoma and its therapeutic application. *Clin Orthop Relat Res* 2008;466(9):2147-55.

6. Burningham Z, Hashibe M, Spector L, Schiffman JD. The epidemiology of sarcoma. *Clin Sarcoma Res* 2012;2(1):14.

7. Haddox CL, Riedel RF. Recent advances in the understanding and management of liposarcoma. *Fac Rev* 2021;10:1.

8. Bacon A, Wong K, Fernando MS, et al. Incidence and survival of soft tissue sarcoma in England between 2013 and 2017, an analysis from the National Cancer Registration and Analysis Service. *Int J Cancer* 2023;152(9):1789-803.

9. Gage MM, Nagarajan N, Ruck JM, et al. Sarcomas in the United States: Recent trends and a call for improved staging. *Oncotarget* 2019;10(25).

10. Bessen T, Caughey GE, Shakib S, et al. A population-based study of soft tissue sarcoma incidence and survival in Australia: An analysis of 26,970 cases. *Cancer Epidemiol* 2019;63:101590.

11. Stiller CA, Trama A, Serraino D, et al. Descriptive epidemiology of sarcomas in Europe: Report from the RARECARE project. *European Journal of Cancer* 2013;49(3):684-95.

12. Bestic JM, Kransdorf MJ, White LM, et al. Sclerosing Variant of Well-Differentiated Liposarcoma: Relative Prevalence and Spectrum of CT and MRI Features. *American Journal of Roentgenology* 2013;201(1):154-61.

13. Moreau L-C, Turcotte R, Ferguson P, et al. Myxoid\Round Cell Liposarcoma (MRCLS) Revisited: An Analysis of 418 Primarily Managed Cases. *Annals of Surgical Oncology* 2012;19(4):1081-88.

14. Machhada A, Emam A, Colavitti G, et al. Liposarcoma subtype recurrence and survival: A UK regional cohort study. *J Plast Reconstr Aesthet Surg* 2022;75(7):2098-107.

15. Bourcier K, Dinart D, Le Cesne A, et al. Outcome of Patients with Soft-Tissue Sarcomas: An Age-Specific Conditional Survival Analysis. *The Oncologist* 2019;24(7):e559-e64.

16. Knebel C, Lenze U, Pohlig F, et al. Prognostic factors and outcome of Liposarcoma patients: a retrospective evaluation over 15 years. *BMC Cancer* 2017;17(1):410.

17. Ducimetiere F, Lurkin A, Ranchere-Vince D, et al. Incidence of sarcoma histotypes and molecular subtypes in a prospective epidemiological study with central pathology review and molecular testing. *PLoS One* 2011;6(8):e20294.

18. Bock S, Hoffmann DG, Jiang Y, et al. Increasing Incidence of Liposarcoma: A Population-Based Study of National Surveillance Databases, 2001-2016. *Int J Environ Res Public Health* 2020;17(8).

19. Mastrangelo G, Coindre JM, Ducimetiere F, et al. Incidence of soft tissue sarcoma and beyond: a population-based prospective study in 3 European regions. *Cancer* 2012;118(21):5339-48.

20. Dermawan JK, Hwang S, Wexler L, et al. Myxoid pleomorphic liposarcoma is distinguished from other liposarcomas by widespread loss of heterozygosity and significantly worse overall survival: a genomic and clinicopathologic study. *Mod Pathol* 2022;35(11):1644-55.

21. Creytens D, Folpe AL, Koelsche C, et al. Myxoid pleomorphic liposarcoma—a clinicopathologic, immunohistochemical, molecular genetic and epigenetic study of 12 cases, suggesting a possible relationship with conventional pleomorphic liposarcoma. *Modern Pathology* 2021;34(11):2043-49.

22. Amer KM, Congiusta DV, Thomson JE, et al. Epidemiology and survival of liposarcoma and its subtypes: A dual database analysis. *J Clin Orthop Trauma* 2020;11(Suppl 4):S479-S84.

23. Lee ATJ, Thway K, Huang PH, Jones RL. Clinical and Molecular Spectrum of Liposarcoma. *J Clin Oncol* 2018;36(2):151-59.

24. Dalal KM, Kattan MW, Antonescu CR, et al. Subtype specific prognostic nomogram for patients with primary liposarcoma of the retroperitoneum, extremity, or trunk. *Ann Surg* 2006;244(3):381-91.

25. Hornick JL, Bosenberg MW, Mentzel T, et al. Pleomorphic Liposarcoma: Clinicopathologic Analysis of 57 Cases. *The American Journal of Surgical Pathology* 2004;28(10):1257-67.

26. Cates JMM. The AJCC 8th Edition Staging System for Soft Tissue Sarcoma of the Extremities or Trunk: A Cohort Study of the SEER Database. *Journal of the National Comprehensive Cancer Network J Natl Compr Canc Netw* 2018;16(2):144-52.

27. Porrino J, Al-Dasuqi K, Irshaid L, et al. Update of pediatric soft tissue tumors with review of conventional MRI appearance—part 1: tumor-like lesions, adipocytic tumors, fibroblastic and myofibroblastic tumors, and perivascular tumors. *Skeletal Radiology* 2022;51(3):477-504.

28. Zhao J, Du W, Tao X, et al. Survival and prognostic factors among different types of liposarcomas based on SEER database. *Scientific Reports* 2025;15(1):1790.

29. Dangoor A, Seddon B, Gerrand C, et al. UK guidelines for the management of soft tissue sarcomas. *Clin Sarcoma Res* 2016;6:20.

30. Santoscoy J, Castillo R, Jose J, et al. Liposarcoma: A Pictorial and Literature Review. *Journal of Clinical Research in Radiology* 2018;1.

31. Demir M, Guven DC, Aktas BY, et al. Clinical features and prognosis of patients with liposarcoma: Single-center experience. *Journal of Clinical Oncology* 2019;37(15_suppl):e22557-e57.

32. Matthyssens LE, Creytens D, Ceelen WP. Retroperitoneal Liposarcoma: Current Insights in Diagnosis and Treatment. *Frontiers in Surgery* 2015;2(4).

33. Setsu N, Miyake M, Wakai S, et al. Primary Retroperitoneal Myxoid Liposarcomas. *The American Journal of Surgical Pathology* 2016;40(9):1286-90.

34. Wang L, Ren W, Zhou X, et al. Pleomorphic liposarcoma: a clinicopathological, immunohistochemical and molecular cytogenetic study of 32 additional cases. *Pathol Int* 2013;63(11):523-31.

35. Smith CA, Martinez SR, Tseng WH, et al. Predicting Survival for Well-Differentiated Liposarcoma: The Importance of Tumor Location1. *Journal of Surgical Research* 2012;175(1):12-17.

36. Gootee J, Aurit S, Curtin C, Silberstein P. Primary anatomical site, adjuvant therapy, and other prognostic variables for dedifferentiated liposarcoma. *J Cancer Res Clin Oncol* 2019;145(1):181-92.

37. Lu J, Wood D, Ingley E, et al. Update on genomic and molecular landscapes of well-differentiated liposarcoma and dedifferentiated liposarcoma. *Mol Biol Rep* 2021;48(4):3637-47.

38. Rauh J, Klein A, Baur-Melnyk A, et al. The role of surgical margins in atypical Lipomatous Tumours of the extremities. *BMC Musculoskeletal Disorders* 2018;19(1):152.

39. Sbaraglia M, Dei Tos AP. The pathology of soft tissue sarcomas. *La radiologia medica* 2019;124(4):266-81.

40. Patrichi AI, Gurzu S. Pathogenetic and molecular classifications of soft tissue and bone tumors: A 2024 update. *Pathology - Research and Practice* 2024;260:155406.

41. Dangoor A, Seddon B, Gerrand C, et al. UK guidelines for the management of soft tissue sarcomas. *Clinical Sarcoma Research* 2016;6(1):20.

42. Sbaraglia M, Dei Tos AP. The pathology of soft tissue sarcomas. *Radiol Med* 2019;124(4):266-81.

43. Dry SM. Dedifferentiation in bone and soft tissue sarcomas: How do we define it? What is prognostically relevant? *Human Pathology* 2024;147:139-47.

44. Sbaraglia M, Dei Tos AP. 12 - Adipocytic Tumors. In: Hornick JL (ed.) *Practical Soft Tissue Pathology: a Diagnostic Approach (Second Edition)*. Philadelphia: Elsevier; 2019 p311-40.

45. Jo VY, Hornick JL. 5 - Tumors With Myxoid Stroma. In: Hornick JL (ed.) *Practical Soft Tissue Pathology: a Diagnostic Approach (Second Edition)*. Philadelphia: Elsevier; 2019 p135-63.

46. Abaricia S, Hirbe AC. Diagnosis and Treatment of Myxoid Liposarcomas: Histology Matters. *Curr Treat Options Oncol* 2018;19(12):64.

47. Muratori F, Bettini L, Frenos F, et al. Myxoid Liposarcoma: Prognostic Factors and Metastatic Pattern in a Series of 148 Patients Treated at a Single Institution. *International Journal of Surgical Oncology* 2018;2018:8928706.

48. Tuzzato G, Laranga R, Ostetto F, et al. Primary High-Grade Myxoid Liposarcoma of the Extremities: Prognostic Factors and Metastatic Pattern. *Cancers (Basel)* 2022;14(11).

49. Keung EZ, Ikoma N, Benjamin R, et al. The clinical behavior of well differentiated liposarcoma can be extremely variable: A retrospective cohort study at a major sarcoma center. *J Surg Oncol* 2018;117(8):1799-805.

50. Hisaoka M. Lipoblast: morphologic features and diagnostic value. *J uoeh* 2014;36(2):115-21.

51. Gjorgova Gjeorgjievski S, Thway K, Dermawan JK, et al. Pleomorphic Liposarcoma: A Series of 120 Cases With Emphasis on Morphologic Variants. *Am J Surg Pathol* 2022;46(12):1700-05.

52. Nessim C, Raut CP, Callegaro D, et al. Analysis of Differentiation Changes and Outcomes at Time of First Recurrence of Retroperitoneal Liposarcoma by Transatlantic Australasian Retroperitoneal Sarcoma Working Group (TARPSWG). *Ann Surg Oncol* 2021;28(12):7854-63.

53. Singer S, Antonescu CR, Riedel E, Brennan MF. Histologic subtype and margin of resection predict pattern of recurrence and survival for retroperitoneal liposarcoma. *Annals of Surgery* 2003;238(3):358-71.

54. Masaki N, Onozawa M, Inoue T, et al. Clinical features of multiply recurrent retroperitoneal liposarcoma: A single-center experience. *Asian Journal of Surgery* 2021;44(1):380-85.

55. Ikoma N, Torres KE, Somaiah N, et al. Accuracy of preoperative percutaneous biopsy for the diagnosis of retroperitoneal liposarcoma subtypes. *Ann Surg Oncol* 2015;22(4):1068-72.

56. Marino-Enriquez A, Fletcher CD, Dal Cin P, Hornick JL. Dedifferentiated liposarcoma with "homologous" lipoblastic (pleomorphic liposarcoma-like) differentiation: clinicopathologic and molecular analysis of a series suggesting revised diagnostic criteria. *Am J Surg Pathol* 2010;34(8):1122-31.

57. Saeed-Chesterman D, Thway K. Homologous Lipoblastic Differentiation in Dedifferentiated Liposarcoma. *International Journal of Surgical Pathology* 2016;24(3):237-39.

58. Gordhandas J, Lin G, Tipps AMP, Zare SY. Osteosarcomatous Divergence in Dedifferentiated Liposarcoma Presenting as a Colonic Mass. *Case Rep Pathol* 2019;2019:8025103.

59. Kurzawa P, Mullen JT, Chen YL, et al. Prognostic Value of Myogenic Differentiation in Dedifferentiated Liposarcoma. *Am J Surg Pathol* 2020;44(6):799-804.

60. Nishio J, Nakayama S, Nabeshima K, Yamamoto T. Biology and Management of Dedifferentiated Liposarcoma: State of the Art and Perspectives. *J Clin Med* 2021;10(15).

61. Grimer R, Judson I, Peake D, Seddon B. Guidelines for the management of soft tissue sarcomas. *Sarcoma* 2010;2010:506182.

62. Coindre JM, Pedeutour F, Aurias A. Well-differentiated and dedifferentiated liposarcomas. *Virchows Arch* 2010;456(2):167-79.

63. Thway K, Jones RL, Noujaim J, et al. Dedifferentiated Liposarcoma: Updates on Morphology, Genetics, and Therapeutic Strategies. *Advances in Anatomic Pathology* 2016;23(1):30-40.

64. Cancer Genome Atlas Research Network. Electronic address edsc, Cancer Genome Atlas Research N. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell* 2017;171(4):950-65 e28.

65. Cairncross L, Snow HA, Strauss DC, et al. Diagnostic performance of MRI and histology in assessment of deep lipomatous tumours. *British Journal of Surgery* 2019;106(13):1794-99.

66. Thway K, Flora R, Shah C, et al. Diagnostic Utility of p16, CDK4, and MDM2 as an Immunohistochemical Panel in Distinguishing Well-differentiated and Dedifferentiated

Liposarcomas From Other Adipocytic Tumors. *The American Journal of Surgical Pathology* 2012;36(3).

67. Coindre J-M. Grading of Soft Tissue Sarcomas: Review and Update. *Archives of Pathology & Laboratory Medicine* 2006;130(10):1448-53.

68. von Mehren M, Kane JM, Agulnik M, et al. Soft Tissue Sarcoma, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology. *Journal of the National Comprehensive Cancer Network* 2022;20(7):815-33.

69. Knebel C, Lenze U, Pohlig F, et al. Prognostic factors and outcome of Liposarcoma patients: a retrospective evaluation over 15 years. *BMC Cancer* 2017;17(1):410.

70. Fujiwara T, Kaneuchi Y, Tsuda Y, et al. Low-grade soft-tissue sarcomas: What is an adequate margin for local disease control? *Surgical Oncology* 2020;35:303-08.

71. Zhou X-P, Xing J-P, Sun L-B, et al. Molecular characteristics and systemic treatment options of liposarcoma: A systematic review. *Biomedicine & Pharmacotherapy* 2024;178:117204.

72. Park JO, Qin LX, Prete FP, et al. Predicting outcome by growth rate of locally recurrent retroperitoneal liposarcoma: the one centimeter per month rule. *Ann Surg* 2009;250(6):977-82.

73. Gronchi A, Miceli R, Allard MA, et al. Personalizing the Approach to Retroperitoneal Soft Tissue Sarcoma: Histology-specific Patterns of Failure and Postrelapse Outcome after Primary Extended Resection. *Annals of Surgical Oncology* 2015;22(5):1447-54.

74. Crago AM, Dickson MA. Liposarcoma: Multimodality Management and Future Targeted Therapies. *Surg Oncol Clin N Am* 2016;25(4):761-73.

75. Masunaga T, Tsukamoto S, Honoki K, et al. Comparison of pre-operative and post-operative radiotherapy in patients with localized myxoid liposarcoma. *Japanese Journal of Clinical Oncology* 2023;53(12):1153-61.

76. Tfayli Y, Baydoun A, Naja AS, Saghieh S. Management of myxoid liposarcoma of the extremity. *Oncol Lett* 2021;22(2):596.

77. Lee LH, Tepper S, Owen G, et al. Radiotherapy, volume reduction, and short-term surgical outcomes in the treatment of large myxoid liposarcomas. *Radiat Oncol J* 2022;40(3):172-79.

78. Stacchiotti S, Van der Graaf WTA, Sanfilippo RG, et al. First-line chemotherapy in advanced intra-abdominal well-differentiated/dedifferentiated liposarcoma: An EORTC Soft Tissue and Bone Sarcoma Group retrospective analysis. *Cancer* 2022;128(15):2932-38.

79. Pan M, Zhou M, Xie L, et al. Recent advances in sarcoma therapy: new agents, strategies and predictive biomarkers. *Journal of Hematology & Oncology* 2024;17(1):124.

80. von Mehren M, Kane JM, Agulnik M, et al. Soft Tissue Sarcoma, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* 2022;20(7):815-33.

81. Gronchi A, Miah AB, Dei Tos AP, et al. Soft tissue and visceral sarcomas: ESMO&#x2013;EURACAN&#x2013;GENTURIS Clinical Practice Guidelines for diagnosis, treatment and follow-up<sup>&#x2606;</sup>. *Annals of Oncology* 2021;32(11):1348-65.

82. Deng H, Gao J, Xu X, et al. Predictors and outcomes of recurrent retroperitoneal liposarcoma: new insights into its recurrence patterns. *BMC Cancer* 2023;23(1):1076.

83. Yu Z-Y, Gao J-W, Liu N, et al. Predictive factors and a novel nomogram for recurrence of primary retroperitoneal liposarcoma: Comprehensive analysis of 128 cases. *Oncol Lett* 2023;25(6):257.

84. Mulita F, Verras GI, Liolis E, et al. Recurrent retroperitoneal liposarcoma: A case report and literature review. *Clin Case Rep* 2021;9(9):e04717.

85. Choi K-Y, Jost E, Mack L, Bouchard-Fortier A. Surgical management of truncal and extremities atypical lipomatous tumors/well-differentiated liposarcoma: A systematic review of the literature. *The American Journal of Surgery* 2020;219(5):823-27.

86. Ishii K, Yokoyama Y, Nishida Y, et al. Characteristics of primary and repeated recurrent retroperitoneal liposarcoma: outcomes after aggressive surgeries at a single institution. *Japanese Journal of Clinical Oncology* 2020;50(12):1412-18.

87. Olson CR, Suarez-Kelly LP, Ethun CG, et al. Resection Status Does Not Impact Recurrence in Well-Differentiated Liposarcoma of the Extremity. *Am Surg* 2021;87(11):1752-59.

88. Arvinius C, Torrecilla E, Beano-Collado J, et al. A clinical review of 11 cases of large-sized well-differentiated liposarcomas. *European Journal of Orthopaedic Surgery & Traumatology* 2017;27(6):837-41.

89. Byerly S, Chopra S, Nassif NA, et al. The role of margins in extremity soft tissue sarcoma. *J Surg Oncol* 2016;113(3):333-8.

90. Dalal KM, Antonescu CR, Singer S. Diagnosis and management of lipomatous tumors. *J Surg Oncol* 2008;97(4):298-313.

91. Callegaro D, Miceli R, Bonvalot S, et al. Development and external validation of two nomograms to predict overall survival and occurrence of distant metastases in adults after surgical resection of localised soft-tissue sarcomas of the extremities: a retrospective analysis. *Lancet Oncol* 2016;17(5):671-80.

92. Álvarez Álvarez R, Manzano A, Agra Pujol C, et al. Updated Review and Clinical Recommendations for the Diagnosis and Treatment of Patients with Retroperitoneal Sarcoma by the Spanish Sarcoma Research Group (GEIS). *Cancers (Basel)* 2023;15(12).

93. Jour G, Gullet A, Liu M, Hoch BL. Prognostic relevance of Fédération Nationale des Centres de Lutte Contre le Cancer grade and MDM2 amplification levels in dedifferentiated liposarcoma: a study of 50 cases. *Modern Pathology* 2015;28(1):37-47.

94. Vos M, Boeve WC, van Ginhoven TM, et al. Impact of primary tumor location on outcome of liposarcoma patients, a retrospective cohort study. *European Journal of Surgical Oncology* 2019;45(12):2437-42.

95. García-Ortega DY, Clara-Altamirano MA, Martín-Tellez KS, et al. Epidemiological profile of soft tissue sarcomas of the extremities: Incidence, histological subtypes, and primary sites. *Journal of Orthopaedics* 2021;25:70-74.

96. Lv X, Zhu L, Lan G, et al. A clinical tool to predict overall survival of elderly patients with soft tissue sarcoma after surgical resection. *Scientific Reports* 2024;14(1):15098.

97. Lazarides AL, Visgauss JD, Nussbaum DP, et al. Race is an independent predictor of survival in patients with soft tissue sarcoma of the extremities. *BMC Cancer* 2018;18(1):488.

98. Wan L, Tu C, Qi L, Li Z. Survivorship and prognostic factors for pleomorphic liposarcoma: a population-based study. *Journal of Orthopaedic Surgery and Research* 2021;16(1):175.

99. Vos M, Koseła-Paterczyk H, Rutkowski P, et al. Differences in recurrence and survival of extremity liposarcoma subtypes. *European Journal of Surgical Oncology* 2018;44(9):1391-97.

100. Wan L, Tu C, Qi L, Li Z. Survivorship and prognostic factors for pleomorphic liposarcoma: a population-based study. *J Orthop Surg Res* 2021;16(1):175.

101. Ren L, Qi Y, Zhao J, et al. Gender Differences in Prognosis After Primary Resection for Retroperitoneal Liposarcoma. *The American Surgeon™* 2024;90(4):575-84.

102. Pan M, Zhou MY, Jiang C, et al. Sex-dependent Prognosis of Patients with Advanced Soft Tissue Sarcoma. *Clin Cancer Res* 2024;30(2):413-19.

103. Daigeler A, Zmarsly I, Hirsch T, et al. Long-term outcome after local recurrence of soft tissue sarcoma: a retrospective analysis of factors predictive of survival in 135 patients with locally recurrent soft tissue sarcoma. *British Journal of Cancer* 2014;110(6):1456-64.

104. Abdelfatah E, Guzzetta AA, Nagarajan N, et al. Long-term outcomes in treatment of retroperitoneal sarcomas: A 15 year single-institution evaluation of prognostic features. *J Surg Oncol* 2016;114(1):56-64.

105. Gamboa AC, Gronchi A, Cardona K. Soft-tissue sarcoma in adults: An update on the current state of histiotype-specific management in an era of personalized medicine. *CA: A Cancer Journal for Clinicians* 2020;70(3):200-29.

106. Pisters PW, Leung DH, Woodruff J, et al. Analysis of prognostic factors in 1,041 patients with localized soft tissue sarcomas of the extremities. *Journal of Clinical Oncology* 1996;14(5):1679-89.

107. Wang S, Zhou Y, Wang H, Ling J. Survival analysis and treatment strategies for limb liposarcoma patients with metastasis at presentation. *Medicine (Baltimore)* 2021;100(13):e25296.

108. Tirumani SH, Tirumani H, Jagannathan JP, et al. Metastasis in dedifferentiated liposarcoma: Predictors and outcome in 148 patients. *Eur J Surg Oncol* 2015;41(7):899-904.

109. Mavrogenis AF, Lesensky J, Romagnoli C, et al. Atypical lipomatous tumors/well-differentiated liposarcomas: clinical outcome of 67 patients. *Orthopedics* 2011;34(12):e893-8.

110. Mussi C, Collini P, Miceli R, et al. The prognostic impact of dedifferentiation in retroperitoneal liposarcoma: a series of surgically treated patients at a single institution. *Cancer* 2008;113(7):1657-65.

111. Binh MB, Sastre-Garau X, Guillou L, et al. MDM2 and CDK4 immunostainings are useful adjuncts in diagnosing well-differentiated and dedifferentiated liposarcoma subtypes: a comparative analysis of 559 soft tissue neoplasms with genetic data. *Am J Surg Pathol* 2005;29(10):1340-7.

112. Conyers R, Young S, Thomas DM. Liposarcoma: Molecular Genetics and Therapeutics. *Sarcoma* 2011;2011:483154.

113. Italiano A, Bianchini L, Keslair F, et al. HMGA2 is the partner of MDM2 in well-differentiated and dedifferentiated liposarcomas whereas CDK4 belongs to a distinct inconsistent amplicon. *Int J Cancer* 2008;122(10):2233-41.

114. Hirata M, Asano N, Katayama K, et al. Integrated exome and RNA sequencing of dedifferentiated liposarcoma. *Nat Commun* 2019;10(1):5683.

115. Wang X, Asmann YW, Erickson-Johnson MR, et al. High-resolution genomic mapping reveals consistent amplification of the fibroblast growth factor receptor substrate 2 gene in well-differentiated and dedifferentiated liposarcoma. *Genes Chromosomes Cancer* 2011;50(11):849-58.

116. Sciot R. MDM2 Amplified Sarcomas: A Literature Review. *Diagnostics (Basel)* 2021;11(3).

117. Nag S, Qin J, Srivenugopal KS, et al. The MDM2-p53 pathway revisited. *J Biomed Res* 2013;27(4):254-71.

118. Levine AJ. The many faces of p53: something for everyone. *J Mol Cell Biol* 2019;11(7):524-30.

119. Duffy MJ, Synnott NC, O'Grady S, Crown J. Targeting p53 for the treatment of cancer. *Semin Cancer Biol* 2022;79:58-67.

120. Oliner JD, Saiki AY, Caenepeel S. The Role of MDM2 Amplification and Overexpression in Tumorigenesis. *Cold Spring Harb Perspect Med* 2016;6(6).

121. Bill KLJ, Seligson ND, Hays JL, et al. Degree of MDM2 Amplification Affects Clinical Outcomes in Dedifferentiated Liposarcoma. *Oncologist* 2019;24(7):989-96.

122. Mansoori B, Mohammadi A, Ditzel HJ, et al. HMGA2 as a Critical Regulator in Cancer Development. *Genes (Basel)* 2021;12(2).

123. Narita M, Narita M, Krizhanovsky V, et al. A Novel Role for High-Mobility Group A Proteins in Cellular Senescence and Heterochromatin Formation. *Cell* 2006;126(3):503-14.

124. Yamashita K, Kohashi K, Yamada Y, et al. Prognostic significance of the MDM2/HMGA2 ratio and histological tumor grade in dedifferentiated liposarcoma. *Genes, Chromosomes and Cancer* 2021;60(1):26-37.

125. Beird HC, Wu CC, Ingram DR, et al. Genomic profiling of dedifferentiated liposarcoma compared to matched well-differentiated liposarcoma reveals higher genomic complexity and a common origin. *Cold Spring Harb Mol Case Stud* 2018;4(2).

126. Saada-Bouzid E, Burel-Vandenbos F, Ranchere-Vince D, et al. Prognostic value of HMGA2, CDK4, and JUN amplification in well-differentiated and dedifferentiated liposarcomas. *Mod Pathol* 2015;28(11):1404-14.

127. Denis CJ, Deiteren K, Hendriks D, et al. Carboxypeptidase M in apoptosis, adipogenesis and cancer. *Clin Chim Acta* 2013;415:306-16.

128. Bujalska IJ, Quinkler M, Tomlinson JW, et al. Expression profiling of 11beta-hydroxysteroid dehydrogenase type-1 and glucocorticoid-target genes in subcutaneous and omental human preadipocytes. *J Mol Endocrinol* 2006;37(2):327-40.

129. Kanojia D, Nagata Y, Garg M, et al. Genomic landscape of liposarcoma. *Oncotarget* 2015;6(40):42429-44.

130. Uribe ML, Marrocco I, Yarden Y. EGFR in Cancer: Signaling Mechanisms, Drugs, and Acquired Resistance. *Cancers (Basel)* 2021;13(11).

131. Yang JL, Gupta RD, Goldstein D, Crowe PJ. Significance of Phosphorylated Epidermal Growth Factor Receptor and Its Signal Transducers in Human Soft Tissue Sarcoma. *Int J Mol Sci* 2017;18(6).

132. Park JH, Roeder RG. GAS41 is required for repression of the p53 tumor suppressor pathway during normal cellular proliferation. *Mol Cell Biol* 2006;26(11):4006-16.

133. Barretina J, Taylor BS, Banerji S, et al. Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nat Genet* 2010;42(8):715-21.

134. Barretina J, Taylor BS, Banerji S, et al. Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nature Genetics* 2010;42(8):715-21.

135. Italiano A, Bianchini L, Keslair F, et al. HMGA2 is the partner of MDM2 in well-differentiated and dedifferentiated liposarcomas whereas CDK4 belongs to a distinct inconsistent amplicon. *Int J Cancer* 2008;122(10):2233-41.

136. Ricciotti RW, Baraff AJ, Jour G, et al. High amplification levels of MDM2 and CDK4 correlate with poor outcome in patients with dedifferentiated liposarcoma: A cytogenomic microarray analysis of 47 cases. *Cancer Genet* 2017;218-219:69-80.

137. Zhang K, Chu K, Wu X, et al. Amplification of FRS2 and activation of FGFR/FRS2 signaling pathway in high-grade liposarcoma. *Cancer Res* 2013;73(4):1298-307.

138. Brooks AN, Kilgour E, Smith PD. Molecular pathways: fibroblast growth factor signaling: a new therapeutic opportunity in cancer. *Clin Cancer Res* 2012;18(7):1855-62.

139. Asano N, Yoshida A, Mitani S, et al. Frequent amplification of receptor tyrosine kinase genes in welldifferentiated/ dedifferentiated liposarcoma. *Oncotarget* 2017;8(8):12941-52.

140. Dickson MA. Molecular pathways: CDK4 inhibitors for cancer therapy. *Clin Cancer Res* 2014;20(13):3379-83.

141. Creytens D, Van Gorp J, Speel EJ, Ferdinande L. Characterization of the 12q amplicons in lipomatous soft tissue tumors by multiplex ligation-dependent probe amplification-based copy number analysis. *Anticancer Res* 2015;35(4):1835-42.

142. Wang J, Zhou Y, Li D, et al. TSPAN31 is a critical regulator on transduction of survival and apoptotic signals in hepatocellular carcinoma cells. *FEBS Lett* 2017;591(18):2905-18.

143. Gruel N, Quignot C, Lesage L, et al. Cellular origin and clonal evolution of human dedifferentiated liposarcoma. *Nature Communications* 2024;15(1):7941.

144. Yoo Y, Park SY, Jo EB, et al. Overexpression of Replication-Dependent Histone Signifies a Subset of Dedifferentiated Liposarcoma with Increased Aggressiveness. *Cancers (Basel)* 2021;13(13).

145. Crago AM, Socci ND, DeCarolis P, et al. Copy number losses define subgroups of dedifferentiated liposarcoma with poor prognosis and genomic instability. *Clin Cancer Res* 2012;18(5):1334-40.

146. Tap WD, Eilber FC, Ginther C, et al. Evaluation of well-differentiated/de-differentiated liposarcomas by high-resolution oligonucleotide array-based comparative genomic hybridization. *Genes Chromosomes Cancer* 2011;50(2):95-112.

147. Mariani O, Brennetot C, Coindre JM, et al. JUN oncogene amplification and overexpression block adipocytic differentiation in highly aggressive sarcomas. *Cancer Cell* 2007;11(4):361-74.

148. Peng T, Zhang P, Liu J, et al. An experimental model for the study of well-differentiated and dedifferentiated liposarcoma; deregulation of targetable tyrosine kinase receptors. *Laboratory Investigation* 2011;91(3):392-403.

149. Horvai AE, DeVries S, Roy R, et al. Similarity in genetic alterations between paired well-differentiated and dedifferentiated components of dedifferentiated liposarcoma. *Modern Pathology* 2009;22(11):1477-88.

150. Snyder EL, Sandstrom DJ, Law K, et al. c-Jun amplification and overexpression are oncogenic in liposarcoma but not always sufficient to inhibit the adipocytic differentiation programme. *J Pathol* 2009;218(3):292-300.

151. Chibon F, Mariani O, Derre J, et al. ASK1 (MAP3K5) as a potential therapeutic target in malignant fibrous histiocytomas with 12q14-q15 and 6q23 amplifications. *Genes Chromosomes Cancer* 2004;40(1):32-7.

152. Nygaard G, Di Paolo JA, Hammaker D, et al. Regulation and function of apoptosis signal-regulating kinase 1 in rheumatoid arthritis. *Biochem Pharmacol* 2018;151:282-90.

153. Ahmadian M, Suh JM, Hah N, et al. PPARgamma signaling and metabolism: the good, the bad and the future. *Nat Med* 2013;19(5):557-66.

154. Siersbaek R, Nielsen R, Mandrup S. PPARgamma in adipocyte differentiation and metabolism--novel insights from genome-wide studies. *FEBS Lett* 2010;584(15):3242-9.

155. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;6(269):pl1.

156. Garsed DW, Marshall OJ, Corbin VD, et al. The architecture and evolution of cancer neochromosomes. *Cancer Cell* 2014;26(5):653-67.

157. Waterfall Joshua J, Meltzer Paul S. Building through Breaking: The Development of Cancer Neochromosomes. *Cancer Cell* 2014;26(5):593-95.

158. Chai H, Xu F, DiAdamo A, et al. Cytogenomic Characterization of Giant Ring or Rod Marker Chromosome in Four Cases of Well-Differentiated and Dedifferentiated Liposarcoma. *Case Rep Genet* 2022;2022:6341207.

159. Amin-Mansour A, George S, Sioletic S, et al. Genomic Evolutionary Patterns of Leiomyosarcoma and Liposarcoma. *Clin Cancer Res* 2019;25(16):5135-42.

160. Chen TWW, Sanfilippo R, Jones RL, et al. 76MO Efficacy and safety findings from MANTRA: A global, randomized, multicenter, phase III study of the MDM2 inhibitor milademetan vs trabectedin in patients with dedifferentiated liposarcomas. *Annals of Oncology* 2023;34:S1496.

161. Bauer TM, Gounder MM, Weise AM, et al. A phase 1 study of MDM2 inhibitor DS-3032b in patients with well/de-differentiated liposarcoma (WD/DD LPS), solid tumors (ST) and lymphomas (L). *Journal of Clinical Oncology* 2018;36(15_suppl):11514-14.

162. Gounder MM, Bauer TM, Schwartz GK, et al. 7LBA Late Breaking - Milademetan, an oral MDM2 inhibitor, in well-differentiated/ dedifferentiated liposarcoma: results from a phase 1 study in patients with solid tumors or lymphomas. *European Journal of Cancer* 2020;138:S3-S4.

163. McKean M, Tolcher AW, Reeves JA, et al. Newly updated activity results of alrizomadlin (APG-115), a novel MDM2/p53 inhibitor, plus pembrolizumab: Phase 2 study in adults and children with various solid tumors. *Journal of Clinical Oncology* 2022;40(16_suppl):9517-17.

164. Saleh MN, Patel MR, Bauer TM, et al. Phase 1 Trial of ALRN-6924, a Dual Inhibitor of MDMX and MDM2, in Patients with Solid Tumors and Lymphomas Bearing Wild-type TP53. *Clinical Cancer Research* 2021;27(19):5236-47.

165. Stein EM, DeAngelo DJ, Chromik J, et al. Results from a First-in-Human Phase I Study of Siremadlin (HDM201) in Patients with Advanced Wild-Type TP53 Solid Tumors and Acute Leukemia. *Clin Cancer Res* 2022;28(5):870-81.

166. Schöffski P, Mehdi L, Anthony L, and Maki RG. Brightline-1: phase II/III trial of the MDM2–p53 antagonist BI 907828 versus doxorubicin in patients with advanced DDLPS. *Future Oncology* 2023;19(9):621-29.

167. LoRusso P, Yamamoto N, Patel MR, et al. The MDM2-p53 Antagonist Brigimadlin (BI 907828) in Patients with Advanced or Metastatic Solid Tumors: Results of a Phase Ia, First-in-Human, Dose-Escalation Study. *Cancer Discov* 2023;13(8):1802-13.

168. Dickson MA, Koff A, D'Angelo SP, et al. Phase 2 study of the CDK4 inhibitor abemaciclib in dedifferentiated liposarcoma. *Journal of Clinical Oncology* 2019;37(15_suppl):11004-04.

169. Dickson MA, Ballman KV, Weiss MC, et al. SARC041: A phase 3 randomized double-blind study of abemaciclib versus placebo in patients with advanced dedifferentiated liposarcoma. *Journal of Clinical Oncology* 2023;41(16_suppl):TPS11587-TPS87.

170. Gleason CE, Dickson MA, Klein Dooley ME, et al. Therapy-Induced Senescence Contributes to the Efficacy of Abemaciclib in Patients with Dedifferentiated Liposarcoma. *Clin Cancer Res* 2024;30(4):703-18.

171. Dickson MA, Schwartz GK, Keohan ML, et al. Progression-Free Survival Among Patients With Well-Differentiated or Dedifferentiated Liposarcoma Treated With CDK4 Inhibitor Palbociclib: A Phase 2 Clinical Trial. *JAMA Oncology* 2016;2(7):937-40.

172. Movva S, Matloob S, Handorf EA, et al. SAR-096: Phase II Clinical Trial of Ribociclib in Combination with Everolimus in Advanced Dedifferentiated Liposarcoma (DDL) and Leiomyosarcoma (LMS). *Clinical Cancer Research* 2024;30(2):315-22.

173. Wilky BA, Trucco MM, Subhawong TK, et al. Axitinib plus pembrolizumab in patients with advanced sarcomas including alveolar soft-part sarcoma: a single-centre, single-arm, phase 2 trial. *Lancet Oncol* 2019;20(6):837-48.

174. D'Angelo SP, Mahoney MR, Van Tine BA, et al. Nivolumab with or without ipilimumab treatment for metastatic sarcoma (Alliance A091401): two open-label, non-comparative, randomised, phase 2 trials. *The Lancet Oncology* 2018;19(3):416-26.

175. Zhou M, Bui N, Bolleddu S, et al. Nivolumab plus ipilimumab for soft tissue sarcoma: a single institution retrospective review. *Immunotherapy* 2020;12(18):1303-12.

176. Verret B, Bellera C, Boudou Rouquette P, et al. 1719O Multisarc: A randomized precision medicine study in advanced soft-tissue sarcomas. *Annals of Oncology* 2024;35:S1030-S31.

177. Italiano A, Dinart D, Soubeyran I, et al. Molecular profiling of advanced soft-tissue sarcomas: the MULTISARC randomized trial. *BMC Cancer* 2021;21(1):1180.

178. Derrible S, Kennedy C. Applications of Graph Theory and Network Science to Transit Network Design. *Transport Reviews* 2011;31(4):495-519.

179. Sadri AM, Hasan S, Ukkusuri SV. Joint inference of user community and interest patterns in social interaction networks. *Social Network Analysis and Mining* 2019;9(1):11.

180. . *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS).*

181. Silverman EK, Schmidt H, Anastasiadou E, et al. Molecular networks in Network Medicine: Development and applications. *Wiley Interdiscip Rev Syst Biol Med* 2020;12(6):e1489.

182. Pe'er D, Hacohen N. Principles and strategies for developing network models in cancer. *Cell* 2011;144(6):864-73.

183. Seo H, Kim W, Lee J, Youn B. Network-based approaches for anticancer therapy (Review). *Int J Oncol* 2013;43(6):1737-44.

184. Zhang W, Chien J, Yong J, Kuang R. Network-based machine learning and graph theory algorithms for precision oncology. *npj Precision Oncology* 2017;1(1):25.

185. Zhang W, Ma J, Ideker T. Classifying tumors by supervised network propagation. *Bioinformatics* 2018;34(13):i484-i93.

186. MotieGhader H, Tabrizi-Nezhadi P, Deldar Abad Paskeh M, et al. Drug repositioning in non-small cell lung cancer (NSCLC) using gene co-expression and drug-gene interaction networks analysis. *Sci Rep* 2022;12(1):9417.

187. Bidkhori G, Benfeitas R, Klevstig M, et al. Metabolic network-based stratification of hepatocellular carcinoma reveals three distinct tumor subtypes. *Proc Natl Acad Sci U S A* 2018;115(50):E11874-E83.

188. Hofree M, Shen JP, Carter H, et al. Network-based stratification of tumor mutations. *Nature Methods* 2013;10(11):1108-15.

189. Dimitrakopoulos C, Hindupur SK, Häfliger L, et al. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics (Oxford, England)* 2018;34(14):2441-48.

190. Shi K, Gao L, Wang B. Discovering potential cancer driver genes by an integrated network-based approach. *Mol Biosyst* 2016;12(9):2921-31.

191. Yu D, Kim M, Xiao G, Hwang TH. Review of biological network data and its applications. *Genomics & informatics* 2013;11(4):200-10.

192. Ozturk K, Dow M, Carlin DE, et al. The Emerging Potential for Network Analysis to Inform Precision Cancer Medicine. *Journal of Molecular Biology* 2018;430(18, Part A):2875-99.

193. Sun K, Gonçalves JP, Larminie C, Pržulj N. Predicting disease associations via biological network analysis. *BMC Bioinformatics* 2014;15(1):304.

194. Tan A, Huang H, Zhang P, Li S. Network-based cancer precision medicine: A new emerging paradigm. *Cancer Letters* 2019;458:39-45.

195. Koutrouli M, Karatzas E, Paez-Espino D, Pavlopoulos GA. A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in Bioengineering and Biotechnology* 2020;8.

196. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005;4:Article17.

197. Kivelä M, Arenas A, Barthelemy M, et al. Multilayer networks. *Journal of Complex Networks* 2014;2(3):203-71.

198. Newman M, Newman M. Mathematics of networks *Networks*: Oxford University Press; 2018 p0.

199. Newman M, Newman M. Chapter 1: Introduction *Networks*: Oxford University Press; 2018 p0.

200. Alm E, Arkin AP. Biological networks. *Current Opinion in Structural Biology* 2003;13(2):193-202.

201. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;5(2):101-13.

202. Lesne A. Complex Networks: from Graph Theory to Biology. *Letters in Mathematical Physics* 2006;78(3):235-62.

203. Zhang W. *Fundamentals of Network Biology*: World Scientific; 2018.

204. Newman M, Newman M. The structure of real-world networks *Networks*: Oxford University Press; 2018 p0.

205. Holme P. Rare and everywhere: Perspectives on scale-free networks. *Nat Commun* 2019;10(1):1016.

206. Khanin R, Wit E. How scale-free are biological networks. *J Comput Biol* 2006;13(3):810-8.

207. Charitou T, Bryan K, Lynn DJ. Using biological networks to integrate, visualize and analyze genomics data. *Genet Sel Evol* 2016;48:27.

208. He X, Zhang J. Why do hubs tend to be essential in protein networks? *PLoS Genet* 2006;2(6):e88.

209. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature* 2001;411(6833):41-2.

210. Koschützki D, Schreiber F. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul Syst Bio* 2008;2:193-201.

211. Koutrouli M, Karatzas E, Paez-Espino D, Pavlopoulos GA. A Guide to Conquer the Biological Network Era Using Graph Theory. *Front Bioeng Biotechnol* 2020;8:34.

212. Rodrigues FA. Network Centrality: An Introduction. In: Macau EEN (ed.) *A Mathematical Modeling Approach from Nonlinear Dynamics to Complex Systems*. Cham: Springer International Publishing; 2019 p177-96.

213. Ashtiani M, Salehzadeh-Yazdi A, Razaghi-Moghadam Z, et al. A systematic survey of centrality measures for protein-protein interaction networks. *BMC Systems Biology* 2018;12(1):80.

214. Jalili M, Salehzadeh-Yazdi A, Gupta S, et al. Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks. *Front Physiol* 2016;7:375.

215. Saito R, Smoot ME, Ono K, et al. A travel guide to Cytoscape plugins. *Nat Methods* 2012;9(11):1069-76.

216. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;13(11):2498-504.

217. Hakes L, Pinney JW, Robertson DL, Lovell SC. Protein-protein interaction networks and biology—what's the connection? *Nature Biotechnology* 2008;26(1):69-72.

218. Ideker T, Sharan R. Protein networks in disease. *Genome Res* 2008;18(4):644-52.

219. Rual JF, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005;437(7062):1173-8.

220. Rasti S, Vogiatzis C. A survey of computational methods in protein–protein interaction networks. *Annals of Operations Research* 2019;276(1-2):35-87.

221. Spirin V, Mirny LA. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A* 2003;100(21):12123-8.

222. Zhou M, Li Q, Wang R. Current Experimental Methods for Characterizing Protein-Protein Interactions. *ChemMedChem* 2016;11(8):738-56.

223. Zhang Y, Fonslow BR, Shan B, et al. Protein analysis by shotgun/bottom-up proteomics. *Chem Rev* 2013;113(4):2343-94.

224. Xing S, Wallmeroth N, Berendzen KW, Grefen C. Techniques for the Analysis of Protein-Protein Interactions in Vivo. *Plant Physiol* 2016;171(2):727-58.

225. Rao VS, Srinivas K, Sujini GN, Kumar GN. Protein-protein interaction detection: methods and analysis. *Int J Proteomics* 2014;2014:147648.

226. Peng X, Wang J, Peng W, et al. Protein-protein interactions: detection, reliability assessment and applications. *Brief Bioinform* 2017;18(5):798-819.

227. Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;49(D1):D605-D12.

228. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* 2005;4(1).

229. Lachmann A, Giorgi FM, Lopez G, Califano A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* 2016;32(14):2233-5.

230. Reverter A, Chan EKF. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics* 2008;24(21):2491-97.

231. Perkins AD, Langston MA. Threshold selection in gene co-expression networks using spectral graph theory techniques. *BMC Bioinformatics* 2009;10 Suppl 11(Suppl 11):S4.

232. Lee HK, Hsu AK, Sajdak J, et al. Coexpression analysis of human genes across many microarray data sets. *Genome Res* 2004;14(6):1085-94.

233. Garcia-Ruiz S, Gil-Martinez AL, Cisterna A, et al. CoExp: A Web Tool for the Exploitation of Co-expression Networks. *Front Genet* 2021;12:630187.

234. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9(1):559.

235. Morselli Gysi D, Voigt A, Fragoso T, et al. wTO: An R package for computing weighted topological overlap and a consensus network with integrated visualization tool. *BMC Bioinformatics* 2018;19.

236. Lemoine GG, Scott-Boyer M-P, Ambroise B, et al. GWENA: gene co-expression networks analysis and extended modules characterization in a single Bioconductor package. *BMC Bioinformatics* 2021;22(1):267.

237. Kuang J, Michel K, Scoglio C. GeCoNet-Tool: a software package for gene co-expression network construction and analysis. *BMC Bioinformatics* 2023;24(1):281.

238. Zhai X, Xue Q, Liu Q, et al. Colon cancer recurrenceassociated genes revealed by WGCNA coexpression network analysis. *Mol Med Rep* 2017;16(5):6499-505.

239. Zhu Z, Jin Z, Deng Y, et al. Co-expression Network Analysis Identifies Four Hub Genes Associated With Prognosis in Soft Tissue Sarcoma. *Frontiers in Genetics* 2019;10(37).

240. Xia WX, Yu Q, Li GH, et al. Identification of four hub genes associated with adrenocortical carcinoma progression by WGCNA. *PeerJ* 2019;7:e6555.

241. Di Y, Chen D, Yu W, Yan L. Bladder cancer stage-associated hub genes revealed by WGCNA co-expression network analysis. *Hereditas* 2019;156:7.

242. Su R, Jin C, Zhou L, et al. Construction of a ceRNA network of hub genes affecting immune infiltration in ovarian cancer identified by WGCNA. *BMC Cancer* 2021;21(1):970.

243. Zhou J, Guo H, Liu L, et al. Construction of co-expression modules related to survival by WGCNA and identification of potential prognostic biomarkers in glioblastoma. *J Cell Mol Med* 2021;25(3):1633-44.

244. Han Z, Ren H, Sun J, et al. Integrated weighted gene coexpression network analysis identifies Frizzled 2 (FZD2) as a key gene in invasive malignant pleomorphic adenoma. *J Transl Med* 2022;20(1):15.

245. Wang L, Liu X, Liu Z, et al. Network models of prostate cancer immune microenvironments identify ROMO1 as heterogeneity and prognostic marker. *Scientific Reports* 2022;12(1):192.

246. Yang M, He H, Peng T, et al. Identification of 9 Gene Signatures by WGCNA to Predict Prognosis for Colon Adenocarcinoma. *Comput Intell Neurosci* 2022;2022:8598046.

247. Xiang J, Gao L, Jing HY, et al. Construction of CeRNA regulatory network based on WGCNA reveals diagnosis biomarkers for colorectal cancer. *BMC Cancer* 2022;22(1):991.

248. Zhang H, Lin Y, Zhuang M, et al. Screening and identification of CNIH4 gene associated with cell proliferation in gastric cancer based on a large-scale CRISPR-Cas9 screening database DepMap. *Gene* 2023;850:146961.

249. Wang Z, Tao P, Fan P, et al. Insight of a lipid metabolism prognostic model to identify immune landscape and potential target for retroperitoneal liposarcoma. *Frontiers in Immunology* 2023;14.

250. Langfelder P, Horvath S. Fast R Functions for Robust Correlations and Hierarchical Clustering. *J Stat Softw* 2012;46(11).

251. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol* 2007;1:54.

252. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? *PLoS Comput Biol* 2011;7(1):e1001057.

253. Li J, Zhou D, Qiu W, et al. Application of Weighted Gene Co-expression Network Analysis for Data from Paired Design. *Scientific Reports* 2018;8(1):622.

254. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2008;24(5):719-20.

255. Gaiteri C, Ding Y, French B, et al. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. *Genes Brain Behav* 2014;13(1):13-24.

256. Zhao W, Langfelder P, Fuller T, et al. Weighted Gene Coexpression Network Analysis: State of the Art. *Journal of Biopharmaceutical Statistics* 2010;20(2):281-300.

257. Deng Y, Jiang Y-H, Yang Y, et al. Molecular ecological network analyses. *BMC Bioinformatics* 2012;13(1):113.

258. Zoppi J, Guillaume JF, Neunlist M, Chaffron S. MiBiOmics: an interactive web application for multi-omics data exploration and integration. *BMC Bioinformatics* 2021;22(1):6.

259. Sanchez-Baizan N, Ribas L, Piferrer F. Improved biomarker discovery through a plot twist in transcriptomic data analysis. *BMC Biol* 2022;20(1):208.

260. Guo W, Calixto CPG, Tzioutziou N, et al. Evaluation and improvement of the regulatory inference for large co-expression networks with limited sample size. *BMC Syst Biol* 2017;11(1):62.

261. Long T, Liu Z, Zhou X, et al. Identification of differentially expressed genes and enriched pathways in lung cancer using bioinformatics analysis. *Mol Med Rep* 2019;19(3):2029-40.

262. Yu H, Pei D, Chen L, et al. Identification of key genes and molecular mechanisms associated with dedifferentiated liposarcoma based on bioinformatic methods. *OncoTargets and therapy* 2017;10:3017-27.

263. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 2013;14(1):91.

264. Vahabi N, Michailidis G. Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review. *Front Genet* 2022;13:854752.

265. Sánchez-Baizán N, Ribas L, Piferrer F. Improved biomarker discovery through a plot twist in transcriptomic data analysis. *BMC Biology* 2022;20(1):208.

266. Feng S, Xu Y, Dai Z, et al. Integrative Analysis From Multicenter Studies Identifies a WGCNA-Derived Cancer-Associated Fibroblast Signature for Ovarian Cancer. *Front Immunol* 2022;13:951582.

267. Li D, Wang L, Wang G, et al. Weighted Gene Co-Expression Network Analysis Reveals a New Survival Model for Prognostic Prediction in Ewing Sarcoma. 2021.

268. Liao Y, Wang Y, Cheng M, et al. Weighted Gene Coexpression Network Analysis of Features That Control Cancer Stem Cells Reveals Prognostic Biomarkers in Lung Adenocarcinoma. *Frontiers in Genetics* 2020;11(311).

269. Mohr T, Katz S, Paulitschke V, et al. Systematic Analysis of the Transcriptome Profiles and Co-Expression Networks of Tumour Endothelial Cells Identifies Several Tumour-Associated Modules and Potential Therapeutic Targets in Hepatocellular Carcinoma. *Cancers (Basel)* 2021;13(8).

270. van Dam S, Võsa U, van der Graaf A, et al. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics* 2017;19(4):575-92.

271. Hammoud Z, Kramer F. Multilayer networks: aspects, implementations, and application in biomedicine. *Big Data Analytics* 2020;5(1):2.

272. De Domenico M, Porter MA, Arenas A. MuxViz: a tool for multilayer analysis and visualization of networks. *Journal of Complex Networks* 2014;3(2):159-76.

273. Didier G, Brun C, Baudot A. Identifying communities from multiplex biological networks. *PeerJ* 2015;3:e1525.

274. Cen Y, Zou X, Zhang J, et al. Representation Learning for Attributed Multiplex Heterogeneous Network *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA*: Association for Computing Machinery; 2019.

275. Zhang Y, Chen J, Wang Y, et al. Multilayer network analysis of miRNA and protein expression profiles in breast cancer patients. *PLoS One* 2019;14(4):e0202311.

276. Mahapatra S, Bhuyan R, Das J, Swarnkar T. Integrated multiplex network based approach for hub gene identification in oral cancer. *Heliyon* 2021;7(7):e07418.

277. Wang H, Zheng H, Wang J, et al. Integrating Omic Data with a Multiplex Network-based Approach for the Identification of Cancer Subtypes. *IEEE Trans Nanobioscience* 2016;15(4):335-42.

278. Di Nanni N, Bersanelli M, Milanesi L, Mosca E. Network Diffusion Promotes the Integrative Analysis of Multiple Omics. *Front Genet* 2020;11:106.

279. Bersanelli M, Mosca E, Remondini D, et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 2016;17 Suppl 2(Suppl 2):15-15.

280. Bersanelli M, Mosca E, Remondini D, et al. Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. *Sci Rep* 2016;6:34841.

281. Bersanelli M, Mosca E, Remondini D, et al. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 2016;17 Suppl 2:15.

282. Valdeolivas A, Tichit L, Navarro C, et al. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics* 2018;35(3):497-505.

283. Wen Y, Song X, Yan B, et al. Multi-dimensional data integration algorithm based on random walk with restart. *BMC Bioinformatics* 2021;22(1):97.

284. Kim SY, Jeong H-H, Kim J, et al. Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies. *Biology Direct* 2019;14(1):8.

285. Jiao Y, Lawler K, Patel GS, et al. DART: Denoising Algorithm based on Relevance network Topology improves molecular pathway activity inference. *BMC Bioinformatics* 2011;12:403.

286. Liu W, Li C, Xu Y, et al. Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics* 2013;29(17):2169-77.

287. Kim SY, Kim TR, Jeong H-H, Sohn K-A. Integrative pathway-based survival prediction utilizing the interaction between gene expression and DNA methylation in breast cancer. *BMC Medical Genomics* 2018;11(Suppl 3):68-68.

288. Kim SY, Choe EK, Shivakumar M, et al. Multi-layered network-based pathway activity inference using directed random walks: application to predicting clinical outcomes in urologic cancer. *bioRxiv* 2020:2020.07.22.163949.

289. Liu J, Li R, Liao X, Jiang W. Comprehensive Bioinformatic Analysis Genes Associated to the Prognosis of Liposarcoma. *Medical science monitor : international medical journal of experimental and clinical research* 2018;24:7329-39.

290. You Y, Lai X, Pan Y, et al. Artificial intelligence in cancer target identification and drug discovery. *Signal Transduction and Targeted Therapy* 2022;7(1):156.

291. Huanhuan Z, Guochuan Z. Identification of Differentiation-Related Biomarkers in Liposarcoma Tissues Using Weighted Gene Co-Expression Network Analysis. *Journal of Biological Regulators and Homeostatic Agents* 2023;37(12):6807-19.

292. Zhao Y, Qin D, Li X, et al. Identification of NINJ1 as a novel prognostic predictor for retroperitoneal liposarcoma. *Discover Oncology* 2024;15(1):155.

293. Shen J, Chen R, Duan S. NINJ1: Bridging lytic cell death and inflammation therapy. *Cell Death & Disease* 2024;15(11):831.

294. Jennewein C, Sowa R, Faber AC, et al. Contribution of Ninjurin1 to Toll-like receptor 4 signaling and systemic inflammation. *Am J Respir Cell Mol Biol* 2015;53(5):656-63.

295. Long Z, Wu T, Tian Q, et al. Expression and prognosis analyses of BUB1, BUB1B and BUB3 in human sarcoma. *Aging (Albany NY)* 2021;13(9):12395-409.

296. Zhang S, Yan L, Cui C, et al. Downregulation of RRM2 Attenuates Retroperitoneal Liposarcoma Progression via the Akt/mTOR/4EBP1 Pathway: Clinical, Biological, and Therapeutic Significance. *Onco Targets Ther* 2020;13:6523-37.

297. Chen J, Lian Y, Zhao B, et al. Deciphering the Prognostic and Therapeutic Significance of Cell Cycle Regulator CENPF: A Potential Biomarker of Prognosis and Immune Microenvironment for Patients with Liposarcoma. *Int J Mol Sci* 2023;24(8).

298. Zhu Z, Jin Z, Zhang H, et al. Knockdown of Kif20a inhibits growth of tumors in soft tissue sarcoma in vitro and in vivo. *J Cancer* 2020;11(17):5088-98.

299. Pandey P, Sliker B, Peters HL, et al. Amyloid precursor protein and amyloid precursor-like protein 2 in cancer. *Oncotarget* 2016;7(15):19430-44.

300. Poelaert BJ, Knoche SM, Larson AC, et al. Amyloid Precursor-like Protein 2 Expression Increases during Pancreatic Cancer Development and Shortens the Survival of a Spontaneous Mouse Model of Pancreatic Cancer. *Cancers* 2021;13(7):1535.

301. Takayama K, Tsutsumi S, Suzuki T, et al. Amyloid precursor protein is a primary androgen target gene that promotes prostate cancer growth. *Cancer Res* 2009;69(1):137-42.

302. Wu X, Chen S, Lu C. Amyloid precursor protein promotes the migration and invasion of breast cancer cells by regulating the MAPK signaling pathway. *Int J Mol Med* 2020;45(1):162-74.

303. Miyazaki T, Ikeda K, Horie-Inoue K, Inoue S. Amyloid precursor protein regulates migration and metalloproteinase gene expression in prostate cancer cells. *Biochem Biophys Res Commun* 2014;452(3):828-33.

304. Uddin MS, Kabir MT, Jeandet P, et al. Novel Anti-Alzheimer's Therapeutic Molecules Targeting Amyloid Precursor Protein Processing. *Oxid Med Cell Longev* 2020;2020:7039138.

305. Bai C, Li S, Tan Z, Fan Z. Targeting MCM2 activates cancer-associated fibroblasts-like phenotype and affects chemo-resistance of liposarcoma cells against doxorubicin. *Anti-Cancer Drugs* 2024;35(10):883-92.

306. Yin K, Zhang Y, Zhang S, et al. Using weighted gene co-expression network analysis to identify key modules and hub genes in tongue squamous cell carcinoma. *Medicine (Baltimore)* 2019;98(37):e17100.

307. Lin Y, Wang S, Yang Q. Identification of hub genes and diagnostic efficacy for triple-negative breast cancer through WGCNA and Mendelian randomization. *Discover Oncology* 2024;15(1):117.

308. Zhao R, Wei W, Zhen L. WGCNA-based identification of potential targets and pathways in response to treatment in locally advanced breast cancer patients. *Open Medicine* 2023;18(1).

309. Chen W, Kang Y, Sheng W, et al. A new 4-gene-based prognostic model accurately predicts breast cancer prognosis and immunotherapy response by integrating WGCNA and bioinformatics analysis. *Frontiers in Immunology* 2024;15.

310. Sun X, Wang Z, Chen X, Shen K. CRISPR-cas9 Screening Identified Lethal Genes Enriched in Cell Cycle Pathway and of Prognosis Significance in Breast Cancer. *Front Cell Dev Biol* 2021;9:646774.

311. Manouchehri L, Zinati Z, Nazari L. Population-Specific gene expression profiles in prostate cancer: insights from Weighted Gene Co-expression Network Analysis (WGCNA). *World Journal of Surgical Oncology* 2024;22(1):177.

312. Luo Y, Liu Z, Hu X. ceRNA Network and WGCNA Analyses of Differentially Expressed Genes in Cervical Cancer Tissues for Association with Survival of Patients. *Reproductive Sciences* 2024.

313. Tsherniak A, Vazquez F, Montgomery PG, et al. Defining a Cancer Dependency Map. *Cell* 2017;170(3):564-76 e16.

314. Huang H-Y, Wu W-R, Wang Y-H, et al. ASS1 as a Novel Tumor Suppressor Gene in Myxofibrosarcomas: Aberrant Loss via Epigenetic DNA Methylation Confers Aggressive Phenotypes, Negative Prognostic Impact, and Therapeutic Relevance. *Clinical Cancer Research* 2013;19(11):2861-72.

315. Dempster JM, Boyle I, Vazquez F, et al. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biology* 2021;22(1):343.

316. Otten ABC, Sun BK. Research Techniques Made Simple: CRISPR Genetic Screens. *J Invest Dermatol* 2020;140(4):723-28.e1.

317. Panditrao G, Bhowmick R, Meena C, Sarkar RR. Emerging landscape of molecular interaction networks: Opportunities, challenges and prospects. *Journal of Biosciences* 2022;47(2):24.

318. di Micco P, Antolin AA, Mitsopoulos C, et al. canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic Acids Res* 2023;51(D1):D1212-D19.

319. Zhou Y, Zhang Y, Zhao D, et al. TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Research* 2023;52(D1):D1465-D77.

320. Antolin AA, Sanfelice D, Crisp A, et al. The Chemical Probes Portal: an expert review-based public resource to empower chemical probe assessment, selection and use. *Nucleic Acids Research* 2022;51(D1):D1492-D502.

321. R Core Team. R: A language and environment for statistical computing. *MSOR connections* 2014;1.

322. Rstudio: Integrated Development for R. [program]: Rstudio, PBC, 2020.

323. Inc A. Anaconda Software Distribution [internet]: Anaconda, 2020.

324. Raybaut P. Spyder-documentation. 2009.

325. Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* 2015;19(1A):A68-77.

326. Zhang J, Baran J, Cros A, et al. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database* 2011;2011.

327. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* 2012;41(D1):D991-D95.

328. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* 2016;44(8):e71.

329. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 2007;23(14):1846-47.

330. Grossman RL, Heath AP, Ferretti V, et al. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine* 2016;375(12):1109-12.

331. Rivest R. RFC1321: The MD5 message-digest algorithm: RFC Editor, 1992.

332. Gobble RM, Qin LX, Brill ER, et al. Expression profiling of liposarcoma yields a multigene predictor of patient outcome and identifies genes that contribute to liposarcomagenesis. *Cancer Res* 2011;71(7):2697-705.

333. Zuco V, Pasquali S, Tortoreto M, et al. Selinexor versus doxorubicin in dedifferentiated liposarcoma PDXs: evidence of greater activity and apoptotic response dependent on p53 nuclear accumulation and survivin down-regulation. *J Exp Clin Cancer Res* 2021;40(1):83.

334. S A. *FastQC: a quality control tool for hig throughput sequence data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

335. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;34(17):i884-i90.

336. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29(1):15-21.

337. Baruzzo G, Hayer KE, Kim EJ, et al. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods* 2017;14(2):135-39.

338. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci U S A* 2010;107(21):9546-51.

339. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26(1):139-40.

340. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 2010;11(3):R25.

341. Zhao Y, Li M-C, Konaté MM, et al. TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. *Journal of Translational Medicine* 2021;19(1):269.

342. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;43(7):e47.

343. Huber W, Carey VJ, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* 2015;12(2):115-21.

344. Kolde R. Pheatmap: Pretty Heatmaps, 2019.

345. .

346. Oldham MC, Langfelder P, Horvath S. Network methods for describing sample relationships in genomic datasets: application to Huntington's disease. *BMC Systems Biology* 2012;6(1):63.

347. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* 2007;8:22.

348. Hao Y, Stuart T, Kowalski MH, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology* 2024;42(2):293-304.

349. Choudhary S, Satija R. Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biology* 2022;23(1):27.

350. Ahlmann-Eltze C, Huber W. Comparison of transformations for single-cell RNA-seq data. *Nature Methods* 2023;20(5):665-72.

351. Becht E, McInnes L, Healy J, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology* 2019;37(1):38-44.

352. Liu R, Cheng Y, Yu J, et al. Identification and validation of gene module associated with lung cancer through coexpression network analysis. *Gene* 2015;563(1):56-62.

353. Cui J, Yi G, Li J, et al. Increased EHHADH Expression Predicting Poor Survival of Osteosarcoma by Integrating Weighted Gene Coexpression Network Analysis and Experimental Validation. *BioMed Research International* 2021;2021:9917060.

354. Altay G, Zapardiel-Gonzalo J, Peters B. RNA-seq preprocessing and sample size considerations for gene network inference. *bioRxiv* 2023.

355. Liang W, Sun F. Weighted gene co-expression network analysis to define pivotal modules and genes in diabetic heart failure. *Biosci Rep* 2020;40(7).

356. Langfelder P. *WGCNA sample size minimum: why?* . https://support.bioconductor.org/p/122944/ (accessed 04/04/2024).

357. Horvarth PLaS. *WGCNA package FAQ*. https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/faq.html (accessed 01/02/2022).

358. Aggarwal R, Ranganathan P. Common pitfalls in statistical analysis: The use of correlation techniques. *Perspect Clin Res* 2016;7(4):187-90.

359. Schönbrodt FD, Perugini M. At what sample size do correlations stabilize? *Journal of Research in Personality* 2013;47(5):609-12.

360. Sun P, Ma R, Liu G, et al. Pathological prognostic factors of retroperitoneal liposarcoma: comprehensive clinicopathological analysis of 124 cases. *Annals of Translational Medicine* 2021;9(7):574.

361. Italiano A, Toulmonde M, Cioffi A, et al. Advanced well-differentiated/dedifferentiated liposarcomas: role of chemotherapy and survival. *Annals of Oncology* 2012;23(6):1601-07.

362. Rai MF, Tycksen ED, Sandell LJ, Brophy RH. Advantages of RNA-seq compared to RNA microarrays for transcriptome profiling of anterior cruciate ligament tears. *J Orthop Res* 2018;36(1):484-97.

363. Romero JP, Ortiz-Estévez M, Muniategui A, et al. Comparison of RNA-seq and microarray platforms for splice event detection using a cross-platform algorithm. *BMC Genomics* 2018;19(1):703.

364. Ballouz S, Verleyen W, Gillis J. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics* 2015;31(13):2123-30.

365. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. *Fam Med Community Health* 2020;8(1):e000262.

366. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part IV: further concepts and methods in survival analysis. *Br J Cancer* 2003;89(5):781-6.

367. Busch EL. Cut points and contexts. *Cancer* 2021;127(23):4348-55.

368. Tustumi F. Choosing the most appropriate cut-point for continuous variables. *Rev Col Bras Cir* 2022;49:e20223346.

369. Therneau TM. *A Package for Survival Analysis in R*. https://CRAN.R-project.org/package=survival.

370. In J, Lee DK. Survival analysis: part II - applied clinical data analysis. *Korean J Anesthesiol* 2019;72(5):441-57.

371. FE HJ. *rms: Regression Modeling Strategies*. . https://hbiostat.org/R/rms/.

372. Forthofer RN, Lee ES, Hernandez M. 13 - Linear Regression. In: Forthofer RN, Lee ES, Hernandez M (eds.) *Biostatistics (Second Edition)*. San Diego: Academic Press; 2007 p349-86.

373. Akinwande MO, Dikko HG, Samson A. Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. *Open Journal of Statistics* 2015;Vol.05No.07:14.

374. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;10(1):57-63.

375. Zhang W, Yu Y, Hertwig F, et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol* 2015;16(1):133.

376. Liesecke F, De Craene JO, Besseau S, et al. Improved gene co-expression network quality through expression dataset down-sampling and network aggregation. *Sci Rep* 2019;9(1):14431.

377. Ovens K, Eames BF, McQuillan I. The impact of sample size and tissue type on the reproducibility of gene co-expression networks *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Virtual Event, USA*: Association for Computing Machinery; 2020.

378. Bush SJ, McCulloch MEB, Summers KM, et al. Integration of quantitated expression estimates from polyA-selected and rRNA-depleted RNA-seq libraries. *BMC Bioinformatics* 2017;18(1):301.

379. Zhao S, Zhang Y, Gamini R, et al. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Scientific Reports* 2018;8(1):4781.

380. Chen L, Yang R, Kwan T, et al. Paired rRNA-depleted and polyA-selected RNA sequencing data and supporting multi-omics data from human T cells. *Scientific Data* 2020;7(1):376.

381. Schroeder B, LaFranzo N, LaFleur B, et al. CD4+ T cell and M2 macrophage infiltration predict dedifferentiated liposarcoma patient outcomes. *Journal for ImmunoTherapy of Cancer* 2021;9:e002812.

382. Ye L, Hu C, Wang C, et al. Nomogram for predicting the overall survival and cancer-specific survival of patients with extremity liposarcoma: a population-based study. *BMC Cancer* 2020;20(1):889.

383. Gronchi A, Strauss DC, Miceli R, et al. Variability in Patterns of Recurrence After Resection of Primary Retroperitoneal Sarcoma (RPS): A Report on 1007 Patients From the Multi-institutional Collaborative RPS Working Group. *Annals of Surgery* 2016;263(5):1002-09.

384. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. *Genome Biology* 2016;17(1):13.

385. Koch CM, Chiu SF, Akbarpour M, et al. A Beginner's Guide to Analysis of RNA Sequencing Data. *Am J Respir Cell Mol Biol* 2018;59(2):145-57.

386. Liu Y, Tingart M, Lecouturier S, et al. Identification of co-expression network correlated with different periods of adipogenic and osteogenic differentiation of BMSCs by weighted gene co-expression network analysis (WGCNA). *BMC Genomics* 2021;22(1):254.

387. Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. *PLoS Comput Biol* 2008;4(8):e1000117.

388. Albert R. Scale-free networks in cell biology. *J Cell Sci* 2005;118(Pt 21):4947-57.

389. Broido AD, Clauset A. Scale-free networks are rare. *Nat Commun* 2019;10(1):1017.

390. Albert R, Jeong H, Barabási A-L. Error and attack tolerance of complex networks. *Nature* 2000;406(6794):378-82.

391. Albert R, Barabási A-L. Statistical mechanics of complex networks. *Reviews of Modern Physics* 2002;74(1):47-97.

392. Ghoshal G, Barabási A-L. Ranking stability and super-stable nodes in complex networks. *Nature Communications* 2011;2(1):394.

393. Saul ZM, Filkov V. Exploring biological network structure using exponential random graph models. *Bioinformatics* 2007;23(19):2604-11.

394. Gillespie CS. Fitting Heavy Tailed Distributions: The poweRlaw Package. *Journal of Statistical Software* 2015;64(2):1 - 16.

395. Alstott J, Bullmore E, Plenz D. Powerlaw: a Python package for analysis of heavy-tailed distributions. *PLoS One* 2014;9(1):e85777.

396. Adolfsson A, Ackerman M, Brownstein NC. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition* 2019;88:13-26.

397. Martin FJ, Amode MR, Aneja A, et al. Ensembl 2023. *Nucleic Acids Research* 2022;51(D1):D933-D41.

398. Dong J, Horvath S. Understanding network concepts in modules. *BMC Systems Biology* 2007;1(1):24.

399. Behan FM, Iorio F, Picco G, et al. Prioritization of cancer therapeutic targets using CRISPR–Cas9 screens. *Nature* 2019;568(7753):511-16.

400. Taylor IW, Linding R, Warde-Farley D, et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology* 2009;27(2):199-204.

401. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Research* 2016;45(4):e22-e22.

402. Langfelder P. *Signed or unsigned: which network type is preferable?* https://peterlangfelder.com/2018/11/25/signed-or-unsigned-which-network-type-is-preferable/.

403. Großwendt A, Röglin H. Improved Analysis of Complete-Linkage Clustering. *Algorithmica* 2017;78(4):1131-50.

404. Zhu D, Zhou M, Zhang H, et al. Network analysis identifies a gene biomarker panel for sepsis-induced acute respiratory distress syndrome. *BMC Med Genomics* 2023;16(1):165.

405. Videlock EJ, Hatami A, Zhu C, et al. Distinct Patterns of Gene Expression Changes in the Colon and Striatum of Young Mice Overexpressing Alpha-Synuclein Support Parkinson's Disease as a Multi-System Process. *J Parkinsons Dis* 2023;13(7):1127-47.

406. Tosadori G, Bestvina I, Spoto F, et al. Creating, generating and comparing random network models with NetworkRandomizer [version 3; peer review: 2 approved, 1 approved with reservations]. *F1000Research* 2017;5(2524).

407. Fredrickson MM, Chen Y. Permutation and randomization tests for network analysis. *Social Networks* 2019;59:171-83.

408. Neidlin M, Dimitrakopoulou S, Alexopoulos LG. Multi-tissue network analysis for drug prioritization in knee osteoarthritis. *Scientific Reports* 2019;9(1):15176.

409. Lombardo MV, Courchesne E, Lewis NE, Pramparo T. Hierarchical cortical transcriptome disorganization in autism. *Molecular Autism* 2017;8(1):29.

410. Jiang J, Ding Y, Wu M, et al. Identification of TYROBP and C1QB as Two Novel Key Genes With Prognostic Value in Gastric Cancer by Network Analysis. *Frontiers in Oncology* 2020;10.

411. Adiconis X, Borges-Rivera D, Satija R, et al. Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat Methods* 2013;10(7):623-9.

412. Wang X, Almet AA, Nie Q. The promising application of cell-cell interaction analysis in cancer from single-cell and spatial transcriptomics. *Semin Cancer Biol* 2023;95:42-51.

413. Hou J, Ye X, Feng W, et al. Distance correlation application to gene co-expression network analysis. *BMC Bioinformatics* 2022;23(1):81.

414. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008;2008(10):P10008.

415. Korotkevich G, Sukhov V, Budin N, et al. Fast gene set enrichment analysis. *bioRxiv* 2021:060012.

416. Liberzon A, Birger C, Thorvaldsdóttir H, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 2015;1(6):417-25.

417. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 2009;37(Web Server issue):W305-11.

418. Sayers EW, Beck J, Bolton EE, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2024;52(D1):D33-d43.

419. Cain JA, Montibus B, Oakey RJ. Intragenic CpG Islands and Their Impact on Gene Regulation. *Frontiers in Cell and Developmental Biology* 2022;10.

420. Contreras O, Rossi FMV, Theret M. Origins, potency, and heterogeneity of skeletal muscle fibro-adipogenic progenitors-time for new definitions. *Skelet Muscle* 2021;11(1):16.

421. Mehran R, Nilsson M, Khajavi M, et al. Tumor endothelial markers define novel subsets of cancer-specific circulating endothelial cells associated with antitumor efficacy. *Cancer Res* 2014;74(10):2731-41.

422. Maddison K, Bowden NA, Graves MC, Tooney PA. Characteristics of vasculogenic mimicry and tumour to endothelial transdifferentiation in human glioblastoma: a systematic review. *BMC Cancer* 2023;23(1):185.

423. Wechman SL, Emdad L, Sarkar D, et al. Vascular mimicry: Triggers, molecular interactions and in vivo models. *Adv Cancer Res* 2020;148:27-67.

424. Cho JG, Lee A, Chang W, et al. Endothelial to Mesenchymal Transition Represents a Key Link in the Interaction between Inflammation and Endothelial Dysfunction. *Front Immunol* 2018;9:294.

425. Han C, Liu T, Yin R. Biomarkers for cancer-associated fibroblasts. *Biomarker Research* 2020;8(1):64.

426. Baer PC. Adipose-derived mesenchymal stromal/stem cells: An update on their phenotype in vivo and in vitro. *World J Stem Cells* 2014;6(3):256-65.

427. Camilleri ET, Gustafson MP, Dudakovic A, et al. Identification and validation of multiple cell surface markers of clinical-grade adipose-derived mesenchymal stromal cells as novel release criteria for good manufacturing practice-compliant production. *Stem Cell Research & Therapy* 2016;7(1):107.

428. Galfrè SG, Morandin F, Pietrosanto M, et al. COTAN: scRNA-seq data analysis based on gene co-expression. *NAR Genomics and Bioinformatics* 2021;3(3).

429. Hanahan D. Hallmarks of Cancer: New Dimensions. *Cancer Discovery* 2022;12(1):31-46.

430. Shen J, Shrestha S, Rao PN, et al. Pericytic mimicry in well-differentiated liposarcoma/atypical lipomatous tumor. *Hum Pathol* 2016;54:92-9.

431. Koncz G, Jenei V, Toth M, et al. Damage-mediated macrophage polarization in sterile inflammation. *Front Immunol* 2023;14:1169560.

432. Xiao M, Xu J, Wang W, et al. Functional significance of cholesterol metabolism in cancer: from threat to treatment. *Experimental & Molecular Medicine* 2023;55(9):1982-95.

433. Viet-Nhi NK, Minh Quan T, Cong Truc V, et al. Multi-Omics Analysis Reveals the IFI6 Gene as a Prognostic Indicator and Therapeutic Target in Esophageal Cancer. *Int J Mol Sci* 2024;25(5).

434. Giraulo C, Turiello R, Orlando L, et al. The CD73 is induced by TGF-β1 triggered by nutrient deprivation and highly expressed in dedifferentiated human melanoma. *Biomedicine & Pharmacotherapy* 2023;165:115225.

435. Kuzu OF, Noory MA, Robertson GP. The Role of Cholesterol in Cancer. *Cancer Res* 2016;76(8):2063-70.

436. Ciesielski O, Biesiekierska M, Panthu B, et al. The Epigenetic Profile of Tumor Endothelial Cells. Effects of Combined Therapy with Antiangiogenic and Epigenetic Drugs on Cancer Progression. *Int J Mol Sci* 2020;21(7).

437. Sangwung P, Zhou G, Nayak L, et al. KLF2 and KLF4 control endothelial identity and vascular integrity. *JCI Insight* 2017;2(4).

438. Hanahan D, Weinberg RA. Biological hallmarks of cancer. *Holland-Frei Cancer Medicine* 2016:1-10.

439. Hanahan D, Weinberg Robert A. Hallmarks of Cancer: The Next Generation. *Cell* 2011;144(5):646-74.

440. Fenton SE, Saleiro D, Platanias LC. Type I and II Interferons in the Anti-Tumor Immune Response. *Cancers* 2021;13(5):1037.

441. Wang W, Lopez McDonald MC, Kim C, et al. The complementary roles of STAT3 and STAT1 in cancer biology: insights into tumor pathogenesis and therapeutic strategies. *Frontiers in Immunology* 2023;14.

442. Resag A, Toffanin G, Benešová I, et al. The Immune Contexture of Liposarcoma and Its Clinical Implications. *Cancers (Basel)* 2022;14(19).

443. Jorgovanovic D, Song M, Wang L, Zhang Y. Roles of IFN-γ in tumor progression and regression: a review. *Biomarker Research* 2020;8(1):49.

444. Castro F, Cardoso AP, Gonçalves RM, et al. Interferon-Gamma at the Crossroads of Tumor Immune Surveillance or Evasion. *Frontiers in Immunology* 2018;9.

445. Wynn TA, Ramalingam TR. Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nat Med* 2012;18(7):1028-40.

446. Hasanov Z, Ruckdeschel T, König C, et al. Endosialin Promotes Atherosclerosis Through Phenotypic Remodeling of Vascular Smooth Muscle Cells. *Arterioscler Thromb Vasc Biol* 2017;37(3):495-505.

447. Mogler C, Wieland M, König C, et al. Hepatic stellate cell-expressed endosialin balances fibrogenesis and hepatocyte proliferation during liver damage. *EMBO Mol Med* 2015;7(3):332-8.

448. Hong YK, Lin YC, Cheng TL, et al. TEM1/endosialin/CD248 promotes pathologic scarring and TGF-β activity through its receptor stability in dermal fibroblasts. *J Biomed Sci* 2024;31(1):12.

449. Simonavicius N, Robertson D, Bax DA, et al. Endosialin (CD248) is a marker of tumor-associated pericytes in high-grade glioma. *Modern Pathology* 2008;21(3):308-15.

450. Petrus P, Fernandez TL, Kwon MM, et al. Specific loss of adipocyte CD248 improves metabolic health via reduced white adipose tissue hypoxia, fibrosis and inflammation. *EBioMedicine* 2019;44:489-501.

451. Brett E, Zielins ER, Chin M, et al. Isolation of CD248-expressing stromal vascular fraction for targeted improvement of wound healing. *Wound Repair Regen* 2017;25(3):414-22.

452. Kretschmer M, Rüdiger D, Zahler S. Mechanical Aspects of Angiogenesis. *Cancers (Basel)* 2021;13(19).

453. O'Shannessy DJ, Somers EB, Chandrasekaran LK, et al. Influence of tumor microenvironment on prognosis in colorectal cancer: Tissue architecture-dependent signature of endosialin (TEM-1) and associated proteins. *Oncotarget* 2014;5(12):3983-95.

454. O'Shannessy DJ, Dai H, Mitchell M, et al. Endosialin and Associated Protein Expression in Soft Tissue Sarcomas: A Potential Target for Anti-Endosialin Therapeutic Strategies. *Sarcoma* 2016;2016(1):5213628.

455. Diaz LA, Jr., Coughlin CM, Weil SC, et al. A first-in-human phase I study of MORAb-004, a monoclonal antibody to endosialin in patients with advanced solid tumors. *Clin Cancer Res* 2015;21(6):1281-8.

456. Jones RL, Chawla SP, Attia S, et al. A phase 1 and randomized controlled phase 2 trial of the safety and efficacy of the combination of gemcitabine and docetaxel with ontuxizumab (MORAb-004) in metastatic soft-tissue sarcomas. *Cancer* 2019;125(14):2445-54.

457. Krock BL, Skuli N, Simon MC. Hypoxia-induced angiogenesis: good and evil. *Genes Cancer* 2011;2(12):1117-33.

458. Dudley AC. Tumor endothelial cells. *Cold Spring Harb Perspect Med* 2012;2(3):a006536.

459. Leone P, Malerba E, Susca N, et al. Endothelial cells in tumor microenvironment: insights and perspectives. *Frontiers in Immunology* 2024;15.

460. Lee S. The association of genetically controlled CpG methylation (cg158269415) of protein tyrosine phosphatase, receptor type N2 (PTPRN2) with childhood obesity. *Scientific Reports* 2019;9(1):4855.

461. Yin J, Guo Y. HOXD13 promotes the malignant progression of colon cancer by upregulating PTPRN2. *Cancer Med* 2021;10(16):5524-33.

462. Sorokin AV, Nair BC, Wei Y, et al. Aberrant Expression of proPTPRN2 in Cancer Cells Confers Resistance to Apoptosis. *Cancer Research* 2015;75(9):1846-58.

463. Sengelaub CA, Navrazhina K, Ross JB, et al. PTPRN2 and PLCβ1 promote metastatic breast cancer cell migration through PI(4,5)P2-dependent actin remodeling. *Embo j* 2016;35(1):62-76.

464. Paniagua-Herranz L, Moreno I, Nieto-Jiménez C, et al. Genomic and Immunologic Correlates in Prostate Cancer with High Expression of KLK2. *International Journal of Molecular Sciences* 2024;25(4):2222.

465. Chen J, Bai Z, Wang Y, et al. Aberrant expression of PTPRN2 promotes malignant transformation of colorectal cancer cells through EMT/TRAF2/STAT3 signaling pathway. *Research Square* 2023.

466. Shen J, LeFave C, Sirosh I, et al. Integrative epigenomic and genomic filtering for methylation markers in hepatocellular carcinomas. *BMC Medical Genomics* 2015;8(1):28.

467. Vargas AC, Gray LA, White CL, et al. Genome wide methylation profiling of selected matched soft tissue sarcomas identifies methylation changes in metastatic and recurrent disease. *Sci Rep* 2021;11(1):667.

468. Chen CL, Mahalingam D, Osmulski P, et al. Single-cell analysis of circulating tumor cells identifies cumulative expression patterns of EMT-related genes in metastatic prostate cancer. *Prostate* 2013;73(8):813-26.

469. Balakrishnan A, Guruprasad KP, Satyamoorthy K, Joshi MB. Interleukin-6 determines protein stabilization of DNA methyltransferases and alters DNA promoter methylation of genes associated with insulin signaling and angiogenesis. *Laboratory Investigation* 2018;98(9):1143-58.

470. Huang B, Song B-l, Xu C. Cholesterol metabolism in cancer: mechanisms and therapeutic opportunities. *Nature Metabolism* 2020;2(2):132-41.

471. Fu Y, Zou T, Shen X, et al. Lipid metabolism in cancer progression and therapeutic strategies. *MedComm (2020)* 2021;2(1):27-59.

472. Koundouros N, Poulogiannis G. Reprogramming of fatty acid metabolism in cancer. *British Journal of Cancer* 2020;122(1):4-22.

473. Currie E, Schulze A, Zechner R, et al. Cellular fatty acid metabolism and cancer. *Cell Metab* 2013;18(2):153-61.

474. Jiang M, Karsenberg R, Bianchi F, van den Bogaart G. CD36 as a double-edged sword in cancer. *Immunology Letters* 2024;265:7-15.

475. Du A, Wang Z, Huang T, et al. Fatty acids in cancer: Metabolic functions and potential treatment. *MedComm – Oncology* 2023;2(1):e25.

476. Huff T, Boyd B, Jialal I. Physiology, Cholesterol *StatPearls*. Treasure Island (FL): StatPearls Publishing

Copyright © 2025, StatPearls Publishing LLC.; 2025.

477. Vasseur S, Guillaumond F. Lipids in cancer: a global view of the contribution of lipid pathways to metastatic formation and treatment resistance. *Oncogenesis* 2022;11(1):46.

478. Ehmsen S, Pedersen MH, Wang G, et al. Increased Cholesterol Biosynthesis Is a Key Characteristic of Breast Cancer Stem Cells Influencing Patient Outcome. *Cell Rep* 2019;27(13):3927-38.e6.

479. Liu M, Xia Y, Ding J, et al. Transcriptional Profiling Reveals a Common Metabolic Program in High-Risk Human Neuroblastoma and Mouse Neuroblastoma Sphere-Forming Cells. *Cell Rep* 2016;17(2):609-23.

480. Liu H, Zhang Z, Song L, et al. Lipid metabolism of cancer stem cells. *Oncol Lett* 2022;23(4):119.

481. Hwang S, Chung KW. Targeting fatty acid metabolism for fibrotic disorders. *Archives of Pharmacal Research* 2021;44(9):839-56.

482. Rajesh R, Atallah R, Bärnthaler T. Dysregulation of metabolic pathways in pulmonary fibrosis. *Pharmacology & Therapeutics* 2023;246:108436.

483. Henderson J, O'Reilly S. The emerging role of metabolism in fibrosis. *Trends in Endocrinology & Metabolism* 2021;32(8):639-53.

484. Suryadevara V, Ramchandran R, Kamp DW, Natarajan V. Lipid Mediators Regulate Pulmonary Fibrosis: Potential Mechanisms and Signaling Pathways. *International Journal of Molecular Sciences* 2020;21(12):4257.

485. Burgy O, Loriod S, Beltramo G, Bonniaud P. Extracellular Lipids in the Lung and Their Role in Pulmonary Fibrosis. *Cells* 2022;11(7):1209.

486. Ehmsen S, Pedersen MH, Wang G, et al. Increased Cholesterol Biosynthesis Is a Key Characteristic of Breast Cancer Stem Cells Influencing Patient Outcome. *Cell Reports* 2019;27(13):3927-38.e6.

487. Chen CL, Uthaya Kumar DB, Punj V, et al. NANOG Metabolically Reprograms Tumor-Initiating Stem-like Cells through Tumorigenic Changes in Oxidative Phosphorylation and Fatty Acid Metabolism. *Cell Metab* 2016;23(1):206-19.

488. Gao S, Soares F, Wang S, et al. CRISPR screens identify cholesterol biosynthesis as a therapeutic target on stemness and drug resistance of colon cancer. *Oncogene* 2021;40(48):6601-13.

489. Esau L, Sagar S, Bangarusamy D, Kaur M. Identification of CETP as a molecular target for estrogen positive breast cancer cell death by cholesterol depleting agents. *Genes Cancer* 2016;7(9-10):309-22.

490. Wu K, Zou L, Lei X, Yang X. Roles of ABCA1 in cancer. *Oncol Lett* 2022;24(4):349.

491. Dufour J, Viennois E, De Boussac H, et al. Oxysterol receptors, AKT and prostate cancer. *Curr Opin Pharmacol* 2012;12(6):724-8.

492. Chan ES, Zhang H, Fernandez P, et al. Effect of cyclooxygenase inhibition on cholesterol efflux proteins and atheromatous foam cell transformation in THP-1 human macrophages: a possible mechanism for increased cardiovascular risk. *Arthritis Res Ther* 2007;9(1):R4.

493. Azuaje FJ. Selecting biologically informative genes in co-expression networks with a centrality score. *Biology Direct* 2014;9(1):12.

494. Alvarez-Socorro AJ, Herrera-Almarza GC, González-Díaz LA. Eigencentrality based on dissimilarity measures reveals central nodes in complex networks. *Scientific Reports* 2015;5(1):17095.

495. Hao M, Liu W, Ding C, et al. Identification of hub genes and small molecule therapeutic drugs related to breast cancer with comprehensive bioinformatics analysis. *PeerJ* 2020;8:e9946.

496. Yang W, Zhao X, Han Y, et al. Identification of hub genes and therapeutic drugs in esophageal squamous cell carcinoma based on integrated bioinformatics strategy. *Cancer Cell International* 2019;19(1):142.

497. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 2012;30(5):413-21.

498. Taylor AM, Shih J, Ha G, et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer Cell* 2018;33(4):676-89.e3.

499. Chibon F, Lesluyes T, Valentin T, Le Guellec S. CINSARC signature as a prognostic marker for clinical outcome in sarcomas and beyond. *Genes, Chromosomes and Cancer* 2019;58(2):124-29.

500. Lesluyes T, Delespaul L, Coindre J-M, Chibon F. The CINSARC signature as a prognostic marker for clinical outcome in multiple neoplasms. *Scientific Reports* 2017;7(1):5480.

501. Brunac A-C, Fourquet J, Perot G, et al. CINSARC signature outperforms gold-standard TNM staging and consensus molecular subtypes for clinical outcome in stage II&#x2013;III colorectal carcinoma. *Modern Pathology* 2022;35(12):2002-10.

502. Callegaro D, Tinè G, Oppong FB, et al. CINSARC and Sarculator in Patients with Primary Retroperitoneal Sarcoma: A Combined Analysis of Single-Institution Data and the EORTC-STBSG-62092 Trial (STRASS). *Clin Cancer Res* 2025;31(15):3239-48.

503. Nielsen CF, Zhang T, Barisic M, et al. Topoisomerase IIα is essential for maintenance of mitotic chromosome structure. *Proc Natl Acad Sci U S A* 2020;117(22):12131-42.

504. Min M, Mayor U, Dittmar G, Lindon C. Using in vivo biotinylated ubiquitin to describe a mitotic exit ubiquitome from human cells. *Mol Cell Proteomics* 2014;13(9):2411-25.

505. Van Tine BA, Agulnik M, Olson RD, et al. A phase II clinical study of 13-deoxy, 5-iminodoxorubicin (GPX-150) with metastatic and unresectable soft tissue sarcoma. *Cancer Med* 2019;8(6):2994-3003.

506. Azizeh A, Nafiseh G, Katayoun D. Identification of druggable hub genes and key pathways associated with cervical cancer by protein-protein interaction analysis: An in silico study. *International Journal of Reproductive BioMedicine (IJRM)* 2023;21(10).

507. Hyer ML, Milhollen MA, Ciavarri J, et al. A small-molecule inhibitor of the ubiquitin activating enzyme for cancer treatment. *Nature Medicine* 2018;24(2):186-93.

508. Liu H, Wang X, Liu L, et al. Targeting liposarcoma: unveiling molecular pathways and therapeutic opportunities. *Front Oncol* 2024;14:1484027.

509. Tsuchiya R, Yoshimatsu Y, Noguchi R, et al. Establishment and characterization of NCC-DDLPS1-C1: a novel patient-derived cell line of dedifferentiated liposarcoma. *Hum Cell* 2021;34(1):260-70.

510. Al Shihabi A, Tebon PJ, Nguyen HTL, et al. The landscape of drug sensitivity and resistance in sarcoma. *Cell Stem Cell* 2024;31(10):1524-42.e4.

511. Gao X, Keller KR, Bonzerato CG, et al. The ubiquitin-proteasome pathway inhibitor TAK-243 has major effects on calcium handling in mammalian cells. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 2024;1871(1):119618.

512. Arakawa Y, Jo U, Kumar S, et al. Activity of the Ubiquitin-activating Enzyme Inhibitor TAK-243 in Adrenocortical Carcinoma Cell Lines, Patient-derived Organoids, and Murine Xenografts. *Cancer Research Communications* 2024;4(3):834-48.

513. Majeed S, Aparnathi MK, Nixon KCJ, et al. Targeting the Ubiquitin-Proteasome System Using the UBA1 Inhibitor TAK-243 is a Potential Therapeutic Strategy for Small-Cell Lung Cancer. *Clin Cancer Res* 2022;28(9):1966-78.

514. Liu Y, Awadia S, Delaney A, et al. UAE1 inhibition mediates the unfolded protein response, DNA damage and caspase-dependent cell death in pancreatic cancer. *Transl Oncol* 2020;13(11):100834.

515. Barghout SH, Patel PS, Wang X, et al. Preclinical evaluation of the selective small-molecule UBA1 inhibitor, TAK-243, in acute myeloid leukemia. *Leukemia* 2019;33(1):37-51.

516. Barghout SH, Patel P, Wang X, et al. TAK-243 Is a Selective UBA1 Inhibitor That Displays Preclinical Activity in Acute Myeloid Leukemia (AML). *Blood* 2017;130(Supplement 1):814-14.

517. Parashar S, Kaushik A, Ambasta RK, Kumar P. E2 conjugating enzymes: A silent but crucial player in ubiquitin biology. *Ageing Research Reviews* 2025;108:102740.

518. Yang Q, Zhao J, Chen D, Wang Y. E3 ubiquitin ligases: styles, structures and functions. *Mol Biomed* 2021;2(1):23.

519. Spano D, Catara G. Targeting the Ubiquitin-Proteasome System and Recent Advances in Cancer Therapy. *Cells* 2023;13(1).

520. Soave CL, Guerin T, Liu J, Dou QP. Targeting the ubiquitin-proteasome system for cancer treatment: discovering novel inhibitors from nature and drug repurposing. *Cancer Metastasis Rev* 2017;36(4):717-36.

521. Clague MJ, Heride C, Urbé S. The demographics of the ubiquitin system. *Trends in Cell Biology* 2015;25(7):417-26.

522. Bansal S, Tiwari S. Mechanisms for the temporal regulation of substrate ubiquitination by the anaphase-promoting complex/cyclosome. *Cell Division* 2019;14(1):14.

523. Chang LF, Zhang Z, Yang J, et al. Molecular architecture and mechanism of the anaphase-promoting complex. *Nature* 2014;513(7518):388-93.

524. Martinez-Chacin RC, Bodrug T, Bolhuis DL, et al. Ubiquitin chain-elongating enzyme UBE2S activates the RING E3 ligase APC/C for substrate priming. *Nat Struct Mol Biol* 2020;27(6):550-60.

525. Bavi P, Uddin S, Ahmed M, et al. Bortezomib stabilizes mitotic cyclins and prevents cell cycle progression via inhibition of UBE2C in colorectal carcinoma. *Am J Pathol* 2011;178(5):2109-20.

526. Kariri Y, Toss MS, Alsaleem M, et al. Ubiquitin-conjugating enzyme 2C (UBE2C) is a poor prognostic biomarker in invasive breast cancer. *Breast Cancer Research and Treatment* 2022;192(3):529-39.

527. Lara-Gonzalez P, Moyle MW, Budrewicz J, et al. The G2-to-M Transition Is Ensured by a Dual Mechanism that Protects Cyclin B from Degradation by Cdc20-Activated APC/C. *Developmental Cell* 2019;51(3):313-25.e10.

528. Zhou Z, He M, Shah AA, Wan Y. Insights into APC/C: from cellular function to diseases and therapeutics. *Cell Division* 2016;11(1):9.

529. Dastsooz H, Cereda M, Donna D, Oliviero S. A Comprehensive Bioinformatics Analysis of UBE2C in Cancers. *Int J Mol Sci* 2019;20(9).

530. Xian F, Zhao C, Huang C, et al. The potential role of CDC20 in tumorigenesis, cancer progression and therapy: A narrative review. *Medicine* 2023;102(36):e35038.

531. Zhang M, Wang J, Zhang Z, et al. Diverse roles of UBE2S in cancer and therapy resistance: Biological functions and mechanisms. *Heliyon* 2024;10(2):e24465.

532. Xian F, Yang X, Xu G. Prognostic significance of CDC20 expression in malignancy patients: A meta-analysis. *Frontiers in Oncology* 2022;12.

533. Jalali P, Samii A, Rezaee M, et al. UBE2C: A pan-cancer diagnostic and prognostic biomarker revealed through bioinformatics analysis. *Cancer Rep (Hoboken)* 2024;7(4):e2032.

534. van Ree JH, Jeganathan KB, Malureanu L, van Deursen JM. Overexpression of the E2 ubiquitin-conjugating enzyme UbcH10 causes chromosome missegregation and tumor formation. *J Cell Biol* 2010;188(1):83-100.

535. Bruno S, Ghelli Luserna di Rorà A, Napolitano R, et al. CDC20 in and out of mitosis: a prognostic factor and therapeutic target in hematological malignancies. *Journal of Experimental & Clinical Cancer Research* 2022;41(1):159.

536. Gu Q, Li F, Ge S, et al. CDC20 Knockdown and Acidic Microenvironment Collaboratively Promote Tumorigenesis through Inhibiting Autophagy and Apoptosis. *Mol Ther Oncolytics* 2020;17:94-106.

537. Sherman DJ, Li J. Proteasome Inhibitors: Harnessing Proteostasis to Combat Disease. *Molecules* 2020;25(3).

538. Jo EB, Hong D, Lee YS, et al. Establishment of a Novel PDX Mouse Model and Evaluation of the Tumor Suppression Efficacy of Bortezomib Against Liposarcoma. *Transl Oncol* 2019;12(2):269-81.

539. Perez M, Peinado-Serrano J, Garcia-Heredia JM, et al. Efficacy of bortezomib in sarcomas with high levels of MAP17 (PDZK1IP1). *Oncotarget* 2016;7(41):67033-46.

540. Hu Y, Wang L, Wang L, et al. Preferential cytotoxicity of bortezomib toward highly malignant human liposarcoma cells via suppression of MDR1 expression and function. *Toxicology and Applied Pharmacology* 2015;283(1):1-8.

541. Ludwig MP, Galbraith MD, Eduthan NP, et al. Proteasome Inhibition Sensitizes Liposarcoma to MDM2 Inhibition with Nutlin-3 by Activating the ATF4/CHOP Stress Response Pathway. *Cancer Research* 2023.

542. Cohen-Sharir Y, McFarland JM, Abdusamad M, et al. Aneuploidy renders cancer cells vulnerable to mitotic checkpoint inhibition. *Nature* 2021;590(7846):486-91.

543. Quinton RJ, DiDomizio A, Vittoria MA, et al. Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature* 2021;590(7846):492-97.

544. Haddox CL, Hornick JL, Roland CL, et al. Diagnosis and management of dedifferentiated liposarcoma: A multidisciplinary position statement. *Cancer Treatment Reviews* 2024;131:102846.

545. Tian Z, Yao W. Chemotherapeutic drugs for soft tissue sarcomas: a review. *Front Pharmacol* 2023;14:1199292.

546. Tan C, Etcubanas E, Wollner N, et al. Adriamycin—an antitumor antibiotic in the treatment of neoplastic diseases. *Cancer* 1973;32(1):9-17.

547. Traweek RS, Cope BM, Roland CL, et al. Targeting the MDM2-p53 pathway in dedifferentiated liposarcoma. *Front Oncol* 2022;12:1006959.

548. Gounder MM, Bauer TM, Schwartz GK, et al. A First-in-Human Phase I Study of Milademetan, an MDM2 Inhibitor, in Patients With Advanced Liposarcoma, Solid Tumors, or Lymphomas. *Journal of Clinical Oncology* 2023;41(9):1714-24.

549. Yu B, O'Toole SA, Trent RJ. Somatic DNA mutation analysis in targeted therapy of solid tumours. *Transl Pediatr* 2015;4(2):125-38.

550. Corsello SM, Nagari RT, Spangler RD, et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nat Cancer* 2020;1(2):235-48.

551. Tieri P, Farina L, Petti M, et al. Network Inference and Reconstruction in Bioinformatics. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (eds.) *Encyclopedia of Bioinformatics and Computational Biology*. Oxford: Academic Press; 2019 p805-13.

552. Fionda V. Networks in Biology. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (eds.) *Encyclopedia of Bioinformatics and Computational Biology*. Oxford: Academic Press; 2019 p915-21.

553. Ghafouri-Fard S, Safarzadeh A, Taheri M, Jamali E. Identification of diagnostic biomarkers via weighted correlation network analysis in colorectal cancer using a system biology approach. *Scientific Reports* 2023;13(1):13637.

554. Farhadian M, Rafat SA, Panahi B, Mayack C. Weighted gene co-expression network analysis identifies modules and functionally enriched pathways in the lactation process. *Scientific Reports* 2021;11(1):2367.

555. Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. *Nature Communications* 2018;9(1):1090.

556. Lv J, Zhou Y, Jin S, et al. WGCNA-ML-MR integration: uncovering immune-related genes in prostate cancer. *Front Oncol* 2025;15:1534612.

557. Shi G, Shen Z, Liu Y, Yin W. Identifying Biomarkers to Predict the Progression and Prognosis of Breast Cancer by Weighted Gene Co-expression Network Analysis. *Frontiers in Genetics* 2020;Volume 11 - 2020.

558. Liu R, Liu J, Cao Q, et al. Identification of crucial genes through WGCNA in the progression of gastric cancer. *Journal of Cancer* 2024;15(11):3284-96.

559. Suski JM, Braun M, Strmiska V, Sicinski P. Targeting cell-cycle machinery in cancer. *Cancer Cell* 2021;39(6):759-78.

560. Singh N, Baby D, Rajguru JP, et al. Inflammation and cancer. *Ann Afr Med* 2019;18(3):121-26.

561. Jirovec A, Flaman A, Godbout E, et al. Immune profiling of dedifferentiated liposarcoma and identification of novel antigens for targeted immunotherapy. *Scientific Reports* 2024;14(1):11254.

562. Petitprez F, de Reyniès A, Keung EZ, et al. B cells are associated with survival and immunotherapy response in sarcoma. *Nature* 2020;577(7791):556-60.

563. Antar SA, Ashour NA, Marawan ME, Al-Karmalawy AA. Fibrosis: Types, Effects, Markers, Mechanisms for Disease Progression, and Its Relation with Oxidative Stress, Immunity, and Inflammation. *Int J Mol Sci* 2023;24(4).

564. Bai C, Li S, Tan Z, Fan Z. Targeting MCM2 activates cancer-associated fibroblasts-like phenotype and affects chemo-resistance of liposarcoma cells against doxorubicin. *Anticancer Drugs* 2024;35(10):883-92.

565. Wang GY, Lucas DR. Dedifferentiated Liposarcoma With Myofibroblastic Differentiation. *Archives of Pathology & Laboratory Medicine* 2018;142(10):1159-63.

566. Paunescu V, Bojin FM, Tatu CA, et al. Tumour-associated fibroblasts and mesenchymal stem cells: more similarities than differences. *J Cell Mol Med* 2011;15(3):635-46.

567. Wrenn ED, Apfelbaum AA, Rudzinski ER, et al. Cancer-Associated Fibroblast-Like Tumor Cells Remodel the Ewing Sarcoma Tumor Microenvironment. *Clinical Cancer Research* 2023;29(24):5140-54.

568. Borriello L, Nakata R, Sheard MA, et al. Cancer-Associated Fibroblasts Share Characteristics and Protumorigenic Activity with Mesenchymal Stromal Cells. *Cancer Res* 2017;77(18):5142-57.

569. Denu RA, Nemcek S, Bloom DD, et al. Fibroblasts and Mesenchymal Stromal/Stem Cells Are Phenotypically Indistinguishable. *Acta Haematol* 2016;136(2):85-97.

570. Yamada Y, Mizoguchi K, Shiba E, et al. A Case of Dedifferentiated Liposarcoma That Contributes to Accompanying Vessels of Various Size. *Diagnostics (Basel)* 2024;14(15).

571. DuBois SG, Marina N, Glade-Bender J. Angiogenesis and vascular targeting in Ewing sarcoma: a review of preclinical and clinical data. *Cancer* 2010;116(3):749-57.

572. Choi KJ, Nam J-K, Kim J-H, et al. Endothelial-to-mesenchymal transition in anticancer therapy and normal tissue damage. *Experimental & Molecular Medicine* 2020;52(5):781-92.

573. Ma X, Geng Z, Wang S, et al. The driving mechanism and targeting value of mimicry between vascular endothelial cells and tumor cells in tumor progression. *Biomedicine & Pharmacotherapy* 2023;165:115029.

574. Ren K, Yao N, Wang G, et al. Vasculogenic mimicry: a new prognostic sign of human osteosarcoma. *Hum Pathol* 2014;45(10):2120-9.

575. Vergani E, Beretta GL, Aloisi M, et al. Targeting of the Lipid Metabolism Impairs Resistance to BRAF Kinase Inhibitor in Melanoma. *Front Cell Dev Biol* 2022;10:927118.

576. Zhang W, Xu Y, Fang Y, et al. Ubiquitination in lipid metabolism reprogramming: implications for pediatric solid tumors. *Front Immunol* 2025;16:1554311.

577. Fromigué O, Haÿ E, Modrowski D, et al. RhoA GTPase inactivation by statins induces osteosarcoma cell apoptosis by inhibiting p42/p44-MAPKs-Bcl-2 signaling independently of BMP-2 and cell differentiation. *Cell Death & Differentiation* 2006;13(11):1845-56.

578. Fernández-Pérez L, Guerra B, M.García-Castellano J, et al. Statins: Are Lipid-lowering Drugs Useful in Sarcomas? In: Amarasekera HW (ed.) *Bone Tumours - A Comprehensive Review of Selected Topics*. Rijeka: IntechOpen; 2022.

579. Hong Y, Zhang L, Lin W, et al. Transcriptome Sequencing Unveils a Molecular-Stratification-Predicting Prognosis of Sarcoma Associated with Lipid Metabolism. *Int J Mol Sci* 2024;25(3).

580. Mason O, Verwoerd M. Graph theory and networks in Biology. *IET Syst Biol* 2007;1(2):89-119.

581. Paci P, Fiscon G, Conte F, et al. Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery. *npj Systems Biology and Applications* 2021;7(1):3.

582. Mares-Quiñones MD, Galán-Vásquez E, Pérez-Rueda E, et al. Identification of modules and key genes associated with breast cancer subtypes through network analysis. *Scientific Reports* 2024;14(1):12350.

583. Jia Y, Yang J, Chen Y, et al. Identification of NCAPG as an Essential Gene for Neuroblastoma Employing CRISPR-Cas9 Screening Database and Experimental Verification. *Int J Mol Sci* 2023;24(19).

584. Wu Z, Lei K, Li H, et al. Transcriptome-based network analysis related to M2-like tumor-associated macrophage infiltration identified VARS1 as a potential target for improving melanoma immunotherapy efficacy. *J Transl Med* 2022;20(1):489.

585. Song G, Liu J, Tang X, et al. Cell cycle checkpoint revolution: targeted therapies in the fight against malignant tumors. *Frontiers in Pharmacology* 2024;Volume 15 - 2024.

586. Cavalu S, Abdelhamid AM, Saber S, et al. Cell cycle machinery in oncology: A comprehensive review of therapeutic targets. *The FASEB Journal* 2024;38(11):e23734.

587. Yan VC, Butterfield HE, Poral AH, et al. Why Great Mitotic Inhibitors Make Poor Cancer Drugs. *Trends in Cancer* 2020;6(11):924-41.

588. Sharpe LJ, Cook EC, Zelcer N, Brown AJ. The UPS and downs of cholesterol homeostasis. *Trends Biochem Sci* 2014;39(11):527-35.

589. Jang HH. Regulation of Protein Degradation by Proteasomes in Cancer. *J Cancer Prev* 2018;23(4):153-61.

590. Bao Y, Cruz G, Zhang Y, et al. The UBA1-STUB1 Axis Mediates Cancer Immune Escape and Resistance to Checkpoint Blockade. *Cancer Discov* 2025;15(2):363-81.

591. Li L, Zhou Y, Zhang Y, et al. A combination therapy of bortezomib, CXCR4 inhibitor, and checkpoint inhibitor is effective in cholangiocarcinoma in vivo. *iScience* 2023;26(3):106095.

592. Xu J, Brosseau J-P, Shi H. Targeted degradation of immune checkpoint proteins: emerging strategies for cancer immunotherapy. *Oncogene* 2020;39(48):7106-13.

593. Pellom ST, Jr., Dudimah DF, Thounaojam MC, et al. Bortezomib augments lymphocyte stimulatory cytokine signaling in the tumor microenvironment to sustain CD8+T cell antitumor function. *Oncotarget* 2017;8(5):8604-21.

594. Zhang X, Meng T, Cui S, et al. Roles of ubiquitination in the crosstalk between tumors and the tumor microenvironment (Review). *Int J Oncol* 2022;61(1).

595. Tawbi HA, Burgess M, Bolejack V, et al. Pembrolizumab in advanced soft-tissue sarcoma and bone sarcoma (SARC028): a multicentre, two-cohort, single-arm, open-label, phase 2 trial. *The Lancet Oncology* 2017;18(11):1493-501.

596. Gahvari Z, Parkes A. Dedifferentiated Liposarcoma: Systemic Therapy Options. *Current Treatment Options in Oncology* 2020;21(2):15.

597. Xu Y, Zeng J, Fu S, et al. The role of ubiquitination and deubiquitination in cancer lipid metabolism. *Front Oncol* 2025;15:1464914.

598. Xu G, Huang S, Peng J, et al. Targeting lipid metabolism in multiple myeloma cells: Rational development of a synergistic strategy with proteasome inhibitors. *British Journal of Pharmacology* 2021;178(23):4741-57.

599. Yan H, Ma Y-l, Gui Y-z, et al. MG132, a proteasome inhibitor, enhances LDL uptake in HepG2 cells in vitro by regulating LDLR and PCSK9 expression. *Acta Pharmacologica Sinica* 2014;35(8):994-1004.

600. Lee EJ, Kim MH, Kim YR, et al. Proteasome inhibition protects against diet-induced gallstone formation through modulation of cholesterol and bile acid homeostasis. *Int J Mol Med* 2018;41(3):1715-23.

601. Wu Z, Yang Y, Lei Z, et al. ABCB1 limits the cytotoxic activity of TAK-243, an inhibitor of the ubiquitin-activating enzyme UBA1. *FBL* 2022;27(1).

602. Glodkowska-Mrowka E, Mrowka P, Basak GW, et al. Statins inhibit ABCB1 and ABCG2 drug transporter activity in chronic myeloid leukemia cells and potentiate antileukemic effects of imatinib. *Experimental Hematology* 2014;42(6):439-47.

603. Patt A, Demoret B, Stets C, et al. MDM2-Dependent Rewiring of Metabolomic and Lipidomic Profiles in Dedifferentiated Liposarcoma Models. *Cancers (Basel)* 2020;12(8).

604. Zhou G, Li S, Xia J. Network-Based Approaches for Multi-omics Integration. In: Li S (ed.) *Computational Methods and Data Analysis for Metabolomics*. New York, NY: Springer US; 2020 p469-87.

605. Saha E, Fanfani V, Mandros P, et al. Bayesian inference of sample-specific coexpression networks. *Genome Res* 2024;34(9):1397-410.

606. Smith HB, Kim H, Walker SI. Scarcity of scale-free topology is universal across biochemical networks. *Sci Rep* 2021;11(1):6542.

607. Clote P. Are RNA networks scale-free? *Journal of Mathematical Biology* 2020;80(5):1291-321.

608. Schwikowski B, Uetz P, Fields S. A network of protein–protein interactions in yeast. *Nature Biotechnology* 2000;18(12):1257-61.

609. Ito T, Tashiro K, Muta S, et al. Toward a protein–protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences* 2000;97(3):1143-47.

610. Raman K, Damaraju N, Joshi GK. The organisational structure of protein networks: revisiting the centrality-lethality hypothesis. *Syst Synth Biol* 2014;8(1):73-81.

611. Schober P, Boer C, Schwarte LA. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia* 2018;126(5):1763-68.

612. Graham DS, Qorbani A, Eckardt MA, et al. Does "Low-Grade" Dedifferentiated Liposarcoma Exist? The Role of Mitotic Index in Separating Dedifferentiated Liposarcoma From Cellular Well-differentiated Liposarcoma. *Am J Surg Pathol* 2023;47(6):649-60.

613. Evans HL. Atypical Lipomatous Tumor, its Variants, and its Combined Forms: A Study of 61 Cases, With a Minimum Follow-up of 10 Years. *The American Journal of Surgical Pathology* 2007;31(1):1-14.

614. Kilpatrick SE. Dedifferentiated Liposarcoma: A Comprehensive Historical Review With Proposed Evidence-based Guidelines Regarding a Diagnosis in Need of Further Clarification. *Advances in Anatomic Pathology* 2021;28(6):426-38.

615. Jeon J, Nim S, Teyra J, et al. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Medicine* 2014;6(7):57.

616. Pacheco MP, Bintener T, Ternes D, et al. Identifying and targeting cancer-specific metabolism with network-based drug target prediction. *EBioMedicine* 2019;43:98-106.

617. Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* 2017;171(6):1437-52.e17.

618. Wang J, Tao X, Zhu J, et al. Tumor organoid-immune co-culture models: exploring a new perspective of tumor immunity. *Cell Death Discovery* 2025;11(1):195.

619. Kafri R, Dahan O, Levy J, Pilpel Y. Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. *Proc Natl Acad Sci U S A* 2008;105(4):1243-8.

620. Franks A, Airoldi E, Slavov N. Post-transcriptional regulation across human tissues. *PLoS Comput Biol* 2017;13(5):e1005535.

621. Edfors F, Danielsson F, Hallström BM, et al. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol* 2016;12(10):883.

622. Ponomarenko EA, Krasnov GS, Kiseleva OI, et al. Workability of mRNA Sequencing for Predicting Protein Abundance. *Genes* 2023;14(11):2065.

623. Greenbaum D, Colangelo C, Williams K, Gerstein M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biology* 2003;4(9):117.

624. Tang S, Wang Y, Luo R, et al. Proteomic characterization identifies clinically relevant subgroups of soft tissue sarcoma. *Nature Communications* 2024;15(1):1381.

625. Milighetti M, Krasny L, Lee ATJ, et al. Proteomic profiling of soft tissue sarcomas with SWATH mass spectrometry. *J Proteomics* 2021;241:104236.

626. Burns J, Wilding CP, Krasny L, et al. The proteomic landscape of soft tissue sarcomas. *Nature Communications* 2023;14(1):3834.

627. Picard M, Scott-Boyer MP, Bodein A, et al. Integration strategies of multi-omics data for machine learning analysis. *Comput Struct Biotechnol J* 2021;19:3735-46.

628. Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. Computational principles and challenges in single-cell data integration. *Nat Biotechnol* 2021;39(10):1202-15.

629. Subramanian I, Verma S, Kumar S, et al. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform Biol Insights* 2020;14:1177932219899051.

630. Argelaguet R, Velten B, Arnol D, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018;14(6):e8124.

631. Tommasini D, Fogel BL. multiWGCNA: an R package for deep mining gene co-expression networks in multi-trait expression data. *BMC Bioinformatics* 2023;24(1):115.

632. Morabito S, Reese F, Rahimzadeh N, et al. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data. *Cell Rep Methods* 2023;3(6):100498.

633. Algabri YA, Li L, Liu ZP. scGENA: A Single-Cell Gene Coexpression Network Analysis Framework for Clustering Cell Types and Revealing Biological Mechanisms. *Bioengineering (Basel)* 2022;9(8).

634. Su C, Xu Z, Shan X, et al. Cell-type-specific co-expression inference from single cell RNA-sequencing data. *Nature Communications* 2023;14(1):4846.

635. Gómez-Pascual A, Rocamora-Pérez G, Ibanez L, Botía JA. Targeted co-expression networks for the study of traits. *Scientific Reports* 2024;14(1):16675.

636. Liu B, Zhou H, Tan L, et al. Exploring treatment options in cancer: tumor treatment strategies. *Signal Transduction and Targeted Therapy* 2024;9(1):175.

637. Zhong L, Li Y, Xiong L, et al. Small molecules in targeted cancer therapy: advances, challenges, and future perspectives. *Signal Transduction and Targeted Therapy* 2021;6(1):201.

638. Wang L, Wang N, Zhang W, et al. Therapeutic peptides: current applications and future directions. *Signal Transduction and Targeted Therapy* 2022;7(1):48.

639. Mohi-ud-din R, Chawla A, Sharma P, et al. Repurposing approved non-oncology drugs for cancer therapy: a comprehensive review of mechanisms, efficacy, and clinical prospects. *European Journal of Medical Research* 2023;28(1):345.

640. Nhàn NTT, Yamada T, Yamada KH. Peptide-Based Agents for Cancer Treatment: Current Applications and Future Directions. *Int J Mol Sci* 2023;24(16).

641. Vadevoo SMP, Gurung S, Lee H-S, et al. Peptides as multifunctional players in cancer therapy. *Experimental & Molecular Medicine* 2023;55(6):1099-109.

642. Li CM, Haratipour P, Lingeman RG, et al. Novel Peptide Therapeutic Approaches for Cancer Treatment. *Cells* 2021;10(11).

643. Roy S, Bhattacharyya DK, Kalita JK. Reconstruction of gene co-expression network from microarray data using local expression patterns. *BMC Bioinformatics* 2014;15 Suppl 7(Suppl 7):S10.

644. Glass K, Huttenhower C, Quackenbush J, Yuan GC. Passing messages between biological networks to refine predicted interactions. *PLoS One* 2013;8(5):e64832.

645. Margolin AA, Nemenman I, Basso K, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 2006;7 Suppl 1(Suppl 1):S7.

646. Ernst J, Beg QK, Kay KA, et al. A semi-supervised method for predicting transcription factor-gene interactions in Escherichia coli. *PLoS Comput Biol* 2008;4(3):e1000044.

647. Faith JJ, Hayete B, Thaden JT, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 2007;5(1):e8.

648. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 2010;5(9).

649. Villaverde AF, Ross J, Banga JR. Reverse engineering cellular networks with information theoretic methods. *Cells* 2013;2(2):306-29.

650. Chou CH, Shrestha S, Yang CD, et al. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res* 2018;46(D1):D296-D302.

651. Portales-Casamar E, Arenillas D, Lim J, et al. The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res* 2009;37(Database issue):D54-60.

652. Pujato M, Kieken F, Skiles AA, et al. Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes. *Nucleic Acids Research* 2014;42(22):13500-12.

653. Alexiou P, Vergoulis T, Gleditzsch M, et al. miRGen 2.0: a database of microRNA genomic information and regulation. *Nucleic Acids Res* 2010;38(Database issue):D137-41.

654. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 2009;19(1):92-105.