The comprehensive proteomic characterisation of soft tissue sarcoma

Jessica Burns

A thesis submitted for the degree of
Doctor of Philosophy

September 2022

The Institute of Cancer Research
University of London

Declaration

The work presented in this thesis was completed under the supervision of Dr. Paul Huang in the Molecular and Systems Oncology team at the Institute of Cancer Research, London, United Kingdom.

I, Jessica Burns, confirm that the work presented within this thesis is my own. Information that has been derived from other sources is indicated within the thesis.

Signed

Jessica Burns

Contributions

The candidate was responsible for:

1. The collation of specimens for proteomic profiling and the extraction of protein/peptide from **most** specimens (number of samples is detailed in **section 3.2.2**).
2. **All** processing of mass spectrometry data after protein identification, including establishing the quality control and normalisation methods, and writing and implementing scripts.
3. **All** bioinformatic and statistical analyses of proteomic data, NanoString data, and immunohistochemistry data (including writing of scripts, interpretation of results, and production of figures).

Contributions were also made by several other people:

1. **All** clinicopathological data was collected and pseudonymised by Mr Chris Wilding, Dr Amani Arthur, Dr Vanessa Djabatey, and Ms Emma Perkins.
2. The construction of **all** tissue microarrays, associated immunohistochemistry staining and scoring was performed by Dr Alex Lee, Dr Cornelia Szecsei, and Ms Nafia Guljar, with guidance from Dr Khin Thway
3. **All** RNA extraction, and NanoString data collection was performed by Dr Alex Lee, Ms Nafia Guljar, and Ms Chanthirika Ragulan.
4. For proteomic analyses, **all** tandem mass tag labelling and high pH fractionation was performed by Dr Lukas Krasny, and **all** proteomic data was acquired by The Institute of Cancer Research Proteomics core facility (Dr Theo Roumeliotis and Prof Jyoti Choudhary)
5. The protein extraction of **some** samples was performed by Ms Martina Milighetti, and Mr Frank McCarthy (number of samples is detailed in **section 3.2.2**)

Abstract

Soft tissue sarcomas (STS) are a group of rare and heterogeneous mesenchymal malignancies. The extensive clinical and biological heterogeneity of STS complicates clinical disease management, and in the advanced setting prognosis is poor. Incomplete biological understanding of STS has long hampered efforts to drive clinical improvements for patients. At present, there is a lack of methods to stratify patients based on risk or their likelihood of treatment response. Additionally, there are limited targeted therapies available for STS patients, and current standard of care is largely a 'one size fits all' approach. Whilst the genomic, epigenomic, and transcriptomic basis of STS has been previously assessed, there is no proteomic understanding of the disease. Herein, my project conducts comprehensive proteomic profiling, by mass spectrometry, of 321 formalin-fixed paraffin-embedded primary tumour specimens from STS patients. This is the largest proteomic characterisation of STS to date and provides an overview of the baseline STS proteome. Specifically, heterogeneity in leiomyosarcoma was investigated, and 3 robust proteomic subtypes were identified. These molecular subtypes showed different functional biology and were associated with different survival outcomes, highlighting potential for risk stratification. Analysis of the immune landscape of undifferentiated pleomorphic sarcoma and dedifferentiated liposarcoma, highlighted a subpopulation of tumours with low lymphocyte infiltration and high complement activity. This revealed this complement cascade as a candidate therapeutic target. Finally, this project defined a protein-centric view of the STS proteome comprised of 'sarcoma proteome modules'. These modules transcended histological subtypes and covered a range of biological activities. Furthermore, modules were found to be associated with clinical outcome, again highlighting the potential for molecular risk stratification in STS. Overall, this project demonstrates the utility of comprehensive proteomic profiling in improving disease understanding, facilitating risk stratification, and identifying candidate therapies. In doing so, it establishes a rich resource for the STS research community.

Acknowledgements

I would like to thank my supervisors, Dr Paul Huang and Prof Robin Jones, for providing me with the opportunity to undertake this project, and for their advice and feedback throughout. My thanks are also extended to Dr Maggie Cheang and her team within the ICR Clinical Trial and Statistics Unit for their guidance.

Further to this, I would like to thank my colleagues, both past and present, within The Molecular and Systems Oncology team at the ICR. Their support, both scientific and personal, has been invaluable throughout my PhD experience. In particular, I would especially like to thank Nafia Guljar, one of my dearest friends. Our friendship has been one of the best things to have come out of the last 4 years.

Outside of work, I would like to thank all of my friends, and especially mention the wonderful women in my life: Maria, Georgie, and Sian. I would also like to thank my mum, brother, and sister for their support and love.

Finally, and by no means least, I would like to thank Joe Williams-Langley, for his continued encouragement and for looking after me whilst I have been writing up. Without him I would not have been able to complete this PhD.

Abbreviations

(A/E)RMS	(Alveolar/embryonal) rhabdomyosarcoma
2D-DIGE	2-dimensional difference gel electrophoresis
ABC	Ammonium bicarbonate
aCGH	Array-based comparative genomic hybridization
AJCC	American Joint Committee on Cancer
ALT	Alternative lengthening of telomeres
AS	Angiosarcoma
ASCO	American Society of Clinical Oncology
ASPS	Alveolar soft part sarcoma
AUC	Area under the curve
BCA	Bicinchoninic acid assay
BM	Basement membrane
BP	Biological process
CC	Consensus clustering
CCLG	Children's Cancer and Leukaemia Group
CCS	Clear cell sarcoma
CDF	Cumulative distribution function
CI	Confidence interval
CID	Collision induced dissociation
CIN	Chromosomal instability
CINSARC	Genome complexity index in sarcomas
CIx	Concordance index
CNA	Copy number alterations
CPTAC	Clinical Proteomic Tumour Analysis Consortium
CTX	Chemotherapy
DAC	Dacarbazine
DC	Dendritic cells
DDA	Data dependent acquisition
DDLPS	Dedifferentiated liposarcoma
DEP	Differentially expressed proteins
DES	Desmoid tumour
DFS	Disease free survival
DIA	Data independent acquisition
DNA	Deoxyribonucleic acid
DOC	Docetaxel
DOX	Doxorubicin
DSB	Double strand break
DSRCT	Desmoplastic small round cell tumour
DSS	Disease specific survival
DTT	Dithiothreitol
DV	Dependent variable

ECM	Extracellular matrix
ECOG	Eastern Cooperative Oncology Group
EMT	Epithelial-mesenchymal transition
EPS	Epithelioid sarcoma
FASP	Filter-Aided Sample Preparation
FDA	United States food and drug administration
FDR	False Discovery Rate
FF	Fresh frozen
FFPE	Formalin fixed paraffin embedded
FGFR	Fibroblast growth factor receptor
FISH	Fluorescence in situ hybridization
FPKM	Fragments per kilobase of exon per million mapped fragments
FSG	French sarcoma group
GeDDiS	*Trial name:* Gemcitabine and docetaxel versus doxorubicin as first-line treatment in previously untreated advanced unresectable or metastatic soft-tissue sarcomas
GEM	Gemcitabine
GGI	Genomic grade index
GIST	Gastrointestinal stromal tumour
GO	Gene ontology
GSEA	Gene Set Enrichment Analysis
H&E	Haematoxylin and eosin
HCD	Higher-energy C-trap dissociation
HDAC(i)	Histone deacetylase (inhibitor)
HIV	Human immunodeficiency virus
HR	Hazard ratio
HRD	HRR deficiencies
HRR	Homologous recombination repair
HHV-8	Human herpesvirus 8
IAA	Iodoacetamide
ICB	Immune checkpoint blockade
ICGC	International Cancer Genome Consortium
ICR	Immune constant of rejection
IFOS	Ifosfamide
IHC	Immunohistochemistry
IL	Interleukin
INT	Istituto Nazionale Tumori
IV	Independent variable
k-NN	k-nearest neighbour
LC	Liquid chromatography
LMS	Leiomyosarcoma
LOH	Loss of heterozygosity
LPS	Liposarcoma
LR	Low risk

LRFS	Local recurrence free survival
LRR	Local recurrence rate
MFH	Malignant fibrous histiocytoma
MFS	Metastasis free survival
MMR	Mismatch repair
MPNST	Malignant peripheral nerve sheath tumour
MRM	Multiple reaction monitoring
MS	Mass spectrometry
MS/MS	Tandem MS
MSC	Mesenchymal stem cell
MSI	Microsatellite instability
MSKCC	Memorial Sloan Kettering Cancer Center
MV	Missing value
MyFS	Myxofibrosarcoma
NF-1	Neurofibromatosis type 1
NCI	National cancer institute
NGS	Next generation sequencing
NHEJ	Non-homologous end joining
NHS	National health service
NK cells	Natural killer cells
NOS	Not otherwise specified
NTRK 1/2/3	Neurotrophic receptor tyrosine kinase 1/2/3
O-PDX	Orthoganol patient derived xenograft
OA	Overrepresentation analysis
OR	Overall response
ORR	Objective response rate
OS	Overall survival
PARP(i)	Poly (ADP-ribose) polymerase (inhibitor)
PC	Prinicpal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
PEComa	Malignant perivascular epithelioid cell tumour
PFS	Progression free survival
PH	Proportional hazard
PPI	Protein-protein interaction
PROSPECTUS	PROgnoStic and PrEdiCTive ImmUnoprofiling of Sarcomas
PS	Performance status
PTM	Post translational modification
QC	Quality control
RCT	Randomised controlled trial
RFS	Recurrence free survival
RMH	Royal Marsden Hospital
RNA	Ribonucleic acid

RNAseq	RNA sequencing
ROC	Reciever operator characteristic
RPPA	Reverse-phase protein microarray
RT	Rhabdoid tumour
RT-PCR	Reverse transcription-polymerase chain reaction
RTX	Radiotherapy
SAM	Significance analysis of microarrays
SIC	Sarcoma immune class
SPM	Sarcoma proteome module
SPS	Synchronous Precursor Selection
SRM	Single reaction monitoring
SS	Synovial sarcoma
ssGSEA	Single sample GSEA
stLMS	Soft tissue LMS (non-uterine)
STS	Soft tissue sarcoma
TAMs	Tumour associated macrophages
TANs	Tumour associated neutrophils
TAPUR	Targeted Agent and Profiling Utilization Registry
TCGA	The cancer genome atlas
TCGA-CDR	TCGA - Clinical Data Resource
TCGA-SARC	TCGA - Sarcoma study
Th (2/17)	T helper (2/17)
TIL	Tumour infiltrating lymphocytes
TKI	Tyrosine kinase inhibitor
TLA	Tumour linked alteration
TLS	Tertiary lymphoid structure
TMA	Tissue microarray
TMB	Tumour mutation burden
TME	Tissue microenvironment
TMT	Tandem mass tag
TOM	Topology overlap matrix
TPM	Transcripts per million
TRAB	Trabectedin
tSNE	t stochastic neighbour embedding
UCL	University College London
UICC	Union for International Cancer Control
uLMS	Uterine leiomyosarcoma
UMAP	Uniform manifold approximation and projection
UPS	Undifferentiated pleomorphic sarcoma
WDLPS	Well differentiated liposarcoma
WES	Whole exome sequencing
WGCNA	Weighted gene correlation network analysis
WGD	Whole genome duplication

WGS	Whole genome sequencing
WHO	World health organisation
KEGG	Kyoto encyclopaedia of genes and genomes
DSigDB	Drug Signature database
MSigDB	Molecular Signatures Database
BC	Bimodality coefficient
HDS	Hartigan's Dip Statistic

Table of Contents

List of Figures

Chapter 1

Chapter 3

Chapter 4

Chapter 5

List of Supplemental Figures

Chapter 4

Chapter 5

Chapter 6

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of Supplemental Tables

Chapter 4

Chapter 5

Chapter 1 Introduction

1.1 Soft tissue sarcoma overview

Soft tissue sarcomas (STS) are a group of rare malignancies accounting for less than 1% of all adult cancer diagnoses annually[1]. Incidence is higher in children under 14 years where STS accounts for 6-8% of cancers[2]. STS are mesenchymal in origin and can develop from any cell derived of the mesenchymal stem cell (MSC) lineage[3]. Accordingly, STS arise throughout the body and are a highly heterogeneous group of pathologies comprising over 80 different histological subtypes[4]. Further to the biological diversity of STS, extensive clinical heterogeneity is also present. STS tumours show variability in responses to treatment, and rates of local recurrence and distant metastasis[5]. This contributes to vast differences in disease progression between patients. The non-specificity of clinical symptoms in STS and its rarity means diagnosis and clinical management is challenging, particularly in non-specialist centres. Patients may present with advanced disease and frequently experience a prolonged period between presentation and confirmed diagnosis[6,7]. The 5-year overall survival (OS) rate for STS is 55-65%, however if distant metastases are present, OS is reportedly as low as 15%[8,9]. Following curative treatment for primary STS, approximately 50% of patients will go on to develop recurrent disease, however the rates of local recurrence and distant metastasis differ vastly based on histological subtype and anatomical site[10–12]. This extensive diversity of STS complicates attempts to better understand the disease and obscures efforts to translate biological findings to the clinic.

1.1.1 The origin and development of STS

In a subset of STS subtypes a cell of origin can be identified (**Figure *1.1***). For example, leiomyosarcoma (LMS) is derived from the myoblasts and is histologically representative of a smooth muscle tissue that has undergone alternate terminal differentiation[4]. Angiosarcoma (AS) arises from the vascular cell lineage and is specifically hypothesised to originate from the endothelial cells of the inner lining of blood and lymph vessels. Similarly, liposarcoma (LPS) is derived from the adipocytic cell lineage. However, many STS, such as undifferentiated pleomorphic sarcoma (UPS) and clear cell sarcoma (CCS) lack a defined cell of origin. Furthermore, the stage at which pathological transformation of mesenchymal cells is initiated is unclear[13]. STS exist along a spectrum of differentiation both between and within specific histological subtypes. Whether this is resultant of mutations acquired in primitive MSCs, partially differentiated progenitors, or both, remains to be defined. Few studies have investigated the evolutionary paths in

STS. In other cancer types, cellular transformation often occurs in a stepwise fashion, progressing from benign to 'pre-

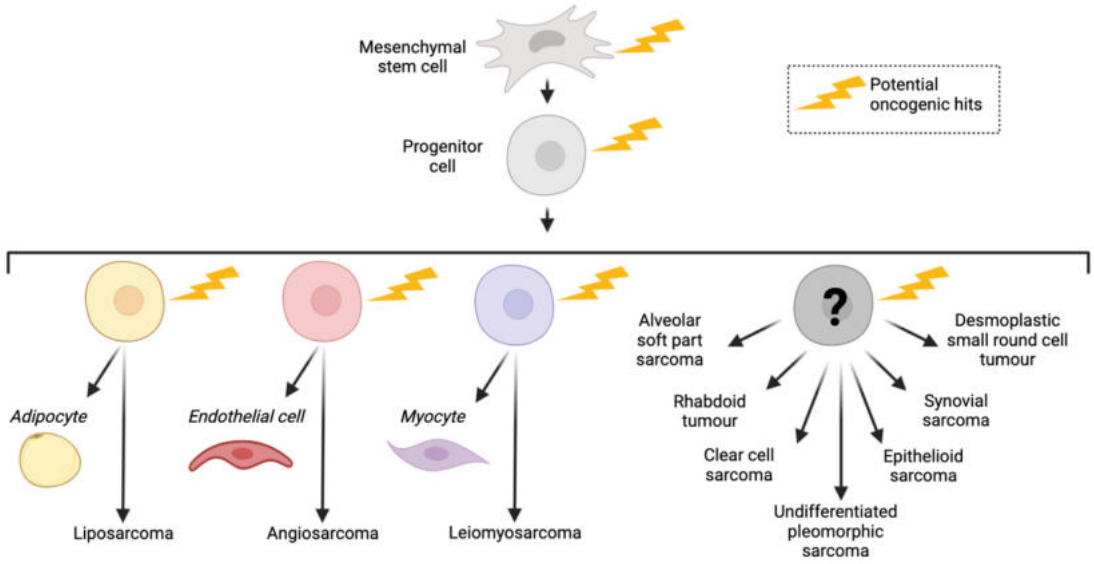


Figure 1.1 Soft tissue sarcoma (STS) cells of origin. Mesenchymal stem cell (MSC) lineage differentiation and associated STS diagnoses. '?' indicates unknown cell type. Schematic adapted from Gaebler et al[14].

cancerous' to malignant. For the most part this is not defined in STS. Moreover, in most cases, there is no identifiable causative risk factor for the development of STS. Unlike in many other cancers, lifestyle factors such as diet and smoking are not implicated in disease risk. Radiation, chemical, and viral exposure, chronic lymphoedema, and inherited syndromes such as familial retinoblastoma, neurofibromatosis type 1 (NF-1), and Li-Fraumeni syndrome can increase the likelihood of a STS diagnosis [4,15]. For example, kaposi sarcoma and LMS can arise as viral-associated STS, resultant of human herpesvirus 8 (HHV-8) and Epstein-Barr virus infections respectively[16,17]. Whilst viral infection can predispose an individual to sarcoma, infection alone is not causative. For example, Kaposi sarcoma arises most commonly in HHV-8-infected individuals with an advanced human immunodeficiency virus (HIV) infection. This is resultant of the weakened immune system of a HIV-positive individual, which allows the HHV-8 virus to multiply largely unchallenged. AS can arise as a radiation-associated STS, occurring secondary to treatment for breast cancer[18,19], and malignant peripheral nerve sheath tumours and gastrointestinal stromal tumours (GIST) can arise as NF-1-associated STS[20,21]. At present, these aetiologically certain tumours represent a minority. However, the recent International Sarcoma Kindred Study identified inherited pathogenic genetic

variants in 55% of patients (n = 1162)[22]. This challenges the current theory that most STS are sporadic and increases the potential utility for disease screening programmes.

1.1.2 Classification of STS

1.1.2.1 Tissue-based definitions

STS classification has historically been based on the tissue type the tumour best represents. The most recent World Health Organisation (WHO) STS classification system describes benign, intermediate, and malignant diagnoses, grouped into 12 categories based on tissue lineage[4]. Classification systems are considered a vital tool in improving diagnostics and therapeutic decision making, and the WHO STS system is implemented worldwide to define STS. Categories include 'adipocytic tumours', '(myo)fibroblastic tumours', 'vascular tumours', 'perivascular tumours', 'smooth muscle tumours', 'skeletal muscle tumours', 'gastrointestinal stromal tumours', 'chondro-osseous tumours', 'fibrohistiocytic tumours', and 'peripheral nerve sheath tumours'. In addition, 'tumours of uncertain differentiation' is used as a category of exclusion; grouping STS that do not resemble a specific tissue. Disease complexity means classification into the WHO categories is ever-changing. There were several key advances between the 2013 WHO classification system and most recent in 2020 [4,23]. Firstly, a new 'undifferentiated small round cell sarcomas' category was established to encompass Ewing sarcoma and 3 molecular subtypes with specific genetic profiles; marking recognition of the different clinicopathologic features of these tumours. Secondly, neurotrophic receptor kinase (*NTRK*)-rearranged spindle cell neoplasms were listed as an emergent diagnosis for the first time. Thirdly, several diagnoses were removed from the 'fibrohistiocytic tumour' category, whose necessity continues to be debated due to the ambiguous role of fibrohistiocytic differentiation.

1.1.2.2 Genomic-based definitions

Complementary to tissue-based definitions, STS can also be classified based on genomic complexity. Genomic complexity in STS exists on a spectrum, within which histological subtypes fall into 2 broad groups: pathologies with simple genomic profiles and pathologies showing high complexity[24–26]. Genomically simple STS are typified by a largely unaltered genome with simple alterations such as translocations or activating mutations. For example, synovial sarcoma (SS), CCS, desmoplastic small round cell tumour (DSRCT), and alveolar soft part sarcoma (ASPS) are all genomically simple STS driven by translocation events that result in aberrant transcriptional activity[27–34]. Whilst gastrointestinal stromal tumours (GIST), rhabdoid tumours (RT) and epithelioid sarcoma

(EPS) are genomically simple STS harbouring highly recurrent mutations. The identification of driver events and specific molecular characteristics in genomically simple STS has enabled the development of robust diagnostic methods and revealed candidate oncogenic pathways for therapeutic intervention.

In contrast to genomically simple STS, STS with complex genomes have high genomic instability resulting in wide-ranging genetic aberrations, unbalanced karyotypes and few recurrent alterations between patients[35]. These tumours show high mutational burden compared to genomically simple STS. Although when compared to other cancer types, mutational burden is still relatively low[36]. In a subset of genomically complex STS subtypes, recurrent genetic aberrations have been identified across patients. For example, ring chromosomes consisting of amplified material of the 12q13-15 region is characteristic of well differentiated LPS (WDLPS) and dedifferentiated LPS (DDLPS) tumours, and its detection used for diagnosis[37–40]. By contrast, in most genomically complex subtypes such as LMS and UPS, extensive chromothripsis, kataegis, genome duplication and aneuploidy/copy number alterations (CNA) result in little genomic concordance between patients[41–43]. The absence of specific molecular features in most genomically complex STS means these patients have been unable to benefit from developments in molecular diagnostics, and therefore are reliant on histological interpretation. Furthermore, due to a lack of common molecular characteristics, identifying actionable targets for therapeutic intervention in these tumours is a demanding task. Accordingly, recent efforts to molecularly profile and target these complex malignancies, have focused on identifying multi-gene signatures or key aberrant signalling axes involved in tumour maintenance and progression[35,36].

1.2 Diagnosis and management of STS

1.2.1 Diagnosis of STS

Histopathological examination by morphological inspection and/or immunohistochemistry (IHC) is the gold standard diagnostic method in STS[4,44]. In recent decades, molecular tests have become more commonplace as companion analyses alongside histopathological review. Routine molecular tests include: fluorescence in situ hybridisation (FISH) where fluorescent-labelled probs are used to establish the presence of absence of a specific DNA/RNA sequence; array-based comparative genomic hybridization (aCGH) where patient and reference DNA samples are compared to identify genome wide copy number changes; and reverse transcription-polymerase chain reaction (RT-PCR) where the presence of specific mRNA regions is assessed through

RNA amplification. In STS, these methods have shown most utility in diagnosing subtypes with simple genomes where an identifiable and characteristic genomic alteration has been described. For example, every suspected SS tumour is molecularly assessed by FISH and/or RT-PCR to assess for *SS18-SSX1/2/4* fusion presence. In addition to establishing and confirming diagnoses, molecular testing is also a vital tool in STS for diagnosis exclusion.

1.2.2 Risk stratification in STS

Present clinical management of STS is complex and, in many cases, poorly defined (**section 1.2.3**). STS surgery can carry a high morbidity risk, particularly in elderly patients or when implemented for large tumours in complex anatomical sites. Moreover, chemotherapies and many of the targeted therapies in development have significant associated toxicities, for which many patients see little to no benefit. Risk stratification aims to quantify the likelihood of a patient experiencing a harmful event, be it treatment complications or disease progression (recurrence, metastasis, death). Medical risk stratification must be carefully balanced with the potential effectiveness of intervention, as well as the psychological and social health of a patient. However, if implemented well, risk stratification has the potential to better inform patient-clinician discussions, support decisions on therapy pathways, enable suitable post-operative planning, and identify disease monitoring needs.

1.2.2.1 Key clinicopathological variables

Risk stratification is not a new concept in oncology. Clinicopathological data such as tumour grade, size, and depth, are routinely recorded for each patient and aid clinical understanding of how advanced a disease is. Whilst this does not provide formal stratification, clinical understanding guides interventional decision-making (ie. predictive stratification; **Figure 1.2A**) and acts as an informal risk assessment for disease-related events (ie. prognostic stratification; **Figure 1.2B**). Large-scale retrospective studies have assessed the prognostic value of clinicopathological variables in STS. In general, male patients with high grade, large, deep tumours, where distant metastases are present at diagnosis, surgical resection is incomplete (positive margins), and Eastern Cooperative Oncology Group (ECOG) performance status (PS) is poor (i.e ≥ 2), have the poorest outcomes (**Table *1.1***)[45,46]. Other important factors in determining outcome include histological subtype and anatomical site. Different STS subtypes show different propensities for metastasis and local recurrence. For example, retrospective analyses report local recurrence free survival (LRFS) for SS patients as low as 6.1% and metastasis free survival (MFS) as 24.1%, whilst LRFS in DDLPS is reported between 41

- 80% and MFS between 14 - 17%[47–50]. Tumours of the same diagnosis in different anatomical locations also show differing outcomes. For example, retroperitoneal DDLPS show a poorer 5-year LRFS than extremity DDLPS (20% vs 62%)[47,51].

At present, tumour size and grade are the single 2 most important measures clinicians use to determine prognosis[44,52]. Tumour size has been demonstrated by multiple studies to be a strong positive predictor for MFS and OS[45,46]. Size is commonly categorised into tumours ≤ 5 cm (at maximum dimension), and those > 5 cm, representing patients with a lower risk and higher risk respectively. However, the relationship between size and

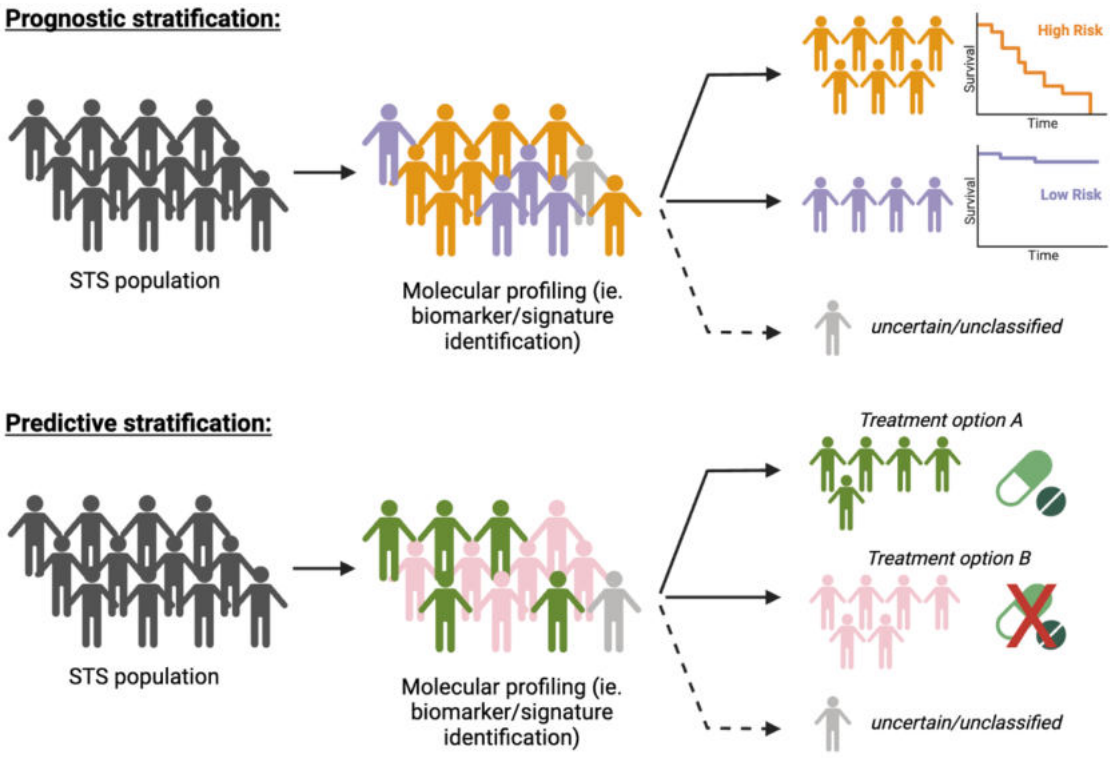


Figure 1.2 Diagrammatic explanation of prognostic and predictive stratification
Prognostic stratification identifies patients at high risk of a particular clinical event (e.g., death), whilst predictive stratification can identify patients most likely to benefit from a treatment intervention.

outcome is more complex than it may seem. The increased risk associated with increased tumour size can taper off or invert in extremely large tumours. This reversal of risk-size relationship is likely due to the largest tumours having a more indolent progression which enables the lesion to persist to such an extreme size. Grading can be performed by the National Cancer institute (NCI) or French Federation of Cancer Center Sarcoma Group (FNCLCC) systems, with the latter most often implemented[53,54].

Table 1.1 Overview of prognostic factors in soft tissue sarcoma (STS). Univariable analysis (UVA) results detailed by p value. Multivariable analysis (MVA) results detailed by comment on the specific significant category within each variable. **.** Data from 2 large-scale retrospective studies[45,46]. Abbreviations: FS = fibrosarcoma; LPS = liposarcoma; MFH = malignant fibrous histiocytoma; LMS = leiomyosarcoma; MPNST = malignant peripheral nerve sheath tumour; SS = synovial sarcoma; NOS = not otherwise specified; RMS = rhabdomyosarcoma; FNCLCC = French Federation of Cancer Center Sarcoma Group; UICC = Union for International Cancer Control; AJCC = American Joint Committee on Cancer; PS = performance status; RP = retroperitoneal; CTX = chemotherapy; RTX = radiotherapy; OS = overall survival; LRFS = local recurrence freesurvival; MFS = metastasis free survival; n.r = not reported

		OS		LRFS		MFS		Study ref
Variable	Categorisation	Significant (UVA)	Significant (MVA)	Significant (UVA)	Significant (MVA)	Significant (UVA)	Significant (MVA)	
Age	≤ 50 or > 50 years	Y 0.05	Y Older = poorer outcome	N n.r	N -	N n.r	N -	46
	≤ 50 or > 50 years	Y 0.004	N -	N 0.900	N -	N 0.500	N -	45
Sex	Female, Male	Y 0.03	Y Male = poorer outcome	N n.r	N -	N n.r	Y Male = poorer outcome	46
	Female, Male	Y 0.002	Y Male = poorer outcome	N 0.300	N -	N 0.200	N -	45
History	Positive, negative	N 0.330	N -	N 0.140	N -	N 0.800	N -	45
Histology	FS, LPS, MFH, LMS, MPNST, SS, NOS, Other	N n.r	N -	Y 0.02	N -	N n.r	Y LPS, MFH, NOS = improved outcome compared to FS	46
	MFH & FS, LPS, LMS, RMS, SS, MPNST, NOS, other	Y 0.002	N -	N 0.070	N -	Y 0.001	N -	45
Grade	1,2,3 (FNCLCC)	Y <0.0001	Y 1 = improved outcome & 3 = poorer outcome	Y 0.002 (for 2 v 3)	Y 1 = improved outcome & 3 = poorer outcome	Y <0.0001	Y 3 = poorer outcome	45
	1,2,3,4	Y <0.001	Y Higher (2,3,4) = poorer outcome	Y 0.04	Y Higher (2,3,4) = poorer outcome	Y <0.001	Y Higher (2,3,4) = poorer outcome	46
Depth	Superficial, deep	Y <0.0001	Y Deep = poorer outcome	Y 0.040	Y Deep = poorer outcome	Y <0.0001	Y Deep = poorer outcome	45
Site	Extremity, trunk wall, head & neck, RP, viscera, other	N n.r	Y Trunk, head & neck, RP, viscera = poorer than extremity	N n.r	N -	N n.r	N -	46
	Extremity, trunk wall, head & neck, internal trunk	N 0.200	N -	Y 0.040	N -	N 0.500	N -	45
Size	≤ 5 cm, > 5cm	Y <0.001	Y Larger = poorer outcome	N n.r	N -	Y <0.01	Y Larger = poorer outcome	46
	< 5 cm, 5-9cm, >10cm	Y <0.0001	Y < 5 cm = improved outcome	Y 0.050	Y > 10 cm = poorer outcome	Y <0.0001	N -	45
Stage (UICC/AJCC)	I,II,III,IV	Y <0.001	Y More advanced (II, III, IV) = poorer outcome	Y 0.02	Y More advanced (II, III, IV) = poorer outcome	Y <0.001	Y More advanced (II, III, IV) = poorer outcome	46
	I,II,III,IV	Y <0.0001	N -	Y 0.02	N -	Y <0.0001	N -	45
Bone/ neurovascular involvement	Positive, negative	Y <0.0001	N -	N 0.900	N -	Y <0.0001	N -	45
Nodal involvement	Positive, negative	Y 0.002	N -	N 0.400	N -	N 0.4	N -	45
PS	0,1,≥2	Y <0.001	Y Highest (≥2) = poorer outcome	N n.r	N -	N n.r	N -	46
Surgery	Good, poor	N 0.35	N -	Y 0.007	N -	N 0.21	Y Poor = poorer outcome	45
	Negative margins, positive margins	Y <0.01	Y Positive = poorer outcome	Y <0.01	Y Positive = poorer outcome	N n.r	N -	46
Adjuvant treatment	CTx: Yes, no	Y 0.07	Y Yes = improved outcome	N 0.2	Y Yes = improved outcome	N 0.3	N -	45
	RTx: Yes, no	N 0.4	N -	Y <0.0001	N -	Y 0.05	Y Yes = improved outcome	45
LR event	Yes, no	N n.r	N -	- -	- -	N n.r	N -	46
DM present	Yes, no	Y <0.001	Y Yes = poorer outcome	N n.r	N -	- -	- -	46

FNCLCC grading integrates cellular differentiation status, mitotic count, and the level of necrosis present. This generates a score mapping to grades of 1, 2, or 3, which show an increasing likelihood of metastasis and death from 1 to 3. This system is predominantly used to select patients for adjuvant chemotherapy, yet limitations exist. Grading offers minimal use for putative high-grade diagnoses (eg., ASPS, CCS, and EPS), and the 3-tier system results in an uninformative intermediate group with high uncertainty regarding tumour aggressiveness. Furthermore, due to intra-tumoural heterogeneity, grading is challenging in the limited diagnostic biopsy material[55]. This often results in under-grading of STS, particularly in LMS, and by extension denies patients treatment that may be beneficial[56].

Whilst individual prognostic factors do enable clinicians to identify high-risk patients, simply summing the risk factors fails to consider cofounding variables, interacting variables, or multicollinearity. Therefore, the relative value of such factors is unclear. In most cancer types, tumour staging following the American Joint Committee on Cancer/Union for International Cancer Control (AJCC/UICC) TNM system is a valuable proxy for general disease risk, which incorporates tumour type, site, size, lymph node status, and metastasis extent. However, until 2017 the TNM system considered STS as a single disease type, and thus had little practical use for the highly heterogeneous population of STS patients[57]. Improvements in the 8th AJCC/UICC edition included delineations between STS of different anatomical sites yet did not integrate this with histology[58]. As a result, staging of STS of not routinely performed.

1.2.2.2 Nomograms

One way to objectively judge risk is with medical nomograms. Nomograms translate complex statistical models into a graphical representation and interpretable numeric value (**Figure 1.3**). Identifying a patient-specific prognostic value offers improved utility than grouped risk stratification (e.g., grading, staging). Nomograms for clinical outcome are most commonly built using the Cox regression model and output the probability of a specific event (e.g., 5-year MFS).

In STS there are numerous pan-subtype and subtype-specific nomograms assessing a range of outcome measures. However, many of the published nomograms have been built from single institution data, thus performances are often assessed by internal validation methods only. Most STS nomograms are therefore more appropriately thought of as exploratory and their use in clinical practice is not recommended. The list of STS

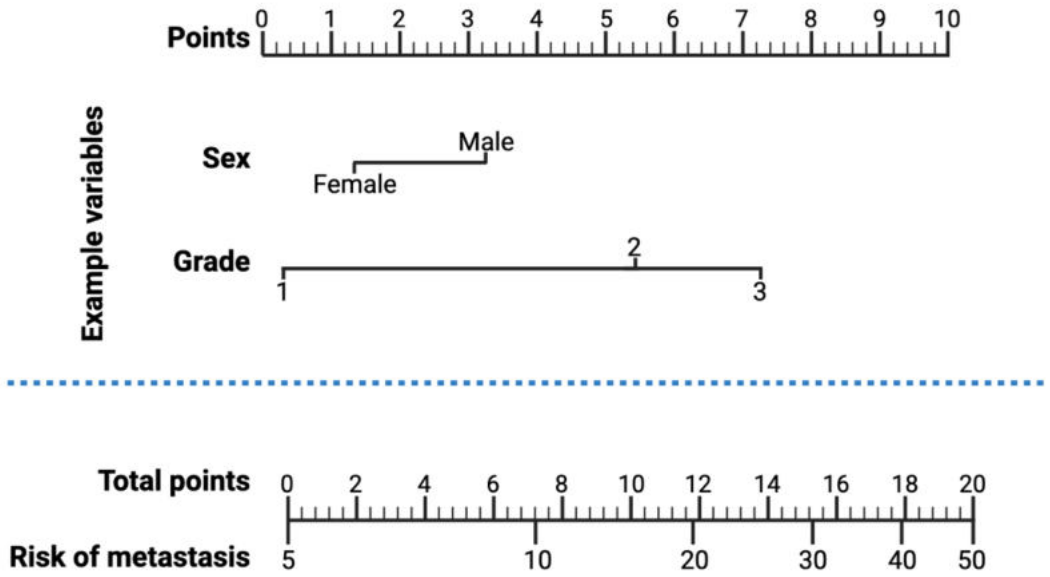


Figure 1.3 Diagrammatic example of a nomogram. For each example variable, a value is assigned based on the top ruler. The values are summed, and the total mapped on the bottom ruler to identify a corresponding risk value.

nomograms which have undergone a higher level of statistical rigor is far shorter. One of the most robust and widely applicable is the Memorial Sloan Kettering Cancer Center (MSKCC) pan-subtype disease specific survival (DSS) nomogram[59]. This utilises age at diagnosis, tumour depth, grade, size, anatomical site, and histology to predict 4/8/12-year DSS following surgery of primary localised disease. Internal validation showed a concordance index (CIx) of 0.77, and subsequent external validation studies achieved CIs from 0.71 to 0.78 in a range of different cohort[10,60–66]. The highest CIx (0.78) was achieved in a head and neck specific cohort, which evidenced nomogram superiority over AJCC/UICC TNM (CIx = 0.71)[66]. Performance was subsequently shown as weaker in paediatric patients and Asian populations, where the nomogram underpredicted mortality[67–69]. This was likely due to the absence of these populations in the training cohort. Further to cohort specific performance differences, the limitations of this nomogram include the use of a malignant fibrous histiocytoma (MFH) diagnosis. Succession of the MFH diagnosis by UPS (**section 1.4.2**) has rendered the histology variable of this nomogram outdated. Furthermore, the nomogram uses high/low grading as opposed to the FNCLCC grading system, which is widely considered a superior correlate for outcome[70–72]. Overall, this pan-subtype nomogram applies well to the general STS population. Yet, some features are now sub-optimal and variations in

performance supports the need for more tailored nomograms reflective of STS heterogeneity.

Site-specific nomograms attempt to account for disease heterogeneity. The most thoroughly validated nomogram for extremity and trunk wall STS were developed at Istituto Nazionale Tumori (INT)[73]. These nomograms predict 5/10-year OS and distant metastasis following surgery for localised primary disease. Importantly, detailed histology data (9 categories) is integrated in the model. Development involved single site data and reported CIx of 0.77 (OS) and 0.76 (distant metastasis). External validation in 3 independent datasets from other institutions showed good nomogram calibration (CIx = 0.64 – 0.81 (OS); CIx = 0.61 – 0.79 (distant metastasis)), highlighting these as reliable methods for prognostication. The same team at INT also developed and validated a retroperitoneal specific nomogram for the prediction of OS and disease-free survival (DFS)[74,75]. As with the extremity nomograms, detailed histological information (7 categories) was included as a covariate. Internal validation reported a CIx of 0.74 (OS) and 0.71 (DFS), with external validation showing good concordance (OS: 0.67 - 0.73, DFS: 0.68 - 0.69). The performance of this retroperitoneal nomogram is superior to any other published method for prognostication, and thus it was included in the 8th edition of the AJCC/UICC manual[58]. Collectively these subtype specific nomograms from INT are referred to as the 'Sarculator' nomograms, and have been rendered in app format, making them readily accessible to clinicians[76]. Akin to the site-specific nomograms, are histology-specific nomograms such as the desmoid tumour (DES) MSKCC 3/5/7-year LRFS nomogram[77]. First-line treatment for DES is active surveillance as opposed to resection[78]. This nomogram is therefore particularly useful as, in contrast to many STS nomograms, all variables required for prediction can be obtained without surgical intervention. There have also been efforts to develop nomograms specific to both site and histology. For example, the MSKCC uterine LMS (uLMS) nomogram for 5-year OS prediction[79]. Here, model performance validated well both internally and externally, however cohort sizes were limited, resulting in poor statistical power. This demonstrates that while capturing STS heterogeneity through increasing specificity may improve the accuracy of the model, the practicality of sub-selection within an already rare disease introduces challenges.

1.2.3 Treatment of STS

1.2.3.1 Surgery and radiotherapy

In the primary localised disease setting, surgical resection is the mainstay of treatment for curative intent[80]. Studies report surgical margins microscopically negative for tumour material (ie. negative/R0 margins) as associated with improved local recurrence rates, and in some cases OS[81,82]. However, optimal margin distances are not well defined due to the anatomical and histological heterogeneity of STS. Furthermore, dependent on the tumour location, achieving negative margins can be technically challenging. By contrast, one of the most controllable factors for improved surgical outcomes that has been reported is centralised surgery within a multidisciplinary care team at specialist sarcoma centres[83–86].

Surgical intervention is challenging for large tumours or those in complex anatomical sites, where an aggressive approach may result in multi-visceral or limb loss. Accordingly, the use of radiotherapy (RTX) to reduce tumour burden prior to surgical intervention has been assessed in STS. The phase III STRASS randomised controlled trial (RCT) is the largest trial to date assessing the use of pre-operative RTX combined with surgery compared to surgery only in retroperitoneal STS[87]. In 2022, STRASS reported that at 3 years follow-up no significant difference was found in median LRFS between surgery only and surgery+RTX (LRFS: hazard ratio (HR) (95% confidence intervals (CI)) = 1.01 (0.71-1.44), p = 0.95). Despite overall negative results, exploratory subgroup analyses did highlight a potential role for pre-operative RTX in LPS and low-grade tumours, however these cohorts were underpowered. Similar trials focused on extremity STS show improved local control in patients who receive surgery+RTX compared to surgery alone (local recurrence rate (LRR) = 1.4% vs 25%)[88]. However, the presence of a link between local control and overall patient outcome is debated, and no significant improvement in OS was seen (20-year OS = 64% surgery alone vs 71% surgery+RTX (p = 0.22)).

In addition to pre-operative use, RTX can also be leveraged in the post-operative setting. In cases where R0 margins are not achievable, RTX has been shown as associated with reduced LRR (surgery only = 23.9 vs surgery+RTX = 1.4%)[89]. However, comparisons between pre- and post-operative regimens have highlighted minimal differences in OS (p = 0.048) and showed pre-operative RTX as associated with a greater risk of wound complications[90].

1.2.3.2 Conventional chemotherapy

For metastatic and unresectable local STS, treatment is palliative in intent and centred on systemic chemotherapy. 1st line therapy in most advanced patients is the anthracycline doxorubicin (DOX), used in combination with ifosfamide (IFOS) where high physical tumour burden is present. RCTs have extensively assessed DOX in combination with other chemotherapeutics (IFOS, cisplatin, cyclophosphamide, dacarbazine (DAC), mitomycin-C, streptozotocin, vincristine, vindesine), and consistently concluded that combination regimens drastically increase toxicity rates, for little to no improvement in patient outcome[91–95]. Where DOX is contraindicated, a gemcitabine (GEM) + docetaxel (DOC) regimen can be used. Phase II studies have shown activity for GEM+DOC, particularly in uLMS; where 53% of patients (n = 34) achieved overall response (OR)[96,97]. However, in unselected STS populations DOX is considered superior. This has been evidenced by the phase III multi-subtype GeDDiS trial[98]. GeDDiS showed no significant difference in progression free survival (PFS) between DOX and GEM+DOC (HR (95% CI) = 1.28 (0.98-1.67), p = 0.07), and reported significant toxicity in the GEM+DOC arm. Accordingly, DOX remains the 1st line choice for most advanced STS patients.

Disease control is a clinically relevant endpoint in advanced STS. However, the high cumulative toxicity of DOX±IFOS and GEM+DOC means such interventions are not suitable for long-term disease stabilisation. In contrast, trabectedin (TRAB) has low cumulative toxicity and as such is considered a 2nd-3rd line therapy choice[99]. OR to TRAB across STS subtypes is low (8%), however TRAB shows improved responses for LPS, LMS and translocation STS[100]. A phase II RCT assessing translocation STS, showed a significantly longer median PFS in myxoid/round-cell LPS and SS patients who received TRAB (3.1 months) compared to best supportive care (1.5 months)[101]. Furthermore, a phase III RCT assessing LPS and LMS patients revealed a superior median PFS in those receiving TRAB (4.2 months) compared to DAC (1.5 months) [102]. However, no significant difference in OS was observed (TRAB = 12.4 months, DAC = 12.9 months). Other frequently used chemotherapies include the microtubule inhibitor eribulin mesylate, which was approved by the United States food and drug administration (FDA) in 2016 for advanced LPS[103], and paclitaxel which is a 1st line choice for AS patients[104].

1.2.3.3 Tyrosine kinase inhibitors

Tumours often show high dysregulation across many cellular signalling pathways regulated by tyrosine kinase activity. Therefore, there has been much investigation into the use of tyrosine kinase inhibitors (TKIs) in STS[105]. TKI specificity ranges extensively, from broad-spectrum and multi-target to highly specific[106]. Multi-target TKIs are beneficial when tumours possess complex genetic profiles and show wide ranging aberrations in signalling. Whereas high specificity TKIs offer improved utility in tumours where kinase driver alterations are identified. At present several TKIs are FDA-approved for use in select STS patients, such as imatinib, pazopanib, larotrectinib, and entrectinib.

Imatinib and pazopanib both show multi-target TKI profiles. Imatinib was developed to inhibit the breakpoint cluster region (*BCR*)-*ABL* fusion in chronic myeloid leukaemia patients and was the first FDA-approved targeted therapy[107]. Subsequent pharmacologic profiling evidenced a targeting profile that extended beyond ABL to include KIT proto-oncogene and platelet derived growth factor receptor alpha/beta (PDGFRA/B)[108]. Approximately 75% of GIST patients harbour a *KIT* mutation, and 10% a *PDGFRA* mutation[109]. Consequently, in the advanced GIST population imatinib can achieve remarkable results, with complete responses seen in approximately 5% of patients, partial responses in 65 - 70%, and stable disease in 15 - 20%[110,111]. Imatinib is therefore the current 1st-line choice for metastatic GIST. Pazopanib has a far broader target spectrum than imatinib. One family of kinases pazopanib has highest specificity for are the vascular endothelial growth factor receptors (VEGFRs). In targeting VEGFRs, pazopanib is considered to elicit most of its anti-tumour effects through the inhibition of angiogenesis[112]. Pazopanib was approved and integrated as a 2nd-3rd line choice in UK care for advanced non-adipocytic STS based on the PALETTE phase III RCT[113]. PALETTE found significantly improved PFS in pazopanib-treated patients compared to placebo-treated (median = 4.6 months vs 1.6 months). However long-term monitoring failed to evidence any improvement in OS, and as a result pazopanib is no longer approved as routine care in the UK[114]. Interestingly, subsequent combined analyses of PALETTE patients and patients on the preceding phase II pazopanib trial[115], identified a sub-population of long term pazopanib responders and survivors; with 34% (166/266) of patients on pazopanib surviving ≥ 18 months[116].

In contrast to imatinib and pazopanib, larotrectinib and entrectinib are highly specific TKIs. These TKIs target NTRK1/2/3 and are only the 2nd and 3rd tissue-agnostic cancer drugs to be granted FDA-approval. Approximately 1% of adult STS are *NTRK* fusion positive and significant anti-tumour effects have been demonstrated in NTRK inhibitor

trials. Phase I/II tissue-agnostic RCTs assessing larotrectinib have shown an 88% objective response rate (ORR) in *NTRK* fusion positive STS patients (n = 68)[117–120]. Tissue agnostic entrectinib trials have reported responses in 46% of sarcoma patients (n = 13)[121], and further trials are ongoing[122,123].

The impressive responses seen to imatinib, larotrectinib and entrectinib exemplify the utility of biomarker-driven precision medicine. Molecular stratification of patients for these interventions based on *KIT/PDFRA* mutational status (imatinib) or *NRTK* fusion status (larotrectinib/entrectinib) can identify populations most likely to benefit. Ultimately, this results in appropriately tailored clinical trials, which translate to both statistically and clinically significant outcomes. Moreover, tissue agnostic use of NTRK inhibitors illustrates how molecular characteristics can transcend diagnoses and function as targets for pan-cancer therapy. However, these examples are few in number and restricted to genomically simple STS. Achieving such impressive outcomes in poorly characterised complex STS tumours is far more complicated.

1.2.3.4 Immunotherapy

Over the recent decades, there has been huge investment in cancer immunotherapy, which has been hailed as a revolutionary treatment approach in oncology. Immunotherapy describes a large group of strategies spanning adoptive cell therapy (ACT), oncolytic viruses, macrophage and cancer antigen targeting, and tumour vaccinations; all of which have been assessed in STS clinical trials[124–128]. Of particular interest in the sarcoma field is ACT, which most commonly involves the engineering of T cells to target tumour-specific antigens. In sarcoma, T cell therapy was first evaluated in osteosarcoma and Ewing's sarcoma patients (n = 17) with the use of a HER2-targetting cells[129]. This phase 1 trial demonstrated safety, and stable disease was observed in 4 patients for up to 4 months. Many more ACT trials have been initiated since, and this technology holds huge promise[130,131]. However, at present it is the immune checkpoint blockade (ICB) drugs that have gained most traction within STS. ICB target checkpoint receptors on immune cells, cancer cells and other tumour supporting cells. Checkpoint receptors inhibit and suppress anti-tumour immune responses, therefore ICB restore immune activity by blocking receptor-ligand interactions within the tumour environment. The expression of checkpoint receptors in STS varies between and within histological subtypes. DDLPS, UPS, myxofibrosarcoma (MyFS) and LMS show generally higher checkpoint receptor expression than translocation-associated STS[132–135]. Although reports are not consistent. This alludes to a potential 'immune hot' STS population of mixed histological subtypes. However, the level of immune activity in even the most

immune hot STS is not comparable to other cancer types such as melanoma where ICB achieves outstanding responses[136–138]. Despite this, is it hoped that a subset of STS patients may benefit from ICB therapy, particularly when used as part of a combination strategy.

One of the most widely assessed ICBs is pembrolizumab, which was the 1st tissue-agnostic drug to be approved by the FDA. Pembrolizumab targets programmed cell death 1 (PD1) on T-cells, to block interactions with programmed cell death ligand 1 (PD-L1) on tumour cells and facilitate an immune response. In STS, the SARC028 RCT evaluated pembrolizumab use in a multi-subtype population[139]. SARC028 reported an ORR of 17.5% corresponding to 7/40 STS patients, 6 of which had a UPS or DDLPS diagnosis. Subsequent expansion of SARC028 to recruit further UPS and LPS patients noted objective responses in 9/40 UPS and 4/39 LPS patients[140]. Pembrolizumab has also shown promising activity as a combination therapy with the VEGFR inhibitor axitinib[141]. In a phase II multi-subtype RCT, over 65% of patients achieved 3 months PFS on pembrolizumab+axitinib. ASPS patients (n = 12) saw most benefit, with 55% achieving partial response and 18% stable disease at 3 months. Recently, the novel PD1 inhibitor TQB2450 has also shown promise in ASPS. When assessed in combination with the TKI anlotinib, ASPS patients showed an ORR of 75% and median PFS of 23 months[142]. Together, these studies highlight the utility of ICB and TKI combination strategies in ASPS.

Alternate to pembrolizumab, is nivolumab, another anti-PD1 ICB. The Alliance A091401 study compared nivolumab±ipilimumab, an anti- cytotoxic T-lymphocyte associated protein 4 (CTLA4) ICB[143]. This study showed a median PFS comparable to standard chemotherapy in the combination arm (~ 4 months). Study expansion revealed UPS and DDLPS patients treated with nivolumab+ipilimumab as the only group to achieve a 6-month response rate[144]. This illustrates similar histology specific results to those seen in pembrolizumab, further supporting the presence of an immune hot STS population. Accordingly, nivolumab+ipilimumab is currently being evaluated in a UPS and DDLPS specific cohort as part of a phase II trial[145]. Nivolumab has also be investigated in combination with the TKI sunitinib as part of the ImmunoSarc trial[146]. ImmunoSarc was a basket trial encompassing 11 different histological subtypes of sarcoma. Impressively, of 14 evaluable patients, PFS at 6 months in the was 50%, with partial responses seen in CCS, ASPS, SS, AS, and chondrosarcoma patients. This again supports ICB-TKI combination strategies for use in STS.

1.2.3.5 Other therapeutic avenues

In addition to conventional chemotherapy, TKIs, and immunotherapy, anti-tumour effects can also be elicited through targeting epigenetic pathways. This is commonly achieved through inhibition of proteins involved in the methylation and acetylation of DNA and histones. Under normal physiological conditions, methylation and acetylation modify the structural organisation of DNA to regulate gene expression. Inhibitors of these processes can therefore alter the oncogenic gene expression profiles of tumours. One example is tazemetostat, a methyltransferase inhibitor targeting enhancer of zest homolog 2 (EZH2). Tazemetostat has been approved for use in advanced EPS based on a phase II basket study, which showed disease control in 26% of patients (n = 62)[147]. Further to methylation-based epigenetic inhibitors, histone deacetylase inhibitors (HDACi) have also been assessed in STS. As monotherapies, HDACi such as panobinostat have shown limited benefit in phase II clinical trials. Yet, early data does suggest value as a combination therapy with epirubicin[148]. Akin to HDACi, Poly (ADP-ribose) polymerase inhibitors (PARPi) such as olaparib are also under investigation for use in combination with other drugs in LMS and osteosarcoma (**section 1.5.3.2**)[149,150]. Finally, one of the most recent STS drug approvals is nab-sirolimus, a nanoparticle bound mammalian target of rapamycin (mTOR) inhibitor approved for malignant perivascular epithelioid cell tumour (PEComa). mTOR regulates a range in cellular functions including growth, metabolism, and survival[151]. Approval was based on a phase II trial analysing 21 patients, where overall stable disease was seen in 52%, and only 10% showed progressive disease[152]

1.2.3.6 Summary

Irrespective of the drug of choice for advanced disease, response rates are low across the general STS population. Moreover, when promising PFS or response rates are observed in the RCT setting these are rarely translated to OS benefits for patients. Whilst accounting for histological subtype in treatment decisions can improve outcome in some cases, varied responses are consistently reported. Heterogeneous responses suggest the need for patient stratification independent of histology and highlights potential for specific therapies in currently undefined STS populations. Indeed, it has long been appreciated that biological heterogeneity within STS: 1) exists at both the inter- and intra-subtype level; and 2) contributes to clinical disease course. However, excluding imatinib in GIST, routine clinical practice in the UK for STS is directed in a largely “one size fits all” manner, which fails to consider heterogeneity. This represents a huge gap in STS care that needs to be addressed. Across oncology practice, there has been a shift

towards precision medicine, and it is vital that patients with rare cancers such as STS can benefit from such advancements.

1.3 Molecular profiling in STS

Molecular profiling aims to analyse the components of a biological sample, be it on the cellular, tissue, or organismal scale. It can span from profiling a single molecule to attempts at capturing the total composition of a sample. There are numerous molecular profiling modalities within biological research. The most commonly employed study the genome, transcriptome and proteome, which together encompass all components of the central dogma of molecular biology; DNA, RNA, and proteins.

Over the last few decades, comprehensive molecular profiling has become a regularly employed tool across oncology research. This was arguably instigated by the Human Genome Project. The Human Genome Project sought to determine the entire human genome sequence, and since its completion in 2001 has marked a new 'post-genomic era' of molecular biology[153,154]. During this time, rapid developments in profiling technologies have driven a surge in the number of molecular studies conducted. In particular, the advent of next generation sequencing (NGS) has supported growth in genomics and transcriptomics by drastically reducing analysis costs and time[155]. In oncology, this expansion has triggered the establishment of consortiums such as The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), and Clinical Proteomic Tumour Analysis Consortium (CPTAC), which aim to comprehensively profile malignancies using -omic approaches[156–158]. Not only has post-genomic expansion transformed the research landscape, but it has also shifted clinical practice in oncology. In recent years clinics have adopted molecular profiling approaches throughout patient care. This includes diagnostics, and disease monitoring and management. For example, UK patients with advanced non-small cell lung cancer receive genetic testing for epidermal growth factor receptor (*EGFR*) mutations, the results of which direct treatment pathways[159]. Similarly, select UK breast cancer patients receive gene expression profiling tests which provide risk assessments for disease progression and guide treatment decisions[160–163]. In STS, molecular profiling technologies have primarily driven improvements in diagnostic accuracy for certain histologic subtypes (**section 1.5.1**). Targeted profiling can be requested for STS patients to screen for established genomic alterations such as *NTRK* fusions or *KIT* mutations and highlight treatment options (**section 1.2.3.3**), yet beyond this there is little integration between molecular profiling and the management of most adult STS patients. To

facilitate further integration, whole genome sequencing (WGS) has been commissioned by National Health Service (NHS) England for sarcoma patients meeting certain criteria[164]. Whilst this is certainly beneficial for the STS community, at present there are no formalised guidelines addressing the translation of WGS results for prognostic or predictive purposes. Therefore, clinician discretion is often relied upon, and the integration of molecular profiling into routine clinical management is inconsistent. There can be little doubt that molecular profiling in STS holds untapped potential to transform patient care, and in line with this, increasing efforts are being made to profile this disease.

1.3.1 Dissecting STS biology and heterogeneity

STS is a heterogeneous group of malignancies. Yet, all STS tumours share a common mesenchymal origin, and therefore pan-subtype profiling studies have been performed to delineate unifying mesenchymal molecular features. By profiling multiple histological subtypes at once, these studies permit comparative assessments across the STS disease space. These have revealed both similar and contrasting features between and within diagnoses. To generalise, in multi-subtype profiling, subtypes with simple genomes tend to show individually distinctive molecular profiles. Complex genome STS, particularly those with undifferentiated phenotypes, show molecularly heterogeneous profiles, which can be challenging to distinguish for each other. Reflective of this, transcriptomics clustering analyses aimed at pan-STS subtyping have repeatedly provided evidence for both tight subtype-specific clusters and more diffuse mixed subtype clusters[36,41,165]. Namely, specific transcriptomic clusters identified include those enriched in SS, GIST, and LMS. Whilst mixed clusters show heterogenous populations of UPS, DDLPS, MyFS, and other tumour types. It is hypothesised that these differences translate to variations in treatment response and outcome, between and within subtypes. Such histology-agnostic profiling therefore lends itself to pan-STS subtyping, which may aid the identification of high risk or treatment responsive STS patients. At present, efforts to comprehensively profile STS have primarily utilised genomic and transcriptomic methods, including WGS, whole exome sequencing (WES), and RNA sequencing (RNAseq). Beyond these, methylation profiling has been conducted to aid STS diagnosis, and less comprehensive methods such as IHC and reverse phase protein arrays (RPPA) have been employed at the protein level.

1.3.1.1 The molecular basis of STS

Gene fusions in STS

Approximately half of all STS fall into a genomically simple classification, many of which are driven by fusion events (**section 1.1.2.2)**. Common fusions across fusion-driven STS

include those involving *EWSR1;* such as *EWSR1-ERG* and *EWSR1-FLI1* in Ewing sarcoma, *EWSR1-ATF1* in CCS, and *EWSR1-WT1* in DSRCT[4,166]. *EWSR1* encodes a TET family RNA binding protein whose normal physiological function is unknown[167]. Evidence suggests EWSR1 to play roles in transcription, DNA repair, cell division and cell ageing[167]. Consequently, the *EWSR1* fusions may drive tumorigenesis through may different mechanisms. Other fusion-driven STS include the *BCOR*-rearranged sarcomas, which as suggested by name are characterised by *BCOR-CCNB3* fusions, the *CIC*-rearranged sarcomas characterised by *CIC-DUX4* fusions, and the *NTRK*-rearranged sarcomas characterised by *NTRK* fusions[4]. One of the most prevalent fusion-driven STS tumours is SS. SS is driven by the reciprocal chromosomal translocation t(X;18)(p11;q11), which results in *SS18-SSX1/2/4* fusion. Approximately 95% of SS patients show a detectable *SS18-SSX* fusion, with 2/3^{rds} presenting with the *SSX1* variant and 1/3rd with the *SSX2*[28,29]. *SS18-SSX4* is exceedingly rare and occurs at a much lower rate than *SS18-SSX1/2*[30]. Under normal physiological conditions, SS18 is a component of the switch/sucrose non-fermentable (SWI/SNF) chromatin-remodelling complex. By facilitating chromatin remodelling, the SWI/SNF complex tightly regulates DNA accessibility and therefore transcriptional activity. In SS, the SS18-SSX fusion protein is incorporated into the SWI/SNF complex, triggering removal and proteasomal degradation of the SMARCB1 subunit. The exact cellular consequence of this altered complex is unclear, however given the genome-wide regulatory role of SWI/SNF complexes it is likely that oncogenic changes are wide ranging. The altered SWI/SNF complex of SS is suggested to bind at *loci* repressed by polycomb repressive complexes (PRC), activating the transcription of normally silenced genes. Indeed, the *Sox2* loci is PRC repressed, and high expression of *Sox2* is observed in SS tumours[168–170]. Conversely, in SMARCB1-deficient sarcoma cell lines, genome-wide SWI/SNF occupancy has been observed, resulting in enhancer activation in opposition to PRC repression[171].[172,173174,175]

Gene fusions are also detectable in complex genome STS; however, these are predominately introduced by intrachromosomal rearrangements introduced by amplifications[176]. These by-product gene fusions are non-recurrent events, which do not map to driver events and are not hypothesised to have pathogenic bearing. By contrast, recurrent gene fusions are more likely to confer a disease advantage. In complex genome STS, the only pan-subtype, recurrently identified fusions involve trio rho guanine nucleotide exchange factor (GEF; *TRIO*)[36,177]. In total, 4 distinct *TRIO* gene fusions have been identified in STS: fusions with telomerase reverse transcriptase (*TERT*) in UPS, DDLPS, and pleomorphic rhabdomyosarcoma; with both cadherin 18 (*CDH18*) and

TERT in UPS; with long intergenic non-protein coding RNA 1504 (*LINC01504*) in UPS; and with zinc finger protein 558 (*ZNF558*) in MyFS[177]. *TRIO* fusions are suggested to be unique to complex genome STS and have so far not been detected in tumours classed as genomically simple, although targeted screening is required for confirmation. Comparative RNAseq profiling of *TRIO*-fusion STS and non-*TRIO*-fusion STS revealed distinct transcriptomes between the two. Specifically, *TRIO*-fusion tumours were enriched in immunity and inflammation related genes, despite no histologically observable difference in immune infiltration. The mechanistic basis for increased immune expression in *TRIO*-fusion tumours is unclear. All *TRIO* fusions characterised in STS result in a truncated TRIO protein that retains its GEF1 domain. GEF1 mediates activation of rac family small GTPase 1 (Rac1) and ras homolog family member G (RhoG), proteins which are implicated in major signalling pathways for cell proliferation and motility [178]. *TERT*, the only *TRIO* fusion partner identified pan-subtype, is critical to the telomerase-mediated mechanism of telomere maintenance; 1 of 2 mutually exclusive pathways, the other mechanism being alternative lengthening of telomeres (ALT)[179,180]. Both telomerase-mediated maintenance and ALT promote telomere stability, which enables cancer cells to avoid senescence, immortalise, and replicate pathologically. In the *TRIO*-*TERT* fusion samples, *TERT* expression is high, and markers for ALT were found to be consistently negative[177]. This suggests telomerase-mediated telomere maintenance is active in *TRIO*-*TERT* fusion STS and reveals a candidate axis for therapeutic intervention. Overall, *TRIO*-fusion STS account for a minority (estimated ~ 5%) of overall STS, however they may represent a key subtype of complex genome tumours. The limited number of *TRIO* fusion cases identified has limited investigations into clinical course, however it is not unreasonable to hypothesise that a molecularly distinct STS subtype will show different patterns of outcome and treatment response, therefore highlighting potential clinical relevance.

Mutational profile of STS

Genomically simple STS characterised by mutational events include RT and EPS. RT and EPS show by loss of function mutations (eg. point mutations or deletions) in *SMARCB1* and by extension present with altered SWI/SNF activity[181–184]. Approximately 95% of RT and 90% of EPS harbour *SMARCB1* mutations[185–187]. As in SS, the exact mechanistic consequence of an altered SWI/SNF not defined for these tumours. It is hypothesised in EPS that the EZH2 axis may play a role in oncogenic activity, as a subset of EPS patients show favourable clinical responses to the EZH2 inhibitor tazemetostat (**section 1.2.3.5**)[147,188].

Beyond specific genomically simple STS, attempts have been made to identify common STS-wide mutational features that may underlie mesenchymal oncogenesis. At present, the only robustly identified feature is the low tumour mutation burden (TMB) rate of STS relative to other cancers. Although there is no conclusively established value, a TMB of > 20 mut/Mb is often considered 'high'[189]. Median TMB for STS is ~ 1 mut/Mb, whilst for other cancers TMB varies from 0.34 - 45.2 mut/Mb [36,189,190]. There are exceptions to the low TMB phenotype. In a group of cardiac STS, 93% were shown to have high TMB, and in more common subtypes such as UPS and LMS, studies often report a minority of samples with high TMB[191–193]. Yet, when considering STS as whole, TMB is low, and accordingly only a handful of recurrently mutated genes are reported. These include the tumour suppressors *TP53*, retinoblastoma 1 (*RB1)*, neurofibromin 1 (*NF1)*, and ATRX chromatin remodeler *(ATRX*).

TP53 is a transcription factor which induces the expression of numerous genes, including those involved in cell cycle arrest and apoptosis[194,195]. Resultantly, aberrant TP53 can drive excessive cell cycle activity and the evasion of cell death. It is therefore unsurprising that *TP53* mutations are seen in approximately half of all cancer types[196–198]. In STS, *TP53* mutations are present in between ~ 5% and 50% of tumours; harboured most frequently in LMS and AS (~ 50%) and to a lesser extent in UPS (30 - 59%), LPS (7 - 20%), and SS (~ 5%)[36,43,175,199–202]. Akin to TP53, RB1 also exerts tumour suppressive effects through cell cycle regulation, and its loss is observed across cancer types[203]. For a subset of soft tissue tumours, the '*RB1*-deleted tumours', *RB1* deletion is the reported putative driver event in nearly all cases[204]. However, excluding pleomorphic LPS, all tumours of this category are benign. *RB1* mutations/deletions are observed across other STS, although such loss of *RB1* is neither a characteristic nor driver event. These 'non *RB1*-deleted tumours' which show *RB1* loss include LMS, UPS/MyFS, and DDLPS, where *RB1* is mutated in approximately 14 - 27%, 16 - 43%, and 19% of tumours respectively, and disrupted in up to 94%, 88%, and 60% respectively[36,43,205–208]. *NF1* encodes a GTPase which negatively regulates Ras signal transduction, by converting active Ras into its inactive form[209]. Ras is involved in both phosphatidylinositol-3-kinase (PI3K) and mitogen-activated protein kinase (MAPK) signalling, pathways which are central to numerous cellular activities, such as differentiation, cell cycle, and apoptosis. In STS, *NF1* mutations/deletions are classically associated with the development of malignant peripheral nerve sheath tumour (MPNST) or UPS. However, recent studies have identified *NF1* loss in multiple other STS subtypes. Specifically, approximately 10.5% of MyFS, 8% of pleomorphic LPS, and 20% of all LPS present with *NF1* loss [199,200]. The final frequently altered gene in STS is *ATRX*. *ATRX* encodes a SWI/SNF family

protein, which plays roles in homologous recombination, PRC2 silencing of genes, and telomeric stability by ALT[210–212]. Defective ATRX is seen across subtypes including in LMS (33%), DDLPS (25 - 30%), UPS (34%), and AS (18%). Investigations into the downstream consequences reveal *ATRX* loss to positively correlate with ALT; with *ATRX* loss observed in 55 - 93% of STS with positive ALT markers[36,213].

Although recurrently altered, *TP53*, *RB1*, *NF1*, and *ATRX* show varied mutation rates across STS populations, often reflective of histology. Whilst most mutations occur sporadically and can represent a driver event in oncogenesis, germline altered *TP53* (Li-Fraumeni syndrome), *RB1* (Retinoblastoma), and *NF1* (Neurofibromatosis), can also result in the development of STS[15]. In addition to mutational alterations, the tumour suppressive roles of TP53, RB1, NF1, and ATRX can be ablated by other mechanisms, such as deletion events, epigenetic regulation, or alterations in genes encoding up/downstream proteins. There is considerable overlap between the phenotypic effects of *TP53*, *RB1*, *NF1* and *ATRX* mutations, all of which in their non-altered form function to maintain central cellular homeostatic activities. Therefore, loss of function in any of these proteins confers considerable tumourigenic effects.

Genome-wide alterations in STS

At the macro level, complex genome STS show high chromosomal instability[36]. This is both numerical chromosomal instability (CIN), displaying aneuploidy and loss of heterozygosity (LOH), and structural, consequent of genome rearrangements. This instability is resultant of macroevolutionary events including whole genome duplication (WGD) and chromothripsis, whereby thousands of clustered genomic rearrangements simultaneously occur. These large-scale events manifest as CNAs. STS show a notably more variant CNA profile than other cancer types that is reflective of histology[36]. DDLPS, UPS, MPNST and MyFS show a high frequency of often genome-wide CNAs, LMS show an intermediate level of CNAs, whilst CNA are rarely observed in genomically simple STS such as SS. Despite histological heterogeneity in STS, genomic alternations are typically centred on the same pathways across tumours, and strikingly, in the same pathways for which recurrent mutations have been identified: the murine double minute 2 (MDM2)-p53 axis and CDKN2A(p16)- cyclin dependent kinase 4 (CDK4)-RB1 axis. For example, *MDM2* amplification is seen in 91 - 100% of WD/DDLPS, and functionally mimics a *TP53* mutation[36,214–217]. Similarly, 91 - 100% of DDLPS patients possess *CDK4* amplifications, whilst deep deletions in *CDKN2A* are identified in 8% of LMS, 20% of

UPS, 18% of MyFS, and 2% of DDLPS, and deep deletions in *RB1* observed in 14% of LMS, 16% of UPS and 24% of MyFS. This illustrates a common disruption of key pathways across STS, which is mediated by multiple mechanisms.

1.3.1.2 Immune profiling of STS

Given the recent advances in immunotherapy and the critical role the immune microenvironment plays in tumour progression, profiling the immune component of STS tumours is increasingly important. Historically, STS have been considered immune-quiescent, particularly when compared to other malignancies. This is closely tied to the observed low TMB in STS (**section 1.3.1.1**), which has previously been identified as reflective of immune activity. High TMB results in increased immunogenic neoantigen expression on the surface of tumours cells. Neoantigen recognition within the tumour microenvironment (TME) can trigger CD8+ T cell activation, inducing T-cell response and tumour cell lysis. However, a low TMB is not necessarily indicative of an immune deserted tumour. In ovarian cancer, low TMB tumours show elevated memory B and plasma cells, illustrating that TMB may dictate the immune microenvironment composition rather than simply the presence or absence of all immune activity[218]. Furthermore, despite the overall low TMB and general low immune activity of STS, a subset of patients show favourable responses to immunotherapy (**section 1.2.3.4**), illustrating the presence of an active and targetable immune module within the STS TME. Accordingly, there is an ever-expanding body of evidence that suggests the simplified classification of STS as a non-immunogenic malignancy is not appropriate.

The immune TME is considered comprised of: 1) an infiltrating immune cell population, made up of tumour associated neutrophils (TANs), TILs (T cells, B cell, natural killer (NK) cells), tumour associated macrophages (TAMs), and dendritic cells (DC); 2) soluble immune factors, including chemokines and growth factors such as the interleukins (ILs) and VEGFR; and 3) immune molecules presented on tumour and tumour-supporting cells, for example the immune checkpoint proteins PD1, PD-L1, and CTLA4[219]. All play integrated and critical roles in determining tumour immunity and by extension tumour progression. For example, the balance between pro- and anti-tumorigenic T cell signalling activity is modulated by immune checkpoint molecules. In STS, IHC has been utilised to assess both T cell TILs and immune checkpoint molecule expression. T cells are characterised as total, helper, cytotoxic, and regulatory cell populations based on CD3, CD4, CD8, and FoxP3, expression respectively. One of the most recent IHC studies profiled 192 tumours, spanning 5 ‘common’ and 12 ‘rare’ STS subtypes, in tissue microarray (TMA) format[220]. This identified T cell (CD3+) infiltration in approximately 50%

of STS, with the infiltrate accounting for an average of 1.02% of the total cellular population. There were a significantly higher number of CD3+ cells in higher grade (grade 3) tumours compared to lower grade (grade 1/2) samples. All TIL measures (CD3, CD4, CD8, FoxP3 in different expression combinations) showed histology-based differences. The highest TIL levels were observed in MyFS and UPS compared to SS, LPS, and the rare subtypes; however statistical significance was not always reached. Histology based differences were also observed in all PD1+ cell types and most PD-L1+ cell types. The checkpoint molecule profiles were more comparable across histology than TIL scores, with significantly higher levels observed only in MyFS when compared to LMS. Survival analyses identified higher regulatory T cells as associated with a poorer LRFS, however multivariable adjustments only considered tumour margin despite other clinicopathological variables being implicated in LRR (**section 1.2.2.1**), and so interpretation is restricted. Overall, this study revealed histology based immune variation, but did not robustly identify any association between immune composition in STS and outcome. Similar reports have also failed to identify any significant relationship with PD-L1 expression, TIL level and outcome[221]. Although the literature is unclear, with others reporting high PD-L1 expression and high TIL as associated with a significantly improved OS and DSS[222]. Often, such inter-study variation is attributed to differences in the study cohort composition, highlighting one of the difficulties of working with a heterogenous and rare disease. In addition to the demonstrated importance of lymphocyte populations in STS, the myeloid component also plays a central role. Indeed, TAMs have been reported to outnumber TILs across most STS types, and within the macrophage population, M2 macrophages (immunosuppressive) outnumber M1 macrophages[223,224]. Furthermore, TCGA reported M2 macrophages to be correlated with DNA damage measures, illustrating potentially important interplay between the tumour and macrophages in the immune TME[36].

Within STS, IHC profiling was, and still is, fundamental in elucidating the immune TME. IHC is an accessible method that provides spatial resolution at the protein (ie. the effector) level. However, antibody-reliant methods can have limited reproducibility and poor specificity. This has led to the development and use of immune deconvolution strategies, which estimate immune infiltrate based on bulk transcriptomic data[225–230]. Deconvolution infers cell type and abundance by referencing the profiles of purified cell types. This provides a cellular signature score as a proxy for cellular infiltrate. In STS, limited spatial profiling studies have been performed, and thus most large-scale profiling data lack the dimensional resolution that IHC can provide. However, gene expression data offers a significant reproducibility advantage over IHC measures, meaning multi-

experiment datasets are more comparable. In addition, the increased sampling depth of gene expression profiling derives a more comprehensive profile, with thousands of transcripts profiled in a single experiment. An immune response is multi-faceted and engages multiple cell types in a coordinated manner to define function; this level of complexity can only be uncovered by -omic scale profiling. TCGA utilised deconvolution methods to report immune estimations in STS and revealed that overall immune cells/functions show common patterns[36]. Tumours high in one immune signature were, for the most part, high in other immune signatures, indicative of a complete immune response. However, supervised histology-focused investigations did identify minor variations. Specifically, high macrophage scores were observed in UPS/MyFS and DDLPS, the highest CD8 levels were estimated specific to DDLPS, and the highest PD-L1 levels were seen in stLMS tumours. These observations were not statistically interrogated. Further to histological subtype specific infiltrations and expression, scores derived from immune deconvolution also showed a subtype specific association with DSS. In brief, a high T helper 2 (T_h2) signature was prognostic for poor DSS in DDLPS; high DC scores were prognostic for improved DSS in UPS/MyFS; high NK cell scores were prognostic for improved DSS in LMS, and UPS/MyFS, and high CD8+ and mast cells were prognostic for improved DSS in uLMS and LMS of other soft tissue sites (stLMS) respectively. However, this analysis was univariable and so did not involve correction for other clinicopathological variables. Interpretation is further limited as survival statistics were calculated based on the top and bottom 1/3rd scoring samples only. Any potential association between outcome in intermediate scoring samples in unknown, and information is lost as a result of categorising this continuous variable.

Beyond descriptive analysis, immune deconvolution has also been utilised to identify histology independent STS molecular subtypes based on immune composition; named the sarcoma immune classes (SIC)[231]. SIC were characterised based on the gene expression profiles of 608 LMS, DDLPS, and UPS samples, and have been applied to SS, myxoid LPS and GIST tumours. This led to the identification of 5 SIC subtypes (A, B, C, D, E) showing variation in immune activity. From A to E, SIC increase in immune activity as measured by T cell activation, chemotaxis, and survival genes, major histocompatibility complex (MHC) class I genes, immune regulatory genes, and the presence of tertiary lymphoid structures (TLS). SIC A represents an immune desert population, SIC B a heterogenous immune-low group, SIC C a highly vascularised group, SIC D a heterogenous immune-high group, and SIC E an immune high group with TLS present in 82% of tumours assessed. TLS are lymphoid organs which develop in areas of chronic inflammation such as at tumour sites[232]. TLS facilitate anti-tumour immunity

through the local generation of autoreactive T/B cells and autoantibodies. It is striking that all except 1 TLS containing tumour was classified as SIC E; reflecting a sustained immune phenotype as characteristic to SIC E[231]. In relation to genomic profile, no difference in CNA level was seen across the SICs, yet mutation frequency in *TP53*, *TTN* and *MUC16* was significantly higher in SIC D and E compared to the other SICs. However, the mutation rates did not increase linearly from A to E. In fact, the mutational occurrence of *MUC16* was highest in SIC D, but lowest in SIC E. The relationship between these mutations and the overall level of immune activity is therefore unclear. Across all STS subtypes investigated, all SICs were represented in each histological subtype, illustrating histological independence. Yet patterns in histological distribution were observed. SIC A and B encompassed most of the LMS samples, C was approximately 50% DDLPS, and D and E showed a more equal representation of all subtypes. This alludes to a pan-subtype 'immune hot' population, which may benefit from immunotherapy interventions. As well as a potential use in predicting treatment response, SICs were also suggested to provide prognostic utility. The highest immune SICs (D and E) were shown to have a significantly longer OS compared to the lowest (A) in multivariable analyses. This supports the TCGA results and supports previous data highlighting TLS as associated with improved outcome in other cancer types[36]. Further analysis revealed the SIC survival difference to be specifically driven by a high B cell signature in high immune SICs[231]. It is hypothesised that this B cell signature is resultant of B cell germinal cores in mature TLS. Whilst B cells were not discussed by TCGA, T_h cells which promote B cell proliferation and differentiation were. Contrary to the B cell-outcome relationship in SIC, the TCGA reported T_h signature correlated with worse outcome[36]. This may suggest the presence a more nuanced relationship between TLS composition, tumour immunity, and disease progression. Indeed, in colorectal cancer, T_h-rich TLS are associated with an increased likelihood of recurrence[233].

Complementary to the STS specific molecular immune subtypes, the representation of pan-cancer defined molecular immune subtypes has also been studied in STS[136]. In 2018, all publicly released TCGA data spanning 33 cancer types, was analysed to characterise pan-cancer immune features. This detailed 6 subtypes: C1 'wound healing', C2 'IFNγ dominant', C3 'inflammatory', C4 'lymphocyte depleted', C5 'immunologically quiet', and C6 'transforming growth factor β (TGFβ) dominant'. Interestingly, none of the 257 STS tumours assessed were classified as immunologically quiet, however STS were over-represented within the lymphocyte depleted group. In addition, STS were shown to display a moderate leukocyte fraction and a larger range of leukocyte fraction relative to other cancer types. This is reflective of the extreme diversity of STS and, in agreement

with STS-specific studies, suggests the presence of an immune hot population that warrants further exploration.

1.4 Clinicopathological and molecular features of select STS subtypes

1.4.1 Leiomyosarcoma

LMS is one of the most common adult STS subtypes, accounting for between 10-25% of STS diagnoses[234,235]. LMS arise from the smooth muscle lineage, most often developing with no identifiable causative factor within the extremities, retroperitoneum, and uterus. LMS of uterine origin (uLMS) occurs in a younger population than stLMS, with peak incidences in the 5th and 7th decades respectively[236]. Histopathological diagnosis of LMS is reliant on 1) morphologically identified fascicles of elongated and spindled cells, and 2) immunohistochemistry (IHC) detection of smooth muscle markers: alpha smooth muscle actin, desmin, and H-caldesmon[237–239]. These IHC proteins however are not disease specific or ubiquitously expressed across LMS. Between 5% and 34% of morphologically-LMS tumours do not stain positive for 1 or more of these routine markers[240]. Resultantly, misdiagnosis in LMS is a persistent risk. Molecular profiling studies often note reclassification of LMS diagnoses upon central pathology review (**section 1.5.1**), and several case reports describe LMS patients misdiagnosed with benign leiomyomas[241,242]. Furthermore, LMS lacking smooth muscle marker expression show an undifferentiated phenotype which can be challenging to discriminate from a UPS tumour[243,244]. Misdiagnosis carries huge risk as LMS is regarded as one of the most aggressive STS subtypes. Tumours show a particularly high propensity for metastasis and recurrence; even when patients present with primary localised disease and receive optimal surgical intervention. The 5-year recurrence rate varies from 10% to 43% dependent on anatomical site, and long-term patient follow up shows late recurrences (> 10 years) can occur in extremity, abdominal and retroperitoneal patients[245]. In addition, only a subset of LMS patients respond to conventional chemotherapy and radiotherapy. Whilst distinct patterns in treatment response and clinical outcome are reported between uLMS and stLMS, the full spectrum of heterogeneity across patients exceeds that introduced by anatomy alone. Therefore, anatomical site is not a consistent factor in predicting disease course. In fact, in multivariable analyses, where multiple clinicopathological variables are adjusted for, only grade and size have been reported as significant prognostic factors for DSS following primary LMS[245]. Understanding of the molecular basis of LMS has been expanded greatly over the last decade and is hypothesised to explain some of these clinically

observed characteristics. As a relatively common disease type, LMS are frequently profiled as part of multi-subtype profiling experiments, and importantly have also been assessed in isolation, in attempts to appreciate intra-subtype heterogeneity.

1.4.1.1 Key molecular features of LMS

As with other STS subtypes, LMS show dysfunctional RB1 and TP53. However more frequent to LMS, these aberrations in *RB1* and *TP53* are often concomitant[246]. These seemingly coupled events are hypothesised to be early or driver occurrences in LMS tumorigenesis. Screening across STS subtypes revealed LMS as the adult tumour type with highest frequency in both Retinoblastoma and Li-Fraumeni populations[247–249]. In accounting for a non-trivial proportion of tumours in patients with hereditary *RB1* or *TP53* loss, this illustrates the central role that RB1 and TP53 can play in LMS progression.

In addition to the recurrent STS-wide genomic aberrations, LMS specifically show a particularly altered level of activity in the phosphatase and tensin homolog (PTEN)/PI3K-AKT Serine/Threonine kinase 1 (AKT) pathway[36,43,246,250–254]. This is not unique to LMS, and is observed in other STS subtypes, but at a comparably low frequency. PTEN is an established tumour suppressor with critical catalytic functions that modulate AKT signalling[255–257]. Briefly, PTEN is the antagonist of PI3K and sits downstream of growth factor receptor tyrosine kinases (**Figure 1.4A**). When activated, PTEN dephosphorylates phosphatidylinositol (3,4,5)-trisphosphate (PIP3) converting it to phosphatidylinositol (4,5)-bisphosphate (PIP2). PIP3 regulates the activation of phosphoinositide-dependent kinase 1 (PDK1), PDK1 activates AKT, and AKT inhibits tuberous sclerosis 1/2 protein (TSC1/2), relieving TSC-mediated inhibition of mTOR complex 1 (mTORC1). When activated, mTORC1 triggers multiple cascading signalling pathways to promote cell growth and tumourigenic survival. The other mTORC, mTORC2, sits upstream of AKT to activate signalling. Cross talk within the pathway is complex, and multiple feedback loops are present, such as mTORC1-mediated inhibition of mTORC2, and TSC-mediated activation of mTORC2, which attenuate and promote AKT signalling respectively. In contrast to *RB1* and *TP53*, germline altered *PTEN* (e.g., Cowden syndrome) has not been reported to translate to a significant predisposition to LMS[258]. In LMS, the PTEN/PI3K-AKT pathway is altered in an estimated 71% of patients, and loss or inactivation of *PTEN* specifically has been observed in 28-57% of LMS[36,43,250,251]. Mutations in *PTEN* itself are an infrequent event seen in approximately 5% of LMS[36]. However, chromosomal deletion of 10q (the region encompassing *PTEN*) occurs at a far higher rate (59% of LMS), and *PTEN* specific deletions are seen in 21-64% of LMS[36,253,254]. These deletions are mostly (~ 85%) predicted to be shallow (ie.

heterogyzous), with homozygous *PTEN* deletions rarer[36]. This is in agreement with *PTEN* alterations in other malignancies, which present mostly with *PTEN* loss of heterozygosity[259–261]. Following the traditional 'two-hit tumour suppressor hypothesis', biallelic inactivation of *PTEN* would be required to promote tumorigenesis[262]. However, studies in other cancers have illustrated haploinsufficiency to confer significant tumorigenic activity[259–261]. The mechanism in LMS is unclear, however irrespective of deletion type, *PTEN* deleted tumours show downstream changes in AKT signalling. TCGA found concordant high AKT pathway scores in both gene expression and RPPA analyses in LMS with *PTEN* loss[36]. Moreover, independent studies have identified overexpression of phosphorylated AKT (ie. activated AKT) in 20-75% of LMS tumours overall[263,264]. This expected relationship between *PTEN* loss of function and increased AKT signalling is more pronounced in well differentiated LMS, where significant overexpression of activated AKT and RICTOR (an mTORC2 component) is seen when compared to other LMS tumours also showing aberrant PTEN[252]. It has been hypothesised that the role RICTOR plays in smooth muscle differentiation may explain this observation, highlighting intra-subtype variations based on the smooth muscle differentiation phenotype. Further intra-LMS variations have also been observed, with one study finding *PTEN* inactivation to be significantly higher in non-primary stLMS compared to non-primary uLMS[246]. This anatomical difference was not observed in primary LMS lesions, and there was no overall significant difference between all primary LMS and all non-primary LMS. This is suggestive that *PTEN* inactivation may promote disease recurrence and/or be acquired during the progression of stLMS. Clinically, the identification of altered AKT signalling in LMS, indicates these patients may be susceptible to mTOR inhibition (eg. everolimus), or the more recently developed dual PI3K/MTOR inhibitors[265–268].

Interestingly, there is evidence which suggests cross talk between the PTEN/PI3K-AKT pathway and oestrogen receptor (ER) signalling, particularly in breast cancer[269,270]. Indeed, high ERα expression is seen in 15 - 60% of LMS, and it is particularly enriched in uLMS[246,271,272]. This contrasts other STS subtypes which show minimal, if any, expression of hormone receptors[273]. Expression of ER is not the only feature LMS tumours appear to have in common with breast cancers. Up to 98% of LMS have also been revealed to show a breast cancer gene (BRCA)ness signature[274]. BRCAness describes a signature characterised homologous recombination repair (HRR) defects that phenotypically mimics a *BRCA1/2* mutation[275]. BRCA1/2 are both key proteins in the HRR pathway, which is relied upon for the repair of double strand DNA breaks (DSB; **Figure 1.4B**). BRCA1 has broad roles in controlling HRR signalling and is integral in the

processing of DNA break ends. BRCA2 has a more defined downstream role, binding RAD51 to facilitate its recruitment to DNA damage sites. There is no singular method for determining BRCAness. Instead, it can be molecularly characterised based on individual mutations in HRR components, broad mutational signatures, or transcriptional signatures corresponding to HRR defects. In LMS, BRCAness has been reported based on the detection of both a BRCAness mutational signature ('Alexandrov-COSMIC 3: associated with defective HRR'), and loss of function mutations or deletions in HRR genes including *BRCA1* (10%), *BRCA2* (53%), *PTEN* (57%), *ATM* (22%), and *RAD51* (10%). One study has suggested *BRCA2* mutations to be more prevalent in uLMS than stLMS (10% vs 1%), and overall alteration rates in HRR pathways to be similar[276]. There is little data on the prognostic repercussion of BRCAness in LMS, however patients with *BRCA1/2* loss show a trend towards a higher mitotic count and more dedifferentiated histology; both suggestive of more aggressive disease[276]. Defects in HRR in LMS have been reported as associated with a significantly shorter OS in the univariable setting[277].

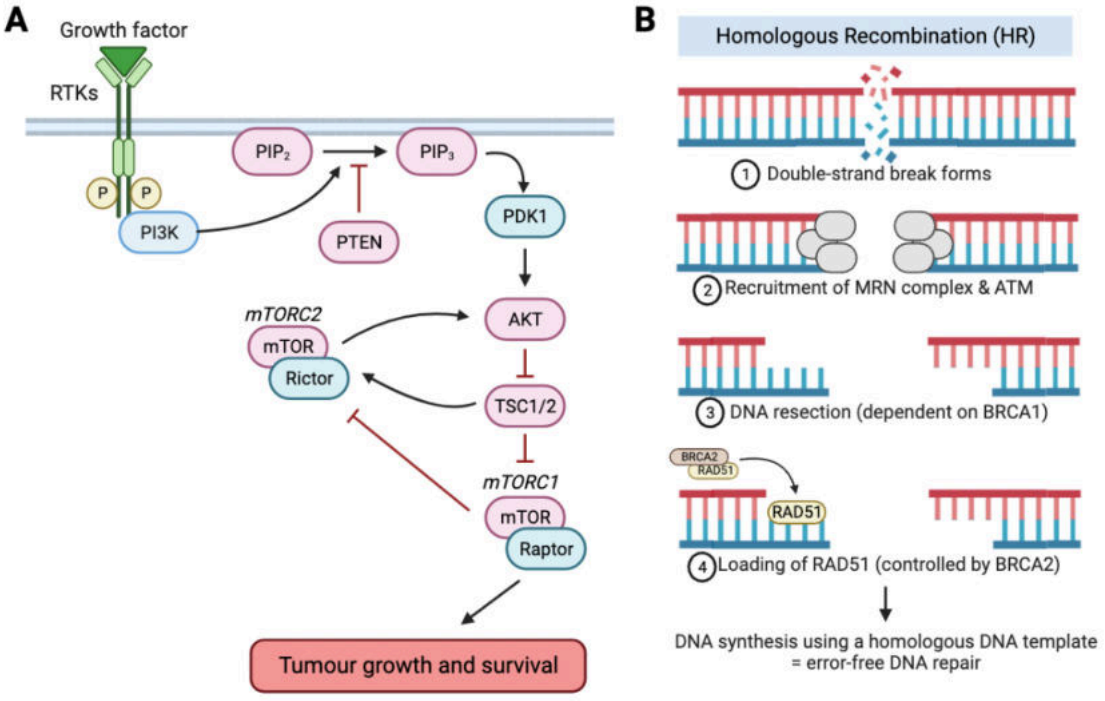


Figure 1.4 Diagrammatic representation of pathways altered in leiomyosarcoma
(A) The phosphoinositide 3-kinase (PI3K) signalling pathway **(B)** The homologous recombination (HR) pathway for the repair for double strand DNA breaks. Figures construction based on the publications of Carracedo *et al* and Lord *et al.*[275]

In multivariable analysis, patients with HRR defects not in *BRCA1/2* show significantly worse PFS than those with *BRCA1/2* HRR defects or an absence of HRR defects. These findings raise the concept of shared biology between breast cancer and LMS and highlights the possibility that currently approved breast cancer drugs may hold utility in LMS. Olaparib is one such drug currently approved in select breast cancer patients with a *BRCA* mutation, which has been investigated in tumours demonstrating BRCAness[278–280] (**section 1.5.3.2**).

1.4.1.2 Molecular subtypes of LMS

The extensive clinical heterogeneity observed across LMS patients has long supported the concept of LMS subtypes. In most pan-STS studies, LMS are identified as a relatively homogeneous group, with high similarity between patient tumours[36,41,165]. However, in 2009, molecular subtypes of LMS were first documented through focused LMS-specific microarray profiling in a cohort of 51 samples[281]. Since, numerous multi-institution studies utilising transcriptomics have repeatedly identified 3 molecular subtypes (**Figure 1.5** and **Table *1.2***)[36,43,274,282,283]. Methodologically, these studies mostly employed RNAseq, as well as microarray-based profiling, methylation profiling, and copy number analysis. Excluding the study by TCGA, all used an unsupervised approach to delineate LMS subtypes. Of note, the most recent study offered further stratification by use of pan-cancer data to identify 4 subtypes, 2 of which (Anderson subtype 2a and 2b) were derived from 1 parent subtype (Anderson subtype 2)[274]. Whilst the relationship between the subtypes identified in different studies has not been formally assessed, these studies are broadly considered to have identified highly similar if not identical LMS subtypes. Across all studies, common subtype-specific features are reported, such as variations in anatomical site, the expression of myogenic markers, immune activity, and potential associations with outcome.

Anatomical site distribution in LMS molecular subtypes

LMS subtypes are reported to show differential distributions in anatomical sites. This includes the repeated detection of a uLMS-enriched subtype (Beck group III, Guo subtype III, Chudasama SG1, Hemming uLMS, Anderson subtype 3)[43,274,281–283]. However, this suggested uterine enrichment is ambiguous. The level of uterine overrepresentation varies greatly, with uLMS accounting for between 34% and 92% of all samples within the putative uLMS-enriched subtype. Moreover, uLMS are present within the other subtypes, comprising between 19% and 59% of samples in the other non-uLMS enriched groups. In support of anatomically driven subtyping, one study

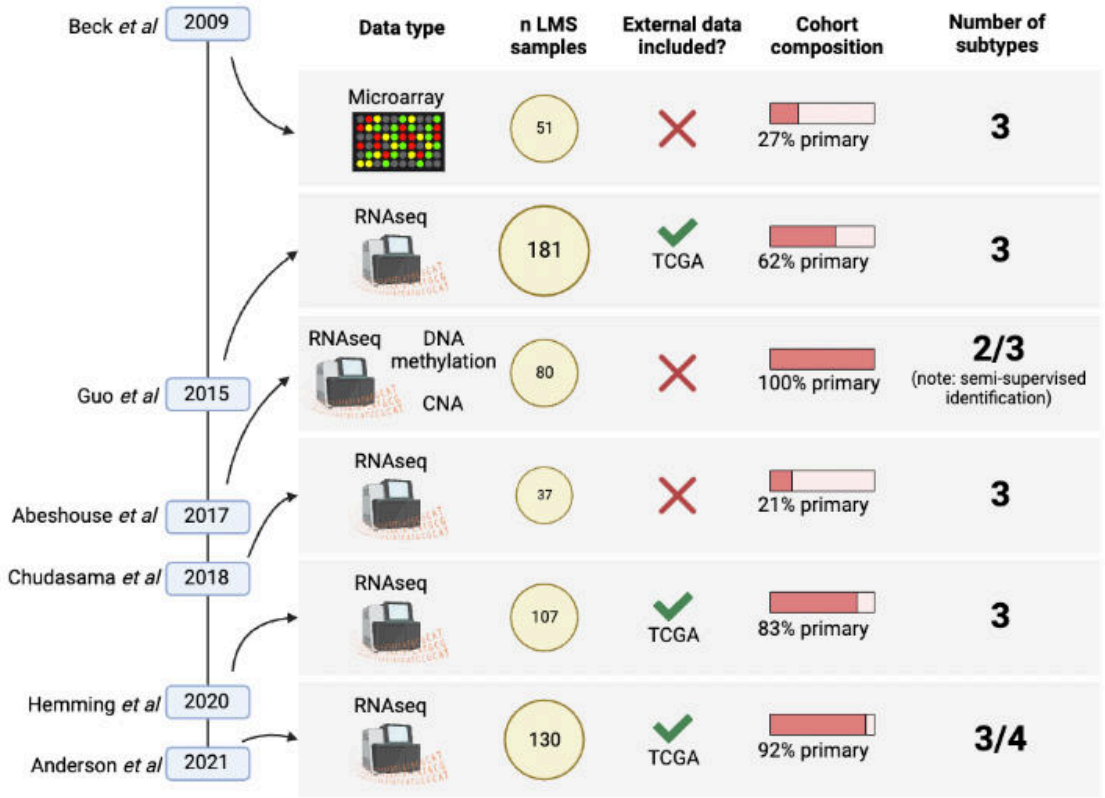


Figure 1.5 Timeline and overview of the leiomyosarcoma transcriptomic subtype literature
For each study, the type of data analysed is shown as well as the number of LMS samples, whether external data was included in the derivation of subtypes, the proportion of primary vs other tumour specimens, and the number of subtypes identified[36,43,274,281–283]. Abbreviations: CNA = copy number alterations

reported preserved expression of uterine-specific transcripts in the uLMS-enriched subtype, and absent or minimal expression of these transcripts in other subtypes[283]. One such transcript, Wilms' tumour gene (*WT1*), has also been reported elsewhere as overexpressed in uLMS, and has been revealed as associated with poorer OS and PFS across high grade uterine sarcomas[284–286]. However, whilst under normal physiological conditions it is rare for *WT1* to be highly expressed beyond the uterus, aberrant *WT1* expression has been noted across non-uterine cancer types[287–290]. The specificity of a relationship between uterogenic transcript expression and a uterine-enriched LMS subtype is therefore unclear. Furthermore, whether the stLMS samples present within the uLMS-enriched subtype, or the uLMS samples in the non-uLMS-enriched subtypes also express *WT1* is not reported[283]. In a separate study, unsupervised clustering illustrated most samples of the putative uLMS subtype to co-localise with normal gynaecological smooth muscle tissue[274]. Clustering in this study also sub-stratified a putative non-uLMS subtype (Anderson subtype 2) into 2 clusters, which appear driven

by anatomical site. Anderson subtype 2a comprised mostly abdominal lesions, and clustered with normal digestive smooth muscle. Whereas 2b comprised a mix of anatomical and extremity lesions, and clustered with normal vascular smooth muscle. Retroperitoneal and extremity LMS frequently arise in association with the vasculature, and therefore the stratification between 2a and 2b may illustrate distinct LMS tissue lineages. Overall, evidence suggests that anatomical site corresponds to differing molecular signatures across LMS. Yet there are notable exceptions, which weakens the hypothesised role of tissue lineage in driving LMS subtypes. This is therefore neither a robust nor definitive finding. Anatomical site may contribute to disease heterogeneity, but it does not fully explain the molecular differences observed across LMS.

The expression of myogenic markers in LMS molecular subtypes

LMS are derived of the smooth muscle lineage and diagnosis entails IHC assessment of smooth muscle markers. However different levels of myogenic expression have been observed across the LMS molecular subtypes. In most subtype studies, a 'high-myogenic' group is reported (Beck group I, Guo subtype I, Abeshouse stLMS C1, Chudasama SG2, Hemming cLMS, Anderson subtype 2)[36,43,274,281–283]. These groups are characterised by overexpression of numerous muscle specific genes and are suggested to be a subtype of low/intermediate grade, majority non-uterine tumours of mostly conventional histology. Genomically, the 'high-myogenic' groups have been characterised by hypermethylation, lower genomic stability compared to other LMS, and myocardin (*MYOCD*) amplifications. MYOCD is a transcriptional co-activator of smooth muscle gene expression, and therefore is implicated in smooth muscle differentiation[291]. Consequentially, *MYOCD* amplicons may representant one underlying mechanism for the high myogenic activity and well differentiated smooth muscle phenotype observed in this subtype. In contrast to the 'high-myogenic' groups, the remaining 'non-uterine' groups show lower expression of myogenic genes. In the Anderson 'low-myogenic' group (subtype 1), a high occurrence of deletions in the smooth muscle marker dystrophin (*DMD*) was observed[274]. Mechanistically, *DMD* deletion may explain the observed lack of a myogenic signature in this subtype. Together, these observations are suggestive of an LMS population with a dedifferentiated phenotype. Indeed, pan-STS clustering showed the majority of Anderson subtype 1 tumours to localise with non-LMS tumours, including the dedifferentiated sarcoma type, UPS. Similarly, a small-scale proteomic study using 2-dimensional difference gel electrophoresis (2D-DIGE) noted co-clustering of a subset of LMS with UPS samples[292]. Moreover, histologically observed dedifferentiation within LMS tumours has also been reported[293–296]. These reports describe tumours with regions of classical LMS tissue, co-occurring alongside

Table 1.2 Overview of leiomyosarcoma (LMS) molecular subtypes identified from transcriptomic studies
Abbreviations: stLMS = soft-tissue LMS (non-uterine); uLMS = uterine LMS; NK = natural killer; DSS = disease specific survival; RFS = recurrence free survival; OS = overall survival.

Subtype	Proportion (%)	Clinical features	Biological features	Survival analysis	Comments
Beck group I[281]	25%	92% stLMS, 77% conventional histology	Enriched in muscle related genes, phosphoproteins, and kinases. Lower genomic stability	Improved DSS in multivariable analysis	Survival analysis performed on seperate cohort using expression measure of unvalidated group I IHC markers
Beck group II[281]	24%	75% stLMS, 50% conventional histology	Enriched in metabolic, cell proliferation and organ development genes	-	
Beck group III[281]	51%	42% uLMS, 79% pleomorphic/mixed histology, mostly non-primary	Enriched in organ development, ribosomal, ECM and wound response genes	-	
Guo subtype I[282]	35%	72% stLMS, similar proportions of low, intermediate, and high grade tumours	Enriched in muscle related genes	Improved DSS in univariable analysis	Survival analysis performed on separate cohort classified based on unvalidated IHC markers
Guo subtype II[282]	22%	59% uLMS, 68% high grade	Enriched in translation & protein localization genes	Poorer DSS in univariable analysis	
Guo subtype III[282]	29%	92% uLMS, 77% high grade	Enriched in metabolic and transcription genes	-	
Guo ungrouped[282]	29%	-	-	-	
Abeshouse uLMS[36]	34%	100% uLMS	High DNA damage response, hypomethylation of ESR1 targets, altered AKT pathway	-	Supervised separation of uLMS from stLMS
Abeshouse stLMS C1[36]	31%	100% stLMS	High HIF1α signalling compared to uLMS, altered AKT pathway, generally hypermethylated, 40% MYOCD amplification,	Poorer RFS & DSS in univariable analysis compared to stLMS C2	
Abeshouse stLMS C2[36]	35%	100% stLMS	High HIF1α signalling compared to uLMS, generally hypomethylated, high inflammatory signatures (NK and mast cells)	-	
Chudasama SG1[43]	16%	34% uLMS	Enriched in platelet degranulation, complement activation and metabolic genes	-	
Chudasama SG2[43]	14%	19% uLMS	Enriched in muscle related genes	-	
Chudasama SG3[43]	70%	19% uLMS	Intermediate expression of muscle related genes, and cell-cell signalling genes	-	

continuation of table from previous page

Hemming cLMS[283]	49%	10% metastasis	High expression of muscle associated transcripts and IGF1R	Improved DSS compared to iLMS in univariable analysis	
Hemming iLMS[283]	28%	10% metastasis	Enriched in immune related genes. Estimated high M2 macrophage and CD8+ T cell infiltration	-	
Hemming uLMS[283]	23%	88% uLMS, 40% metastasis	Expression of uterogenic transcripts	-	
Anderson Subtype 1[274]	18%	43% gLMS	High occurrence of DMD deletions (evidence of dedifferentiation), high in immune activity (M2 macrophages)	-	Subtype 2 split in to 2a (31%; mostly abdominal) and 2b (69%; mixed abdominal and extremity)
Anderson Subtype 2[274]	65%	81% abdominal or extremity LMS	MYOCD amplifications	Improved OS & DSS compared to combined subtype 1&3 in univariable analysis	
Anderson Subtype 3[274]	17%	91% gLMS	DMD deletions & MYOCD amplifications	-	

de-differentiated non-myogenic components; reminiscent of mixed WDLPS and DDLPS tumours (**section 1.4.3**). Dedifferentiation is a well-studied phenomenon across oncology and often confers a higher grade more aggressive tumour type[297–299]. In line with this, dedifferentiated LMS tend to show a high mitotic index[293,296]. However due to the paucity of dedifferentiated LMS reports, the general aggressiveness of these tumours and overall clinical outcome for patients is not well defined. It is evident that the extent of differentiation in LMS tumours is molecularly rooted, which translates to the identification of molecular LMS subtypes. The presence of a dedifferentiated LMS tumour type is particularly pertinent, as in other cancer types, cellular dedifferentiation and a stemness phenotype identifies a high-risk patient population.

The immune component of LMS molecular subtypes

Dedifferentiation and increased cellular stemness has also been shown to be associated with immune cell exclusion, and therefore immune evasion across carcinomas[300,301]. Contrary to this, in LMS, the 'low-myogenic' subtypes have been shown to possess higher immune activity (Abeshouse stLMS C2, Chudasama SG1, Hemming iLMS, Anderson subtype 1)[36,43,274,283]. Across these subtypes, immune activity has been described through in-silico deconvolution estimation algorithms, which have reported high M2 macrophage, NK cell, CD8+ T cell, and mast cell infiltrations. In addition, over-representation analyses have noted enrichment in platelet degranulation and complement activation pathways. This illustrates a wide-ranging increase in immune activity within this subtype, concurrent with the TCGA-observed 'complete immune response' in STS (**section 1.3.1.2**). High immune activity is also known to be associated with altered DNA methylation. In line with this, TCGA reported global hypomethylation in the 'low-myogenic'/'high-immune' LMS subtype[36]. Yet, akin to the dedifferentiation-immune relationship in LMS, this appears contrary to the current carcinoma-centric literature. In carcinomas, global loss of methylation has been shown as correlated with immune escape mechanisms and low immune infiltrate[302]. The basis and consequence of an inverse immune-methylation and immune-dedifferentiation relationship as observed in this LMS subtype are therefore poorly understood.

The clinical significance of LMS molecular subtypes

Molecular subtyping of LMS has consistently identified a dedifferentiated subset of tumours. Across oncology, dedifferentiation is associated with a more aggressive tumour type. Accordingly, some studies report improved outcomes in the high-myogenic groups compared to the low-myogenic groups. Although these are neither consistent nor robust. Beck *et al.* and Guo *et al.* utilised TMA IHC to report survival analyses[281,282]. Both

suggested significantly improved DSS in the high-myogenic group. However, tumour classification into LMS subtypes was based on unvalidated IHC markers differentially expressed across subtypes, thus interpretation is highly tentative. More robust analyses were performed by Hemming *et al.* and Anderson *et al.*, who directly analysed the survival of the profiled and subtyped LMS cohort[274,283]. The former found improved DSS for cLMS (high-myogenic) compared to iLMS (low-myogenic), and the latter similarly found improved DSS for subtype 2 (high-myogenic) compared to the other subtypes combined. However, neither subtype variables remained independent prognosticators upon the adjustment of key clinicopathological features in multivariable analyses. Contrary to all other studies, the TCGA study revealed a significantly poorer RFS and DSS in stLMS C1 (high-myogenic) than stLMS C2 (low-myogenic)[36]. However, significance of the subtype feature was again lost in multivariable analyses. The inconsistent survival observations made by TCGA compared to other studies may be explained in part by the semi-supervised approach TCGA took to LMS subtyping, where uLMS and stLMS were separated prior to analysis. Overall, no robust association has been found between outcome and LMS subtype, thus at present there appears little prognostic utility in LMS molecular subtyping. More promising however, is this use of LMS molecular subtyping for predictive stratification. In other cancer types, high immune activity can be a favourable indictor for response to immunotherapy-based interventions. The observation of a high-immune LMS subtype may therefore reveal a candidate population for immunotherapy. However, given the inconsistency between immune findings in LMS and immune findings in other cancer types, assessment of this hypothesis would require significant in-depth immune profiling to be performed.

1.4.2 Undifferentiated pleomorphic sarcoma

UPS is a heterogenous group of pleomorphic tumours accounting for approximately 16-17% of STS diagnoses[235,303]. Tumours most often arise in the extremities, and the risk of UPS development increases with age. UPS possess no identifiable differentiation lineage, and there are no definable criteria for diagnosis. Instead, a UPS diagnosis is established through the exclusion of other STS subtypes[304]. Historically UPS were grouped under the MFH diagnosis. However seminal analyses in 1992, which leveraged developments in IHC technology, revealed differentiation lineages for approximately 2/3^{rds} of MFH (n = 159)[244]. This led to more routine implementation of IHC in STS diagnostics, and the reclassification of many MFH as pleomorphic LPS, LMS, or other poorly differentiated STS. Numerous MFH were also identified as non-mesenchymal tumours, and the remaining MFH with no discernible lineage were termed UPS. As with many STS, there are no subtype-specific guidelines established for UPS management,

and outcomes are poor. At 5-years, OS for UPS patients is 53-60%, LRFS is approximately 55%, and MFS approximately 70%[305,306]. However due to the 'catch-all' exclusion-based diagnosis of UPS, patients within this group show extreme heterogeneity and outcomes can vary greatly. Due to this high heterogeneity, UPS is often hypothesised as a group of multiple yet-to-be-defined STS, as opposed to a single disease type[307]. This has led to a dearth of molecular profiling studies; thus, UPS is one of the least molecularly characterised STS subtypes. As a subtype with huge unmet need, molecular profiling could have significant utility.

1.4.2.1 Gene fusions in UPS

Recurrent fusions have been characterised in a minority of UPS tumours (< 5%)[308,309]. A majority involve PR/SET domain 10 (*PRDM10*) fusions with either mediator complex subunit 12 (*MED12*) or Cbp/P300-interacting transactivator 2 (*CITED2*). Functionally, PRDM10 itself is poorly characterised, thus the mechanistic consequences of such fusions are speculative. Notably, other PRDM family members have been implicated in tumorigenesis. For example, inactivating mutations in *PRDM1* are noted in lymphoma, deletion of *PRDM4* prevalent in ovarian, gastric, and pancreatic cancer, and *PRDM16* fusions detectable in both myeloid and lymphoid cancers[310–315]. In these tumours, the PRDM family show functional duality with context-dependent tumour suppressive and oncogenic activity, further complicating the ability to hypothesise a role for PRDM in UPS[316]. The *PRDM10* fusion partners, *MED12* and *CITED2*, are better characterised and play major transcriptional, and developmental roles[317–319]. MED12 is a component of the mediator complex which can activate or repress transcription, dependent on its interacting factors[320]. Mutations in *MED12* have been observed in several other cancer types including in uLMS[319,321]. CITED2 is known to regulate ER activity and is upregulated in breast cancer[322,323]. Comprehensive profiling of *PRDM10*-fusion UPS has revealed a lack of co-occurring genomic alterations, indicating the fusions likely represent driver events, as in genomically simple STS[324]. Profiling also revealed *PRDM10*-fusion UPS to have a distinctive transcriptome, inconsistent with other UPS tumours, MyFS, dermatofibrosarcoma protuberans, and myxoinflammatory fibroblastic sarcoma. Differences have also been observed histopathologically, with *PRDM10*-fusion tumours showing low mitotic count and an absence of necrosis, indicative of a low-grade lesion. Whilst data is scant, these tumours appear to correlate to a more indolent clinical progression compared to other typically high-grade UPS. This therefore may reveal a subset of patient who require less aggressive treatment plans to achieve clinical benefit. Beyond the *PRDM10* gene fusions, further novel gene fusions have been identified in

UPS tumours, although these appear non-recurrent[309]. At present, the low occurrence rate of fusions in UPS, and incomplete understanding of downstream fusion-effects currently limits clinical applications.

1.4.2.2 Key molecular features of UPS

The molecular basis of UPS has been assessed in the context of other STS types. Specifically, TCGA profiled a significant number of UPS tumours (n = 44) in a mixed cohort of 206 samples[36]. Across CNA, miRNA, mRNA, methylation and RPPA analyses, most UPS samples were found to be indistinguishable from MyFS. Both UPS and MyFS historically fell under the MFH diagnosis; the discriminatory diagnostic feature being that MyFS possess a myxoid stromal component whilst UPS do not. In line with this, UPS and MyFS could be differentiated by review of genes differentially expressed based on the myxoid stroma. The similarities between UPS and MyFS have led to a hypothesised UPS-MyFS spectrum of disease. Supporting this are histological observations which report the detection of UPS-like (ie. myxoid stroma absent) areas within MyFS tumours[36,325]. Considering UPS and MyFS as a single broad disease type, common recurrent amplifications have been identified. These include 2 components of the Hippo signalling pathway: vestigial like family member 3 (*VGLL3*) on chromosome 3p and yes1 associated transcriptional regulator (*YAP1*) on chromosome 11q, amplifications of which occur in ~ 10-25% and ~ 3-10% of UPS/MyFS respectively[36,208,326,327]. *VGLL3* and *YAP1* encode cofactors of the TEA domain containing transcription factors and functionally enhance Hippo signalling activity to promote proliferation[328]. Interestingly, VGLL3 plays a role in both adipocytic and skeletal muscle differentiation, and *VGLL3* amplifications have also been noted in DDLPS and LMS, albeit to a far lesser extent[326,327,329,330]. It is notable that amplification is observed in the dedifferentiated form of LPS, and it would be of interest to assess *VGLL3* amplification relative to the LMS molecular subtypes (**section 1.4.1.2**), to investigate any association and dedifferentiation/undifferentiation. However, irrespective of the absolute subtype specificity of *VGLL3* amplifications, this highlights the Hippo axis as a candidate for therapeutic targeting in UPS and can be hypothesised as potentially targetable across complex genome STS with an undifferentiated phenotype.

In addition to highlighting candidate therapy targets, the TCGA molecular profiling of UPS/MyFS has also shown prognostic value. Several miRNAs have been revealed as significant independent prognostic factors in multivariable analyses. These include miR194-5p, which was identified as associated with a significantly improved MFS and DSS. In glioblastoma and breast cancer, miR194-5p plays a tumour suppressive role by

promoting apoptosis and inhibiting epithelial-mesenchymal transition (EMT) respectively[331,332]. However, whether these mechanisms are active in STS is unclear, particularly considering the paradoxical nature of EMT in a mesenchymal tumour. Similarly, miR-22-3p was also identified as associated with a significantly improved DSS. This has been previously reported in osteosarcoma, bladder cancer, cervical cancer and acute myeloid leukemia[333–336]. miR-22 is inhibitory towards many signalling pathways, yet the exact mechanistic consequences in STS are undefined. Interestingly, miR-22 downregulates PTEN, which is mutationally inactivated in a minority of UPS (~ 7%)[254,337]. Upregulation of miR-22 in UPS may therefore represent an alternative mechanism by which PTEN activity is lost in these tumours.

On the genome-wide scale, UPS show high chromosomal instability. WGD is observed in an exceptionally high proportion of UPS (~ 90%) and is a putative driver event in tumour development. The high genomic complexity in UPS coupled with high inter-patient variation has led to the hypothesis that UPS develop along distinct evolutionary paths. Investigations into this underlying evolutionary trajectory of UPS guided a proposal for 4 distinct routes of development (haploidization, genomic loss, chromothripsis, and endoreduplication; **Figure 1.6**)[207]. All pathways involve an early driver mutation within *TP53* or *RB1*, and WGD and/or chromothripsis. As part of the haploidization pathway, early *TP53/RB1* mutations are followed by extreme anaphase mis-segregation resulting in 1 hyperploid and 1 hypoploid daughter cell. The resultant genome-wide haploidy of the hypoploid daughter is then rescued by WGD resulting in a UPS tumour cell with a copy neutral LOH signature. In cases where the anaphase mis-segregation is minor, 1 daughter will exhibit large regions of LOH which can be rescued by single or multiple sequential WGD events (genomic loss pathway). Alternatively, mild anaphase mis-segregation or anaphase lagging can trigger chromothripsis followed by WGD (chromothripsis pathway), or WGD can occur spontaneously following *TP53/RB1* mutations, without a LOH/CNA trigger (endoreduplication pathway). Each of these pathways corresponds to tumours with unique CNA signatures. It is hypothesised that these map to 4 distinct UPS subtypes, which may explain the genomic diversity observed across this histology, however these subtypes are yet to be correlated to independent datasets beyond the discovery cohort. Clinically, improved evolutionary understanding for UPS which undergo multiple WGD events may have revealed an actionable window intervention if early drivers/WGD can be detected. This illustrates the potential for clinical translation of basic biology research that is rooted in attempts to understand the molecular basis of a disease.

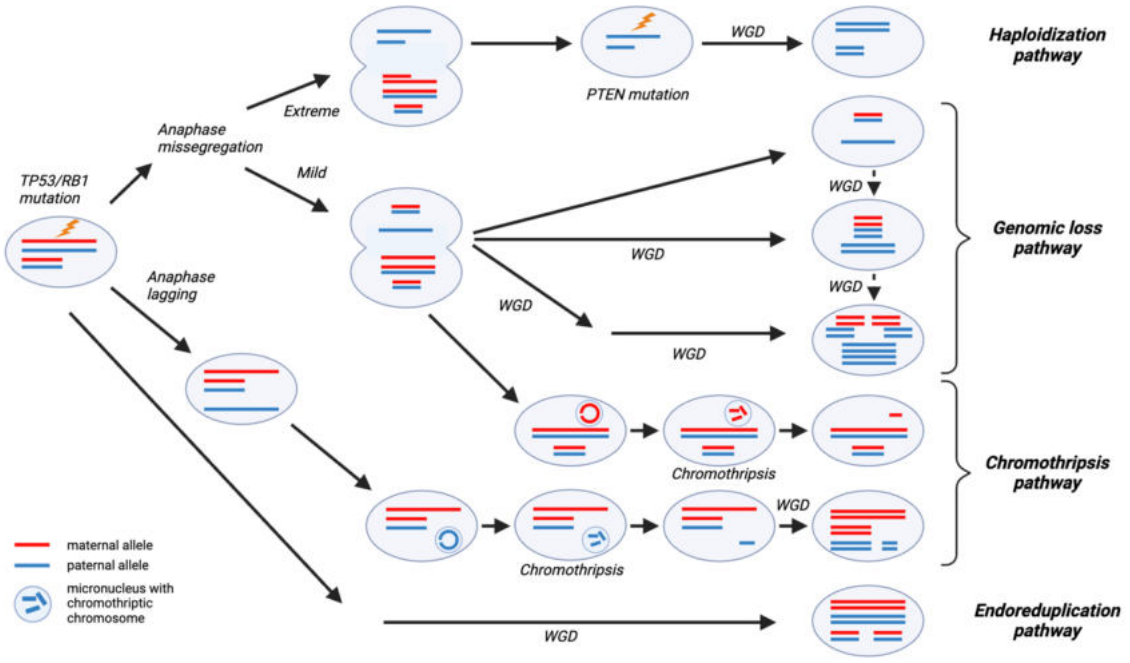


Figure 1.6 Overview of the hypothesised evolutionary development routes for undifferentiated pleomorphic sarcoma (UPS)
Adapted from Steele *et al*[207]. Abbreviations: WGD = whole genome duplication

1.4.2.3 Immune profiling in UPS

UPS is consistently reported to show high immune activity compared to other STS types, across IHC, RNAseq, and proteomic data[36,220]. UPS-specific assessments however have revealed a spectrum of immune activity within this histology. Molecular profiling of UPS (n = 25) identified molecular subtypes of UPS which correspond to variable CD8+ TIL infiltrate levels[338,339]. Namely, 3 subtypes (A, B, C) were identified based on RNAseq data and recapitulated using proteomic data with 82% precision. Of the 3, 2 (A, B) comprised 88% of samples and were assessed further. Subtype A was identified as enriched in genes for normal development and stemness. Subtype B showed enrichment of an array of immune activities and components including inflammatory and interferon gamma (IFNγ) response pathways. Furthermore, immune deconvolution highlighted signatures corresponding to CD4+ TILs CD8+ TIL, monocyte, NK cell, DC, memory B cell, and regulatory T cell infiltration as higher in subtype B. IHC was used to confirm subtype B as a CD8+ TIL high group. Associations between subtypes and outcome were interrogated, revealing a significantly improved MFS in subtype B compared to subtype A. A comparable correlation was observed in the TCGA cohort where the immune high UPS population showed a significantly superior OS compared to immune low UPS[36].

Interrogation of the WES data revealed an overall low mutational burden, as is expected in STS, with no recurrently mutated genes identified[339]. Notably, the tumours with the highest mutational burden rates (> 5 mut/Mb) were all classified as immune high. However, the TMB high samples account for only 36% (4/11) of the total immune high population, and relative to other cancer types would still be considered 'TMB low'. The importance of an immune-TMB relationship in this cohort is therefore not clear. The immune low tumours show higher CNA rates than the immune high; frequently showing deletion events in tumour suppressors involved in DNA repair, cell cycle, apoptosis, and the PI3K/mTOR signalling pathway. This reflects the high genomic complexity of UPS reported elsewhere, and furthermore delineates between 1) a subset of UPS showing complete immune response and a marginally higher mutational burden, and 2) a subset of UPS with low immune activity and higher CNA-driven genomic complexity. This relationship between aneuploidy and immune activity mimics reports in other cancer types. High CNA, and particularly high whole-arm or whole-chromosome CNA has been shown to correspond to lower expression of immune signatures and promote an immune evasive phenotype[340]. Furthermore, aneuploidy has been shown as a predictor of response to ICB. Immune-based molecular profiling in UPS has therefore identified subgroups of patients that may be vulnerable to different therapeutic strategies, and which may correspond to differences in patient outcome.

1.4.3 Dedifferentiated liposarcoma

LPS tumours arise from the adipocytic differentiation lineage, and can be sub-divided into pleomorphic LPS, myxoid/round cell LPS, well-differentiated LPS (WDLPS), and dedifferentiated LPS (DDLPS)[4]. Overall LPS account for between 15% - 20% of all STS, with WDLPS representing the most common LPS type (~ 50%)[234,235]. WDLPS is indolent is nature (7-year OS > 80%) and does not distally metastasise[341,342]. DDLPS, by contrast is an aggressive, higher-grade tumour that shows rapid growth and a high metastasis risk[341.] DDLPS arises subsequent to WDLPS and the 2 frequently co-occur as WD/DDLPS disease. DDLPS typically presents at the primary WDLPS disease site, although, in some cases (10%) can present alone as an independent recurrent lesion following primary WDLPS[343]. Diagnosis of WD/DDLPS is based on the molecular detection of ring or giant marker/rod chromosomes containing genetic material from the 12q13-15 region[344,345]. Histologically, WD/DDLPS show WD regions of low cellularity with mostly mature fat with fibrotic stroma and DD regions of dedifferentiated pleomorphic components. The most pressing risk for DDLPS patients is local recurrence, with patients often experiencing multiple recurrence events over many years. Distant metastasis can occur but at a much lower frequency than local recurrence. DDLPS of intermediate and

high grade show the same propensity for local recurrence (~ 40% at 7 years) but vary in likelihood of metastasis[341]. At 7-years post-surgery, approximately 10% of intermediate grade DDLPS will metastasise compared to 30% of high grade DDLPS. OS for DDLPS is comparable to other STS subtypes (~ 50% at 7 years). Molecular profiling in DDLPS has identified unique molecular features for this complex genome subtype, improving diagnostic confidence and accuracy. Beyond this however, molecular profiling DDLPS is limited. DDLPS has been profiled in the context of WDLPS to better understand malignant progression and help identify markers of progression. Yet compared to the comprehensive evolutionary studies in UPS and the transcriptomic subtyping of LMS, DDLPS tumours are yet to be comprehensively investigated.

1.4.3.1 Key molecular features of WD/DDLPS

WD/DDLPS is diagnostically characterised by amplicons of the 12q13-15 region. This region encompasses many genes reported as amplified in WD/DDLPS. Those of highest prevalence include *MDM2*, *CDK4*, and high mobility group AT-hook 2 (*HMGA2*), amplified in 96%, 96%, and 91% of WDLPS, and 91 - 100%, 91 - 100%, and 76 - 87% of DDLPS respectively[36,214–217]. Other genes of note include *CPM* and *YEATS2*, although data on their occurrence rate is highly inconsistent[216,346–348]. Irrespective of the exact composition of WD/DDLPS amplicons, amplification of these genes is central to WD/DDLPS tumorigenesis and alludes to common mechanisms driving tumour development. The molecular basis of WD-to-DD progression is less well understood. WD-to-DD progression is a time-dependent event that occurs in a subset of WDLPS patients. The co-occurrent nature of dedifferentated and well differentiated components has enabled studies to perform matched profiling of WD and DD samples from the same patient. One such study analysed a series of 17 patients and reported the WD and DD components to share only ~8% of mutations, suggesting early divergence between WD and DD lineages[349]. In general, WD-to-DD progression has been shown to correlate with increased genomic complexity, increased CNA, and an elevated level of *MDM2* amplification[350]. *MDM2* amplification is a near universal characteristic, however the extent of amplification has been shown to follow a log-normal distribution and varies across patients[351]. *MDM2* amplification level also appears to correlate with outcome. Higher amplification is significantly associated with a shorter recurrence free survival (RFS; n = 16), and a short OS (n = 25). The increased frequency of aberrations in DDLPS compared to WDLPS, is predominately due to losses on chromosomes 11, 13, and 15, and genome-wide amplifications[349]. Notable genes more recurrently amplified in DDLPS compared to WDLPS include *JUN* and mitogen-activated protein 3 kinase 5 (*MAP3K5*)[349,350,352–355]. *JUN* encodes c-Jun, a component of the activator protein 1 (AP-

1) transcription factor complex. MAP3K5 activates the c-Jun N-terminal kinase (JNK) signalling cascade, which triggers phosphorylation and activation of AP-1. AP-1 exhibits wide ranging transcriptional control, and specific to DDLPS is implicated in the differentiation of adipocytes through *JUN* and *MAP3K5*. Overexpression of *JUN* in mouse models results in suppression of adipocyte differentiation, and overexpression of *JUN* in liposarcoma specific models leads to aggressive and undifferentiated tumours[352,354]. In *in vitro* work illustrates *MAP3K5* overexpression to result in suppression of functional adipocyte maturation, and amplification and coordinate overexpression and *MAP3K5* has also been suggested to inhibit adipocyte differentiation in MFH[355]. The predominance of *JUN* and *MAP3K5* amplification in DDLPS has led to the hypothesis that these alterations facilitate progression from a WDLPS to DDLPS disease state through suppressing differentiation. *HMGA2* and *CPM* also show differentially altered patterns between WDLPS and DDLPS. Both *HMGA2* and *CPM* amplifications are detectable in WDLPS[353]. In fact, amplifications at the proximal regions of *HMGA2* are associated with WDLPS over DDLPS. However, unique to DDLPS, *HMGA2* and *CPM* fusions have been identified[349]. *HMGA2* fusions are predicted to retain protein function, leading to overexpression of functional HMGA2 in DDLPS compared to WDLPS. By contrast, *CPM* fusions are predicted to generate a truncated non-functional *CPM* transcript, resulting in a lower expression in DDLPS. *HMGA2* encodes a transcriptional regulator and *CPM* a membrane bound enzyme with important functions in monocyte to macrophage differentiation. The mechanistic consequences of altered *HMGA2* and *CPM* in WD/DDLPS is unknown, and whether a link between *HMGA2* or *CPM* and adipocytic differentiation exists is not reported. Interestingly, the ratio of *MDM2* to *HMGA2* amplification is prognostic[356]. *MDM2* amplification twice or more the level of *HMGA2* is associated with a shorter OS and MFS, although the mechanistic reasoning is unknown.

Differences in RB1, a recurrently altered gene across STS, have also been revealed between WD and DD. Matched profiling of different areas of the same LPS tumour identified DD regions to possess higher rates of *RB1* LOH (60% vs 12.5% in WD), *RB1* mutations (19% vs 0% in WD), and *RB1* promoter methylation than WD regions (11% vs 0% in WD)[357]. Accordingly, DD also showed generally lower and more heterogeneous RB1 expression as measured by IHC. Given the established tumour suppressive role of RB1 across cancer, it follows that *RB1* is increasingly altered in the more aggressive LPS type (DDLPS). WDLPS and DDLPS also show differences in telomeric maintenance. The ALT mechanism, detected by heterogeneity in telomere length, is active in DDLPS (30%) yet absent from WDLPS (0%)[358]. In other cancer types, inactivating mutations in *ATRX* or death domain associated protein (*DAXX*) have been

shown to promote ALT[359]. Accordingly, loss of ATRX or DAXX expression was observed ubiquitously across DDLPS with ALT and was not seen in either ALT negative DDLPS or WDLPS[358]. Across LPS, detection of ALT activity, as measured by heterogeneity in telomere length, correlates with a poorer prognosis, and in DDLPS specifically, ALT is associated with poorer PFS and OS[358,360]. Further to inactivation of *ATRX* or *DAXX*, ALT is also suggested to be induced by hypomethylation of telomeres[361]. Whether this mechanism is active in DDLPS is unclear. Genome wide hypomethylation within DDLPS has been revealed to contrarily associate with an increased DSS[36]. Although, this is notably a global hypomethylation status as opposed to localised telomeric hypomethylation, and ALT was not specifically assessed in this cohort. Hypomethylated tumours however did show fewer genome doublings, a lower leukocyte fraction, and lower T_h2 signature, alluding to immune heterogeneity across DDLPS tumours.

1.5 Clinical applications of molecular profiling in STS

1.5.1 Molecular profiling in STS diagnostics

It is widely agreed that molecular profiling can vastly improve accuracy and confidence in diagnostics. The impact of integrating molecular testing with classical histopathology in STS was assessed formally by the GENSARC trial[362]. As part of GENSARC, expert pathologists reviewed 384 tumours by histology and standard-of-care IHC, identifying 'certain' diagnoses for 43%, where 'certain' indicated the diagnosis as the only one possible. The tumours were then molecularly tested by FISH, aCGH, and/or RT-PCR, and reviewed again. Secondary review revealed 13.8% of diagnoses required modification, and 6% of 'certain' diagnoses could not be confirmed molecularly. Similar discordance rates between individual institution diagnoses and centralised diagnoses have been reported by TCGA (12%; 28/237), the French Sarcoma Group (FSG; 14%; 341/2,425), and at MSKCC (10.5%; 789/7,494)[36,363,364]. It follows that centralised histological review and molecular testing at specialist centres well-practiced in molecular diagnostics is recommended for routine STS care[44,365]. Whilst such efforts can improve the diagnosis of STS with established genomic alterations, many STS subtypes lack unequivocal molecular features. This leaves room for misdiagnosis or the absence of a specific diagnosis (not otherwise specified (NOS) disease), and subsequent incorrect disease management[366]. Accurate STS diagnosis is therefore a continual challenge and is a particularly acute problem for poorly characterised subtypes.

Novel tools are needed to address the limitations in STS diagnosis. One such effort employed DNA methylation profiling to improve sarcoma diagnosis accuracy[367].

Methylation profiling of 1,077 reference tumours, including STS with both simple and complex genomes, revealed 62 tumour methylation classes. The identified methylation classes showed high agreement with the STS diagnoses established by WHO; with 48 mapping to WHO classification entities, 9 mapping to subgroups of WHO classification entities, and 3 mapping to combined WHO classification entities. Building on the methylation classes as a reference point, the authors developed a random forest based classifier, which when applied to 428 further tumours identified diagnoses for 322 (75%). Of the 322, 263 matched the original diagnosis, however 59 were classified with high confidence to alternate diagnoses. Review of the 59 led to 55 being reclassified based on the methylation classifier (discordance rate = 17%). Notably, the classifier could not identify a diagnosis for 25% of tumours, a rate higher than that seen in pathologist-led diagnostics (7%, 0%, and 2% unclassified by TCGA, FSG, and MSKCC studies respectively)[36,363,364]. This may be resultant of methylation data failing capture complete biology, or may be attributable to an incomplete reference tumour cohort which did not span all STS subtypes.

1.5.2 Molecular profiling in prognostic stratification

It is notable that current prognostic tools used in clinic (**section 1.2.2**) do not incorporate molecular features. They therefore do not make use of the advancements in molecular profiling seen in cancer research. Across STS research, there are multiple molecular markers detailed as potentially prognostic, such as *PDRM10* fusions and select miRNAs in UPS (**section 1.4.2.1** and **1.4.2.2**), and immune markers across STS subtypes (**section 1.3.1.2**). However, translation from bench to bedside is a challenge across cancer types. In particular, poor reproducibility across cohorts and limited benefits in unselected populations means biomarkers rarely achieve clinical adoption. Translation is undoubtedly complicated further in rare disease types such as STS, where cohort accrual is often limited. Heterogeneity across STS has limited the use of single biomarkers. Alternate to biomarkers are multi-molecule signatures, which comprise a set of biomolecular features. Signatures capture a more comprehensive picture of disease state than single biomarkers and have shown promise in the prognostic stratification of complex genome and non-translocation associated STS. The identification of low/high-risk patients by such methods can enable informed decisions to be made regarding treatment pathways. High risk patients can be highlighted for adjuvant therapy, and patients classed as low risk can avoid unnecessary treatment.

1.5.2.1 Genomic signatures

The most advanced multi-gene prognostic signature in STS is the genome complexity index in sarcomas (CINSARC) signature. Developed and validated by the French Sarcoma Group, CINSARC is a 67-gene expression index with prognostic value for metastasis in non-translocation associated STS[368]. CINSARC classifies patients as either low risk (LR) or high risk (HR) based on metastasis likelihood and has been shown to outperform the FNCLCC grading system. CINSARC is comprised mostly of genes encoding cell cycle and chromosome integrity regulators. Whilst it is evident that these components are central to tumour metastasis, mutations in the CINSARC genes themselves are rare[36]. Global profiling has revealed associations between a high CINSARC score and genomic instability, WGD events, and high CNA[369]. Focused analyses on key gene regulatory axes (micro RNAs (miRNA) and DNA methylation) have also shown differential patterns based on CINSARC. HR tumours show overexpression of putative onco-miRNAs, and anti-correlations with miRNAs related to tumour suppressors (eg. *PTEN)*. Altered methylation has been observed between HR and LR on the global level, however no focal differences are identified at CINSARC gene *loci*. Due to this high genomic complexity, the exact regulatory mechanisms of CINSARC genes are undefined.

In recent years, CINSARC has been retrospectively applied in the clinical trial setting. Specifically, to trial material from the phase III ISG-STS 1001 RCT[370,371]. ISG-STS 1001 compared histology-directed chemotherapy with untailored anthracycline chemotherapy. Profiling found no difference in outcome between CINSARC LR and HR groups. This was unexpected and suggests CINSARC may not perform well in this population. However, considering the extensive validation CINSARC had undergone and high confidence in its ability to predict risk, another hypothesis raised is that the HR patients responded to chemotherapy and thus in post-trial analysis show outcomes comparable to LR patients. CINSARC was not developed to distinguish chemo-responders from non-responders. Yet, the components of CINSARC span conserved biological processes intrinsic to tumour aggressiveness. It is therefore hypothesised that the same genes may dictate response to therapy in addition to disease progression. Accordingly, the phase III CHIC-STS and CIRSARC RCTs are underway to assess CINSARC stratification for peri-operative care[372,373]. This in-clinic stratification has only recently been made possible by technological developments that permit CINSARC profiling on FFPE diagnostic biopsy material. This highlights the importance of parallel biological and technological developments to facilitate biomarker translation to the clinical setting[374,375]. CINSARC has also shown utility outside of STS, including in carcinomas and haematologic

malignancies[376]. This is particularly remarkable given the stark differences in origin between these tumour types: mesenchymal, epithelial, and blood-forming tissue respectively. CINSARC components are therefore hypothesised to show a high level of conservation across cancer types. It has been proposed that CINSARC could be used as a general marker for cancer aggressiveness.

There are limitations to the application of CINSARC. CINSARC was developed on STS with complex genomes, thus it's translatability across all STS subtypes is not clear. STS with complex genomes frequently show WGD (**section 1.3.1.1**), therefore in its current form CINSARC may not be optimal in tumours lacking aneuploidy. Indeed, in near diploid tumours, integration of miRNA and methylation data with CINSARC facilitated improved sub-stratification within HR and LR groups. Furthermore, application of CINSARC to a genomically simple STS (SS), revealed space for signature refinement. CINSARC did possess prognostic utility for SS, however this was driven largely by only 2 genes[377]. Cell division cycle A2 (*CDCA1)* and kinesin family member 14 (*KIF14)* were both independently associated with MFS at a level comparable to the complete CINSARC signature. Thus, in the application of CINSARC to STS with simple genomes, profiling all 67 genes may prove unnecessary. Such refinement of multi-gene signatures is an attractive avenue of research. Whilst technological advancements have abated many issues with multi-molecule profiling, as the number of genes in a signature increases, often so does the analysis time, associated costs, and amount of tumour material required for profiling.

Another limitation of CINSARC is rooted in its original purpose. The signature was developed to address specific limitations in the FNCLCC grading system, such as inter-pathologist variation, and limited utility in intermediate grade tumours, the neoadjuvant setting, and biopsies[70–72]. Therefore, it does not incorporate other known prognostic factors in STS. Since the publication of CINSARC, it has been explored in the context of other prognostic tools, such as the Sarculator nomogram. When applied an STS cohort, both CINSARC and Sarculator held independent prognostic value under multivariable analysis[378]. Within Sarculator groups (high-, intermediate-, and low-risk), sub-stratification by CINSARC revealed significant differences between MFS. Hybrid use of CINSARC with Sarculator improved prognostic performance with respect to both MFS and OS. Whether CINSARC can replace FNCLCC grade as a variable within the Sarculator nomogram is yet to be assessed and would be of interest if CINSARC were to be considered an alternative grading system, as per its original aim.

CINSARC has also been compared to the Genomic Grade Index (GGI). GGI was developed to better stratify early/intermediate breast cancer tumours[379]. GGI is a 97-gene expression signature, which spans 58% of the CINSARC genes. GGI offers prognostic utility for breast cancer that is superior to standard pathologic assessment alone, by classifying tumours as at a high or low risk of recurrence. In STS, GGI has been applied to a multi-subtype series of 678 tumours[380]. GGI classified 275 (41%) as GGI-low and 403 (59%) as GGI-high, illustrating good representation outside of breast cancer. Classification was found to be significantly associated with MFS (multivariable Cox regression HR = 2.23, 95% CI = 1.34-3.74, p = 0.0021). Comparative assessment showed significant overlap between GGI and CINSARC classification, with 71% of tumours assigned to the comparable risk groups. Inclusion of both GGI and CINSARC in multivariable analyses, revealed both as independent significant prognosticators for MFS, illustrating complementarity between the 2 signatures. The successful application of GGI to STS is reflective of the translation of CINSARC to non-STS malignancies, and conveys a generalisability in prognostic signatures, which often represent fundamental processes in tumourigenesis.

1.5.2.2 Tumour microenvironment signatures

Both GGI and CINSARC are genomic prognostic signatures, derived from characteristics of the tumour cells themselves. However, tumour cells do not exist in isolation, but sit within a TME inclusive of an immune component. Leveraging on the role of non-tumour cells in disease progression, the immune constant of rejection (ICR) signature uses immune features to predict outcome events in breast cancer[381]. ICR is a 20-gene signature encompassing genes encoding T_h1 signalling, chemoattraction, cytotoxic activity, and immune checkpoints. ICR classifies tumours as ICR1, ICR2, ICR3, or ICR4 where ICR1 has the lowest immune activity and ICR4 the highest. ICR4 possessed a notably strong T_h1 response with profiles enriched in cytotoxic and T_h1 cells. ICR4 showed a lack of any adaptive immune signatures, and were enriched in T_h17 cells, a pro-inflammatory T_h cell type characterised by production of IL-17. Retrospective application of ICR to a series of 678 STS revealed significantly poorer MFS in ICR1 compared to a pooled ICR2-4 class[382]. Lymphocyte infiltration as calculated by IHC was not associated with MFS. In contrast to Sarculator and GGI which show concordance with CINSARC classification, no significant associations were observed between ICR and CINSARC groups. Integration of ICR with CINSARC improved stratification, identifying 4 subgroups with differential MFS (the poorest being CINSARC HR/ICR1). Furthermore, integration of ICR, CINSARC, and histological subtype enabled the construction of a prognostic model to delineate ‘good prognosis MFS’ from ‘poor

prognosis MFS' patients (receiver operating characteristic (ROC) area under the curve (AUC) = 0.659).

Whilst the immune contexture is a TME module hypothesised to be a consequence of genomic complexity, hypoxia is a TME feature hypothesised to be a driver of genomic complexity[383]. Reports concerning individual hypoxia markers such as pO_2 and carbonic anhydrase 9 (CAIX) do reveal associations with MFS, however these are inconsistently reported within STS cohorts[384,385]. Multi-molecule hypoxic signatures have therefore been applied to STS. One of these signatures was developed for head and neck cancer and comprised 15 genes involved in 'hypoxia-influenced' pathways such as extracellular matrix (ECM) regulation and glycolysis[386]. This was applied to 132, most LPS and UPS, STS tumours, which were split into training and validation cohorts[387]. In both cohorts, DSS and RFS were significantly poorer in high hypoxia tumours, suggesting prognostic utility. However, the cohort size is insufficient to confidently claim, and integration with pO2 data (available for 16 tumours) found high pO2 (ie. low hypoxia) as associated with the high hypoxia gene signature, casting doubt on whether this signature is a measure of hypoxia applicable to STS. The hypoxia signature developed by Yang *et al* may be more appropriate as it was developed *de novo* using RNAseq data from STS cell lines exposed in a normoxic environment and 1% oxygen (ie. hypoxic) environment[388]. The resultant signature contained 24 genes. Overall, this signature separated 555 STS tumours into 'normal' (ie. low hypoxia) and hypoxic and reported the hypoxic patients to have a significantly poorer MFS in the training, validation, and external (TCGA) datasets. The 15-gene head and neck signature was applied to data in this study, however prognostic utility was only reported in 2 out of the 3 cohorts profiled (training and external). Application of CINSARC alongside this hypoxia signature revealed significantly highly representation of CINSARC HR (indicative of high genomic complexity) in the hypoxic group (78%) compared to the 'normal' group (48%). Moreover, within the TCGA cohort, CNA were also higher in samples categorised as hypoxic.

1.5.3 Molecular profiling in predictive stratification

In addition to prognostic stratification, patients can also be stratified for predictive purposes to differentiate between patients who will benefit from a particular therapy and those who will not. In oncology, it is well established that predictive stratification can provide improved cohort outcomes through prospectively selecting high-likelihood responders for treatment. Meta-analysis of 346 phase I and 570 phase II trials spanning over 43,000 patients has demonstrated the superiority of biomarker-guided intervention. Overall, phase I trials have shown a median PFS in biomarker arms of 5.7 months

compared to 2.95 months in non-personalised arms ($p < 0.001$)[389]. This gap is widened further in phase II trials where the overall median PFS in biomarker arms is 6.8 months compared to 2.8 months ($p < 0.001$)[390]. In STS, cohort sizes are drastically reduced, however similar observations have been made. Within genomically simple STS, established driver mutations or genomic alterations are often clear biomarkers for targeted therapies. For example, *KIT/PDFRA* mutations in GIST can dictate response to imatinib and *NTRK* fusion status can dictate response to larotrectinib/entrectinib (**section 1.2.3.3**). Furthermore, in *MDM2*-amplified tumours such as LPS, amplification status may act as predictive for response to MDM2 inhibitors. The first MDM2 inhibitor to be clinically evaluated was RG7112 in *MDM2*-amplified LPS, and there has since been extensive development in MDM2 inhibition[391,392]. More complex is predicting response to therapies with undefined or broad-spectrum mechanisms of action. As is predicting responses in STS with complex genomes. For example, impressive results to pazopanib are seen in a subset of patients (**section 1.2.3.3**). Yet in the trial, these results were masked by heterogeneity. Administering pazopanib to the STS-wide population is therefore not a viable option in the UK due to an overall poor response rate. The absence of a method to identify high likelihood responders underscores the withdrawal of pazopanib from routine UK clinical practice. Moreover, this illustrates a major challenge in improving STS patient outcomes, where the rarity and heterogeneity of the disease complicates clinical trial design and limits cohort sizes. In the absence of effective predictive stratification, patients continue to miss out on therapies which could be of benefit. One approach to address a rare patient population is to translate observations from other cancer types. Indeed, 2 of the drugs where predictive stratification in clinic appears most tangible for STS patients are the pan-cancer approved ICB pembrolizumab and breast and ovarian cancer approved PARPi.

1.5.3.1 Pembrolizumab

Pembrolizumab achieves low responses across STS (ORR 15.1%), although has been shown to elicit durable results in a subset of patients[393]. Pembrolizumab responders possess putative immune hot tumours. The definition of 'immune hot' however is ambiguous and often relative. Several biomarkers have been reported to characterise immune hot tumours and thus identify patients who may benefit form pembrolizumab. The earliest attempts in STS were based on the SARC028 trial[139]. Post-trial analysis of patient tissue reported higher PD-L1 expression in pembrolizumab responders[394]. However, this was based on only 2 evaluable tumours showing PD-L1 expression, restricting interpretation. Subsequent pooled analysis has shown an ORR of 28.5% in PD-L1 positive tumours compared to 6.7% in PD-L1 negative[393]. Yet as in SARC028, the

PD-L1 positive population here was small (< 16%) in comparison to the PD-L1 negative. As an alternative to transmembrane PD-L1, soluble PD-L1 (sPD-L1) has been shown to correlate well with ICB response in NSCLC and melanoma. Although sPD-L1 is yet to be comprehensively assessed in STS[395,396].

Despite apparent correlations between PD-L1 expression and ICB response, PD-L1 proved inconsistent as a biomarker for response across cancer types[397–402]. Immune response is complex and shows extension cross-communication between cell types. Therefore, an alternative to relying on the expression of a single molecule for prediction is to use multi-marker signatures. PD-L1 expression is associated with an increased immune infiltrate of PD1+, T_h1 CD4+, CD8+, and FoxP3+ Treg TILs, B cells, and DCs[132,403–405]. In STS, the SICs encompass many of these features (**section 1.3.1.2**), and thus may lend themselves to use as a ICB predictive signature. Indeed, the authors of SIC assessed the predictive capability of these subgroups in SARC028 trial tissue. Response to pembrolizumab progressively decreased from SIC E, to D, C, B, and A, reflective of the decreasing levels of immune expression across SICs. SIC E patients achieved highest benefit (ORR 50%), and notably no responses were seen in patients from SIC A and B[231]. SIC E is not only characterised by a high cellular immune infiltrate, but also by the presence of TLS. Interestingly, TLS presence has been utilised for patient enrolment in a phase II ICB trial. PEMBROSARC assessed pembrolizumab with cyclophosphamide in STS and was amended whilst underway to include a TLS-selected cohort[406]. Compared to the earlier recruited cohort where 40/41 patients were TLS negative, the TLS-selected cohort showed significantly longer PFS (4.9 vs 1.5 months) and a superior response rate (30% vs 2%). Since these observations, RCTs have been established to assess TLS-based selection for ICB therapy[407,408].

There is a well-established relationship between immune activity and genomic complexity. This includes the expression of PD-L1. Translocation associated STS are almost exclusively PD-L1 negative, and a higher TMB is reported in PD-L1 positive tumours[132,405]. Investigations into the TMB of PD-L1 positive tumours have shown high mutation rates in genes responsible for antigen presentation and T-cell infiltration[132], suggestive of a coordinated increase in immune activity. Tumours with high TMB are therefore hypothesised to be primed for ICB intervention. Indeed, tumours which show favourable response to ICB such as melanoma have a high TMB[409–411]. Conversely, a low TMB, as observed in STS, has been identified as an independent correlate for poor response to ICB. In addition to TMB, microsatellite instability (MSI) has also been highlighted as a candidate biomarker for ICB response[412–414]. MSI describes a

hypermutated phenotype where short repetitive DNA regions (known as microsatellites) accumulate deletion and insertion mutations, altering microsatellite length. This results in increased neoantigen presentation, a major determinant for PD-1 inhibitor response[415,416]. MSI is resultant of a defective mismatch repair (MMR) pathway. In colorectal cancer, a malignancy with frequent MSI, increased MSI is positively correlated with increased TMB and higher TILs[412]. The frameshift-inducing mutations of MSI often induce structural protein changes which can create antigen epitopes increasing tumour immunogenicity, as demonstrated by increased TILs. In STS, MSI is rare, but has been noted in case reports of ASPS, an ultra-rare fusion-positive STS, possibly offering an explanation for the strikingly high response rates that ASPS have shown to ICB[417–419] (**section 1.2.3.4**).

1.5.3.2 Poly (ADP-ribose) polymerase inhibitors (PARPi)

PARPi target the PARP family of enzymes, which are central to DNA damage repair pathways, and specifically to base excision repair of single strand DNA breaks. If single strand DNA breaks persist, DSB occur, requiring repair by HRR or non-homologous end joining (NHEJ). Whilst HRR is a conservative mechanism, NHEJ is highly error-prone and can lead to accrual of genomic instability and cell death. In cells with HRR deficiencies (HRD), inhibition of PARPs induces synthetic lethality due to a reliance on NHEJ activity. Oncogenic alterations in the HRR pathway are typically resultant of an altered *BRCA1/2*. It is therefore unsurprising that PARPi were first investigated for use in breast and ovarian cancer; where a subset of patients harbour germline *BRCA1/2* mutations[420–423]. Since, PARPi have been approved for select breast and ovarian cancer patients, and explored in other cancer types[278–280].

In STS, mutational rates in BRCA1/2 are low (~ 1-12% and ~ 1-6% respectively), however mutations in HRD and *BRCA* associated genes do occur at higher frequency. For example, BRCA1-associated protein 1 (*BAP1*) and fanconi anemia complementation group C (*FANCC*) mutations have been reported to occur in up to 29% of patients[424]. Notably, these number are not reflective of PARPi sensitivity, but provide insight into pathway alteration rates. Beyond considering single mutations, objective analysis aimed at defining somatic mutational signatures of STS has also been performed using WES data. This revealed 30 signatures, 1 of which corresponded to defects in DNA-DSB repair by HRR. Application of these signatures in the TCGA cohort revealed 37.05% of tumours to show BRCAness characteristics of high HRD and high CNA. BRCAness signatures present across subtypes but appear more enriched in osteosarcoma populations (> 80%)

and LMS populations (**section 1.4.1.1**)[425]. In agreement with this, pre-clinical work has long demonstrated a high sensitivity of osteosarcoma cell lines to PARPi, and favourable results are seen in clinic with PARPi intervention in LMS[426–428]. An early case study report details 4 heavily pre-treated advanced LMS patients selected for olaparib therapy based on detection of pathogenic *BRCA2*[276]. At the time of publication, 3 patients remained on olaparib with stable disease at 16 weeks, 16 months, and 17 months. Olaparib stabilised disease in the 4th patient for 15 months prior to progression. A phase II trial of olaparib with temozolomide in advanced uLMS (n = 22) has been conducted. This showed positive and durable responses (ORR = 27%; median duration of response = 12 months)[150]. A subsequent larger scale RCT is planned to follow these results, as well as post trial analysis to assess the interplay between HRR deficiency, PARPi resistance, and response in LMS; however, at present this data is not reported.

In addition to the use of BRCAness as a biomarker, *ATRX* has also been suggested as predictive of PARPi response. ATRX is implicated in DDR but its exact role is not defined. Pre-clinical *in vitro* work has shown loss of *ATRX* to promote ATR signalling and induce replication stress, which can be amplified by PARPi to induce cell death[429,430]. Significant further work needs to be conducted to establish whether pathogenic *ATRX* confers PARPi sensitivity in STS, however this is particularly interesting given *ATRX* is one of the few recurrently mutated genes in STS. Notably, a phase II trial assessing a combination of PARPi and ATR inhibitor ceralasertib in osteosarcoma is underway (NCT04417062), as are similar trials in prostate and ovarian cancer[431–434]. PARPi have shown limited benefit as monotherapies, and it is hypothesised that dual targeting of the DDR pathway may elicit a more potent effect.

1.5.3.3 Clinical integration of predictive stratification

In pembrolizumab and PARPi, biomarkers have been investigated and often validated across cancer types by large retrospective profiling experiments. Yet, STS is a heterogeneous disease, and thus the target population for these drugs is small. If strides are to be made in STS treatment pathways and advances in molecular profiling are to be leveraged effectively in the near future, profiling must be integrated into care. In line with these ambitions, several trials are underway to assess the feasibility and effectiveness of using NGS to identify treatment-linked alterations (TLA) and guide therapy. TLA encompass well validated biomarkers, as well as less comprehensively studied molecular features. Whilst founded in research, these features are often not robustly assessed in the target population. Tissue-agnostic studies assessing the integration of NGS to identify TLAs include the NCI MATCH and American Society of

Clinical Oncology (ASCO) TAPUR phase II basket trials. In these trials molecular profiling, the method of which is dependent on the trial arm, is performed to identify TLAs matching a drug in the treatment arms[435,436]. Similarly, the NCI comboMATCH trial aims to identify TLAs for combination therapy regimens, and the NCI iMATCH trial aims to stratify for immunotherapies[437,438]. As of 2020, NCI MATCH had accrued data from 5,954 patients, identifying a TLA in 37.6%, and assigning a treatment arm to 17.8%[439]. Notably, if all NCI MATCH treatment arms were open at once, it would have been possible to treat 26.4% of patients within this trial. As present, 3 NCI MATCH subprotocol arms have reported. Briefly, 48 patients with fibroblast growth factor receptor (FGFR) pathway TLAs were assigned to the FGFR inhibitor AZD4547, 61 patients with *PIK3CA* TLA were assigned taselisib, and 25 patients with *PIK3CA* TLA were assigned copanlisib[440–442]. Copanlisib achieved an ORR of 16%, however the results for taselisib and AZD4547 have been underwhelming; the former showed only limited activity, and the latter failed to meet the primary endpoint. Poor responses may be attributable to the high heterogeneity of pan-cancer cohorts and/or the insufficient performance of TLAs in predicting response. However, more data is needed before conclusions as to the effectiveness of TLA-guided care can be drawn.

Whilst STS patients are eligible for inclusion in NCI MATCH and ASCO TAPUR, no reports noting recruitment or response of STS cases have been made. Early profiling experiments specific to STS have detected an abundance of TLAs. One study reported 60% of patients (n = 25) to harbour one or more TLA for which clinical trials were ongoing at the time[443]. Whilst another study identified TLAs in 61% of patients (n = 102) and assigned 16% a targeted therapy based on TLA detection, 50% of whom showed stable disease at the time of publication[444]. A more recent and more comprehensive (n = 5,635) retrospective study utilising targeted NGS revealed 16% of sarcoma patients harboured a FDA-approved drug TLA, 7% a study drug TLA, 42% a TLA within the MATCH or TAPUR trials[445]. The authors went on to screen 107 patients who were alive with advanced disease, finding 57% had at least one TLA. Of the 57%, 30% were enrolled on corresponding clinical trials. In addition to guiding therapeutic decisions, profiling also identified resistance-associated mutations in 5% of patients, therefore avoiding ineffective treatment. Since these reports, the STS specific MULTISARC phase II/III trial has been established[446,447]. MULTISARC aims to formally compare NGS guided treatment to standard of care in advanced STS patients. At present the trial is still recruiting and no results are reported.

Integrated NGS trials have demonstrated good feasibility, identifying a high proportion of patients with TLAs, and assigning treatments to 16 – 30% of patients. Moreover, these studies show that with cross functional collaboration, tissue profiling, analysis of the results, and treatment decisions can be conducted in a reasonable timeframe for the patient. Questions on the futility of such an approach remain. At present it is unclear whether patients will significantly benefit from TLA-guided treatment. In patients with advanced STS the difficulty in establishing robust and appropriate RCTs has led to treatment decisions being made based on small datasets of case reports and early phase clinical trial results[44]. It is therefore not unreasonable to assume that using TLAs reported in small datasets to guide treatment decisions could improve such practice and aid the treatment advanced STS patients where no evidence-based treatment guidelines exist.

1.6 Proteomic profiling in STS

Molecular profiling in STS has predominately utilised genomic and transcriptomic methods. This has greatly advanced STS disease understanding and contributed to the improvements in clinical practice. Yet, relative to the amount of research conducted, findings have been rarely translated to clinic. Proteins are the mediators of cell communication and activity, and therefore are key effectors of a cell. One explanation for the ineffective translation of current research in STS may be the lack of a proteomic disease understanding. Indeed, correlation between genomic/transcriptomic readouts and protein-level data are poor, thus using genomics/transcriptomics to describe protein-based activity may not be appropriate[448,449]. Using genomics and transcriptomics to describe a protein-based cellular function or activity is therefore not always appropriate. Proteins govern a wide range of cellular activities, and are central to cell structure, function, and regulation[450]. Sitting downstream of the genome and transcriptome, proteins are a readout of gene expression, and as such are sensitive to the mutational alterations observed in cancer. Proteins are also responsive to changes in both the intracellular and extracellular environments and are under dynamic regulation through post-translational modifications (PTM)[451]. The proteome is therefore not static. As a result, the final form and expression levels of a protein can differ vastly from properties inferred at the gene or transcript level. Proteomics, the study of the proteome, offers many uses and applications[452] (**Figure 1.7**). It can provide an accurate representation of the tumour and better understanding of disease understanding. In achieving a fundamental understanding of the 'active' component of a tumour through proteomics, biological findings may be more readily translated to the clinic, aiding improvements for patients. Specifically, proteomics can be used to identify candidate biomarkers, whether

prognostic, predictive, or diagnostic. Proteins are an attractive biomarker molecule, as they can be readily assessed by IHC, a well-established method already routinely used in clinical diagnostics[453,454]. In addition, given most drugs are directed towards proteins, proteomics is uniquely positioned to aid the identification of candidate drug targets[455].

It is important to acknowledge that proteomics, akin to all single-omic modules, encompasses only 1 component of a complex biological system. Therefore, whilst proteomics can contribute to disease understanding, alone it cannot comprehensively cover all tumour biology. There is a therefore need for multi-omics work spanning multiple biological modalities. Ideally, multi-omics involves profiling the same sample by multiple methods. However multi-omics can also be thought of as the integration of knowledge learnt from independent studies. Given the plethora of genomic and transcriptomic studies in STS, proteomics can complement this current literature.

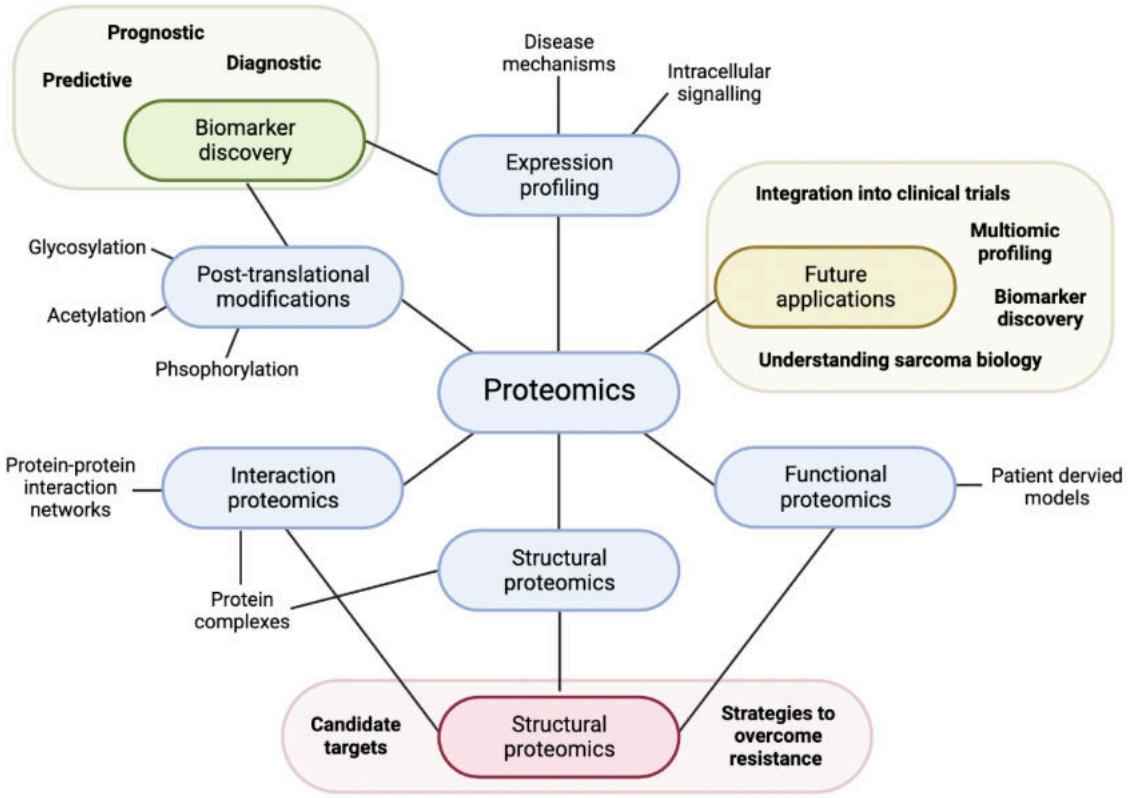


Figure 1.7 Overview of the applications of proteomics

1.6.1 Proteomic methods

Due to technical challenges and the expanse of proteins and different proteoforms encoded by the same gene , the human proteome itself is not yet fully characterised[456]. Thus, the term proteome loosely refers to global protein analyses which attempts to capture a relatively large proportion of the human proteome. This is typically in the order of several hundred to several thousand proteins. One of the major tools for unbiased analysis of the proteome is mass spectrometry (MS). Many different variations of MS exist; however, the fundamental principles remain the same. MS measures the mass-to-charge (m/z) ratio of molecules[457,458]. There are also non-MS-based proteomic methods such as antibody arrays. Whilst these are generally less comprehensive than MS, they are easily implemented and relatively low-cost; thus, their use in research is common[459,460]. All methods used to study the proteome, can be broadly separated into targeted approaches and unbiased approaches.

1.6.1.1 Targeted proteomics

Targeted proteomics describes a supervised profiling approach, whereby specific prior knowledge of a protein or proteins of interest is required to facilitate their identification and quantification. Targeted proteomics can be non-MS-based or MS-based. The non-MS-based methods used in STS research typically involve microarrays such as reverse-phase protein arrays (RPPAs) or antibody arrays. The former entails loading of tumour lysate onto a microarray and probing with antibodies (**Figure 1.8A**), whilst the latter entails the reverse, loading of antibodies onto a microarray and probing with tumour lysate427. These allow for the simultaneous assessment of numerous proteins/samples in a way that is rapid and requires minimal sample material. However, the use of antibodies has its limitations. Antibodies may not be truly specific to a target or may not be available for a protein of interest428. Microarray profiling is therefore restricted and often does not exceed several hundred proteins. By contrast, targeted MS does not suffer from antibody-reliance. Key targeted MS methods include selected reaction monitoring (SRM) and multiple reaction monitoring (MRM; **Figure 1.8B**). SRM scans a single fixed m/z window to isolate ions (i.e., precursor ions) of a particular m/z value[462]. These are then fragmented, and the fragmented ions (i.e., product ions) isolated and measured. MRM follows the same procedure but scans multiple m/z windows to isolate multiple ions. SRM and MRM have higher specificity and sensitivity compared to non-MS approaches. However, as with microarrays, targeted MS requires prior biological knowledge to identify proteins of interest. Additionally, technical knowledge of the peptide fragmentation profile of a protein of interest is also needed; to identify the correct m/z

window for scanning. Whilst targeted proteomics is useful for defined hypotheses, the dependency of these methods on prior understanding means they are not appropriate for comprehensive discovery-based profiling.

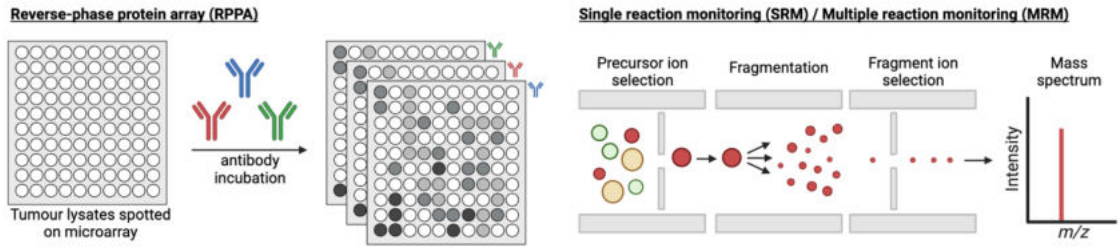


Figure 1.8 Overview of targeted proteomics approaches

1.6.1.2 Unbiased proteomics

Instead, comprehensive discovery-based proteomics, or 'unbiased proteomics' primarily involves the use of shotgun MS. In a typical shotgun MS workflow, peptides are separated based on polarity by liquid chromatography (LC)[458]. Peptides are injected into an LC column coupled to a mass spectrometer and are ionised as they elute to generate gas-phase ions. Inside the mass spectrometer, peptide ions (i.e., precursor ions) of the highest intensity are selected at the MS1 scan for fragmentation. The resultant fragmented ions (i.e., product ions) are then analysed in the MS2 scan, generating a tandem MS (MS/MS) spectrum. The MS/MS spectra produced are searched against known spectra in protein sequence databases, to assign a peptide and subsequently protein of origin. This is known as data dependent acquisition (DDA). Relative peptide quantification information can be extracted based on precursor signal intensities or spectral counting[463]. However, these methods of quantification, known as 'label-free', show low accuracy[464]. Superior approaches to relative quantification utilise multiplexed isobaric labels such tandem mass tags (TMT). TMT labels comprise an MS/MS reporter group, spacer arm, and an amine reactive group (**Figure 1.9A**)[465]. When labels are incubated with peptides, the amine reactive group binds to a peptide at either the N-terminus or a lysine residue. The TMT tags are isobaric and thus have identical masses. This means the same peptide labelled with different tags will show the same behaviour inside the LC column and mass spectrometer. As a result, identical peptides are co-isolated irrespective of labelling. Following precursor ion selection, fragmentation induces cleavage of the TMT labels, generating a unique reporter ion that is detectable

within a low m/z ratio spectral range (**Figure 1.9B**). The reporter ions can be used for relative quantification. Recent developments have led to the production of up to 18

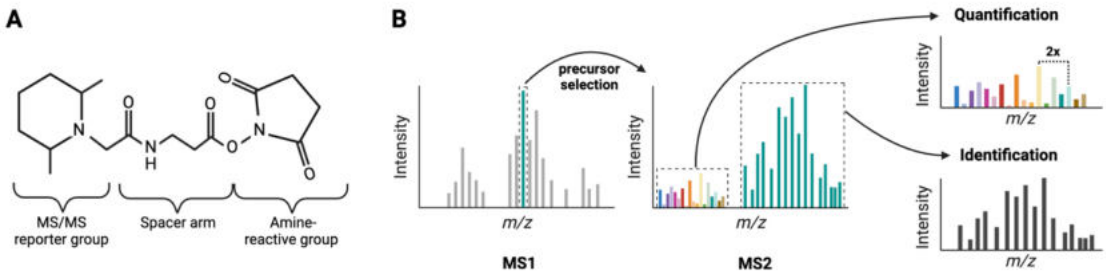


Figure 1.9 Overview of Tandem Mass Tag (TMT) quantitation in mass spectrometry (MS)
(A) The structural basis of TMT labels. **(B)** Diagrammatic representation of how TMT is used for quantitation in tandem MS (MS/MS)

different TMT labels, each with a unique MS/MS reporter group[465]. The 18 labels can be pooled and analysed simultaneously within the mass spectrometer, enabling relative quantification of up to 18 samples at once. The major advantage of simultaneous profiling is the low missingness achieved across samples[464,466]. Missingness describes the situation where a peptide/protein is identified/quantified in one sample but not another. Missingness is prevalent within DDA MS, due to stochastic selection of precursor ions in MS1. This introduces difficulties in downstream data handling as it is not possible to determine whether the missing value represents an unexpressed protein or an unselected precursor ion. Many bioinformatic tools for MS data analysis require complete data and therefore missingness must be addressed, often by removal, imputation, or dimension reduction[467–470]. However, in multiplexing (e.g., with TMT) the precursor ions selected are derived from all samples the original peptide was present in, thus missingness within TMT batches is low. Where experiments require more samples than the TMT limit (currently 18), multiple batches can be performed. In multi-batch TMT, a reference sample containing material representative of the other samples is typically included and occupies 1 label channel within each batch. The reference sample is subsequently used in downstream data processing to adjust for inter-batch variations and facilitate the merging of datasets from multiple TMT batches.

Another MS-based proteomic analysis method is data independent acquisition (DIA). In DIA, all peptides within sequential m/z windows are fragmented (**Figure 1.10**)[471,472]. This generates complex MS/MS spectra from multiple peptides, which require deconvolution.

As peptide ions are not stochastically sampled based on intensity (as in DDA), label free quantification in DIA is considered more accurate than DDA methods. DIA also shows improved reproducibility of peptide identifications due to low missingness and requires far less sample material than DDA with label quantification. For example, in-house approximately 1-2 ug peptide is required for DIA injection, whereas, although only 1-2 ug of peptide is injected in DDA analyses, to facilitate fractionation, 25-100 ug peptide is recommended for TMT labelling approaches[473]. DIA is therefore particularly useful for samples where peptide yields are small. The main limitation of DIA is the lower proteome coverage compared to most DDA experiments[472]. One reason for this is the ability to couple DDA methods with additional orthogonal fractionation prior to LC. In these cases, samples are fractionated off-line, and each fraction is injected into the LC-coupled mass

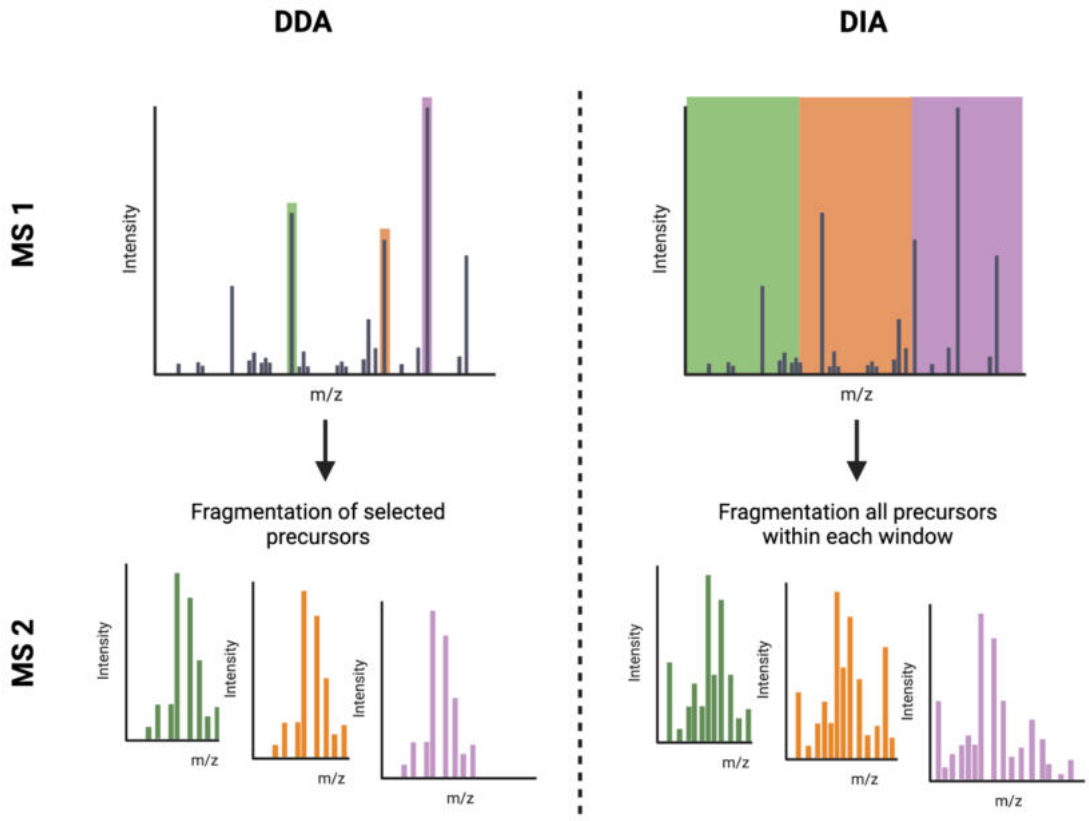


Figure 1.10 Diagrammatic comparison of data dependent acquisition (DDA) and data independent acquisition (DIA) in mass spectrometry (MS)

spectrometer separately. Fractionation in DIA analysis is possible, yet only minor improvements in the number of proteins identified are seen, thus due to the extensive additional work required this it is not routinely performed[472]. Dependent of the additional fractionation steps, DDA experiments can identify upwards of 10,000 proteins, whilst state-of-the-art DIA detects up to approximately 4,000 proteins[474–478]. Recent developments in computational deconvolution strategies have demonstrated an increase in the number of proteins identified by DIA[479]. Yet at present these deconvolution methods identify more proteins at the cost of data completeness, by introducing a high number of missing values.

1.6.1.3 Sample types for proteomics

In proteomic workflows, one of the most crucial steps is protein extraction. Unlike nucleic acid, proteins cannot be amplified (i.e., by polymerase chain reaction (PCR)) prior to analysis. As such, the extraction of proteins is far more challenging than DNA or RNA, and sample requirements for proteomics are often difficult to meet. Small samples such as biopsies often do not contain enough protein for comprehensive analysis, and the high risk of sample exhaustion restricts extraction attempts. Specifically, this has limited the use of DDA with label quantification; the method which can provide the deepest proteome coverage, but demands some of the highest peptide input material.

To maximise protein yields, studies profiling tumour proteomes by TMT DDA (e.g., CPTAC) often utilise fresh frozen (FF) tissue[157]. FF has undergone minimal processing, and therefore proteins can be readily extracted in high yields. Unfortunately, obtaining such material in sufficient numbers for a rare cancer is often not feasible. By contrast, formalin-fixed paraffin-embedded (FFPE) material is widely available. FFPE storage is routine in biobanking to ensure long-term sample stability at room temperature. However, formalin-processing introduces crosslinks between biomolecules which further complicates protein extraction[480–482]. Formaldehyde reacts with primary amine groups or thiol groups in proteins and nucleic acids to form stable inter-molecular methylene bridges. For example, formaldehyde can covalently bond a primary amine group of lysine to form an aminomethylol group. The methylol group can then condense with a free residue (e.g., primary amide, secondary amine or guanidyl) to form a crosslink. Whilst practical advancements now allow for effective reversal of crosslinks, yields from FFPE tissues remain significantly lower than those from FF tissue. In STS, the proteomes of FFPE tissue have been characterised, but at present this has only been achieve through utilising DIA which has exceptionally low input material requirements[483].

1.6.2 Overview of the current status of proteomics in STS

Comprehensive (i.e., MS-based) proteomic profiling in STS is limited[452]. TCGA incorporated RPPA analysis of 206 STS tumours using 192 antibodies; yet this covers only a small proportion of the proteome[36]. STS is included in the list of malignancies selected for CPTAC MS profiling, although data has yet to be deposited or published. Thus far, only smaller-scale proteomic studies focused on specific histological subtypes are reported.

Protein and phosphoprotein profiling has been performed using TMT MS/MS in 17 rhabdomyosarcoma (RMS) orthotopic patient-derived xenograft (O-PDX) models[484]. Clustering showed distinct (phospho)proteomes between embryonal RMS (ERMS), alveolar RMS (ARMS), and human myoblasts and myotubes (the putative cells of origin for RMS). ERMS and ARMS showed significantly different expression profiles of several key muscle development pathways and proteins. Specifically, whilst ARMS showed consistently high myogenin and low myogenic factor 5 expression, ERMS showed a more varied profiled. This suggests ARMS may arise further along the muscle development lineage than ERMS. Many proteins were also differentially expressed in both ARMS and ERMS relative to myoblasts and myotubes. One such protein was Wee1, a tumour suppressor that regulates cell cycle progression through the G2/M checkpoint. Wee1 is also a therapeutic target in many tumours which rely on the G2/M checkpoint to maintain genome stability[485–487]. Upon, *in vitro* treatment of RMS cell lines with a Wee1 inhibitor (AZD1775), cell cycle arrest, mitotic catastrophe and nuclear fragmentation were observed. Moreover, when AZD1775 treatment was combined with RMS standard of care (irinotecan and vincristine) extensive DNA damage was induced. The authors hypothesise DNA damage to be suggestive of genomic instability, which can induce cell death; illustrating therapeutic potential. Notably, this work does not profile human tumour material. Whether the O-PDX findings will be recapitulated in human tissue is yet to be seen.

Tumour material from GIST patients has also been profiled[488]. One study employed TMT MS/MS to profile FF tumours specimens and matched normal tissue from 13 GIST patients. In total, 704 proteins were identified as differentially expressed between tumour and normal tissue. Of these there was a noted enrichment in spliceosome components and an underrepresentation of carbon metabolism (Krebs cycle) proteins. Additionally, expression of protein tyrosine phosphatase non-receptor type 1 (PTPN1) was identified to correlate with risk. Low risk patients, as defined by the NIH GIST risk criteria, showed significantly higher PTPN1 expression than intermediate and high-risk patients. This

association was validated by IHC in an independent series of 131 patients. PTPN1 is an established tumour suppressor that mediates cell adhesion, motility, and invasion[489,490]. Loss of PTPN1 is therefore hypothesised to confer metastatic potential. However, this could not conclusively be determined due to the low numbers of patients included and the unavailability of clinical outcome data in this study[488]. PTM profiling of GIST tumours has also revealed protein acetylation differences between risk levels[491]. TMT MS/MS was employed to assess FF GIST tumour specimens from 9 patients; 3 of low risk, and 6 of intermediate/high risk. Key findings included the upregulation of acetylated Ki-67 (K1063Ac) in intermediate/high risk GIST. Ki-67 is a nucleic antigen and marker of cell proliferation. The impact of K1063 acetylation in Ki-67 is unknown, however acetylation can dramatically modify protein function, altering the hydrophobicity and solubility[492,493]. This effects a range of protein functions including the ability to interact with other molecules. Ki-67 has been reported as prognostic in GIST, whether the acetylation status drives or alters the association between Ki-67 and outcome is unknown[494,495].

Proteomics profiling has also been conducted to profile a multi-histology cohort. Milighetti *et al* profiled FFPE tissue from 36 STS patients spanning LMS, DDLPS, UPS, and SS diagnoses[483]. This study noted distinctive proteomes in SS and LMS patients. LMS showed an enrichment of muscle related ontologies, and SS were enriched in splicing ontologies. DDLPS and UPS showed more mixed proteomic profiles. Clustering failed to distinguish DDLPS and UPS from each other, however supervised analyses did reveal UPS as specifically enriched in immune activity. This study also identified numerous proteins with prognostic significance for OS. These proteins were used to stratify the patient population into 3 groups, identifying 1 group, containing UPS, DDLPS, and LMS patients, with a significantly poorer OS. This not only demonstrates the clinical potential of proteomic profiling, but also the utility of profiling multiple histological subtypes together to reveal histology intendent patterns across STS.

1.7 Conclusions, hypothesis, and aims

The current understanding of STS biology is incomplete which has hindered improvements in patient management and outcome. Numerous studies provide evidence that molecular variation both within and between histological subtypes exists. Heterogeneity in treatment responses and clinical course is also observed across patients. Whilst histology can explain some of this variation, the relationship between molecular and clinical heterogeneity is mostly undefined. As a result, current standard of care for most STS fails to consider biological heterogeneity. Accordingly, there is a

pressing unmet need to develop in depth disease understanding, and to leverage such knowledge to improve patient care. Specifically, there is a need to identify candidate biomarkers and drug targets in STS. At present, many candidate prognostic and predictive biomarkers have been reported, yet data is often not consistent and rarely have these candidates been successfully translated to clinic.

Biological understanding can be greatly improved through comprehensive molecular profiling of tumour specimens. Indeed, increasing efforts to molecularly profile STS have been made over the recent decades. Such attempts have predominately utilised genomic and transcriptomic methods, yet there remains an absence of any proteomic understanding of the disease. Proteins are the central effectors of cellular processes and are targets for the vast majority of drugs. As such, it is crucial that the protein complement of STS is understood. Protein-level data can also be integrated with the current genomic and transcriptomic dominant literature. Biological systems are complex and multi-faceted, thus multi-omic profiling is necessary for a complete disease understanding to be developed. The lack of proteomic understanding in STS is undoubtedly impeding advancements in disease understanding and by extension clinical care.

In line with this, the hypothesis of my thesis project is that **deep characterisation of the proteomic profiles of STS across multiple histological subtypes will reveal oncogenic pathways, and candidate biomarkers and drug targets of clinical relevance**. This hypothesis will be addressed with the following aims:

Aim 1: To profile the STS proteome of multiple histological subtypes (**Chapter 3** and **Chapter 4**)

Aim 2: To investigate intra-subtype heterogeneity in LMS, DDLPS, and UPS (**Chapter 5**)

Aim 3: To assess and characterise the unbiased, protein-centric pan-subtype STS proteome (**Chapter 6**)

Chapter 2 Materials and methods

2.1 Research ethics and data management

Collection of FFPE tissue and clinical data was approved as part of the Royal Marsden Hospital (RMH) PROgnoStic and PrEdiCTive ImmUnoprofiling of Sarcomas (PROSPECTUS) study (RMH Committee for Clinical Research reference 4371, NHS Research Ethic Committee reference 16/EE/0213), National Taiwan University Hospital (Research Ethics Committee Reference 201912226RINB), and as part of Children's Cancer and Leukaemia Group (CCLG) Biological Study 2012 BS 05 (Research Ethics Committee Reference 8/EM/0134). FreezerPro laboratory management software (Brooks Automation, Chelmsford, MA, USA) was used for logging tissue and tracking sample usage, in accordance with the Human Tissue Authority Codes of Practice and Standards. Pseudonymised clinicopathological data was stored in a locally maintained, and password-protected MySQL database (Oracle, Austin, TX, USA), and analyses performed blind to personal identifiable data. Samples obtained through external collaborators were obtained under Material Transfer Agreements.

2.2 Cohort generation

2.2.1 Patient selection and sample retrieval

Patients were selected for inclusion based on the following criteria: 1) histopathologically confirmed diagnosis of AS, ASPS, CCS, DDLPS, DES, DSRCT, EPS, LMS, RT, SS, or UPS, 2) > 18 years of age at the time of sample collection (excluding RT), 3) FFPE tumour material available in quantities sufficient for analyses. Patients were excluded if the primary tumour specimen was FNCLCC grade 1. All RT samples and 2 AS samples were obtained externally from Newcastle University, England (Dr Daniel Williamson, Dr Stephen Crosier) and National Taiwan University Hospital (Dr Tom Wei-Wu Chen), respectively. All other samples were retrieved through RMH. Diagnoses were confirmed by expert histopathological review by soft tissue pathologists (Dr Khin Thway, Prof Cyril Fisher). Eligible patients were identified by retrospective search of the hospital databases, and inclusion finalised upon inspection of medical and histopathology records. Baseline clinicopathological characteristics and survival data were collected by retrospective review of medical records by persons independent of analyses performed.

2.2.2 Histological review and FFPE tissue sampling

Each FFPE block retrieved underwent histologic assessment though review of haematoxylin and eosin (H&E) stained sections. Tumour blocks were sectioned (20 um)

and where identified by review as < 75% tumour-containing, were macrodissected to enrich for tumour content. Liposarcoma (LPS) samples were assessed histologically for well-differentiated (WD) and de-differentiated (DD) areas, and macrodissected to enrich for DD histology. Samples with insufficient material were excluded from downstream processing and subsequent analyses.

2.3 Mass spectrometry proteomics

2.3.1 Protein extraction and digestion

Each tumour sample was deparaffinised with 3 xylene washes, rehydrated twice in a decreasing ethanol gradient (100%, 96%, 70%), and dried in a SpeedVac concentrator (Thermo Scientific, Waltham, MA, USA). Lysis buffer (0.1 M Tris-HCL pH 8.8, 0.5% (w/v) sodium deoxycholate, 0.35% (w/v) sodium lauryl sulphate) was added at 200 ul/mg of dried tissue, samples homogenised by 3x 30 s pulses with a LabGen700 blender (ColePalmer, Vernon Hills, IL, USA), sonicated on ice for 10 min, and heated to 95 ºC for 1 h to reverse formalin crosslinks. Lysis was performed for 2 h by shaking at 750 rpm at 80 ºC. Samples were centrifuged at 14,000 x g at 4 ºC for 15 min, the supernatant retained, and protein concentration measured by bicinchoninic acid (BCA) assay (Thermo Scientific Pierce, Waltham, MA, USA). Tissue extracts were digested by Filter-Aided Sample Preparation (FASP), as previously described[496]. Briefly, samples were concentrated in Amicon-Ultra 4 centrifugal filter units (Merck Group, Darmstadt, Germany), and detergents removed by washing with 8 M urea. Samples were transferred to Amicon-Ulta 0.5 filters (Merck Group, Darmstadt, Germany), reduced with 10 mM dithiothreitol (DTT) for 1 h at 56 ºC, and alkylated with 55 mM iodoacetamide (IAA) for 45 min at room temperature in the dark. Samples were washed with 100 mM ammonium bicarbonate (ABC) and digested with trypsin (Promega, Madison, WI, USA) at a ratio of 1:100 ug sample at 37 ºC overnight. Peptides were collected by three centrifugations at 14,000 xg with 100 mM ABC, desalted using SepPak C18 Plus cartridges (Waters, Milford, MA, USA), and dried in a SpeedVac concentrator (Thermo Fisher Scientific, Waltham, MA, USA).

2.3.2 Tandem-Mass-Tag labelling

Tumour sample peptides and a pooled reference sample containing representative LMS, DDLPS, UPS, and SS material were labelled with TMT 11-Plex reagents (Thermo Scientific, Waltham, MA, USA) as per manufacturer's guidelines. Briefly, dried peptides were labelled as per manufacturer's guidelines. For the 11th (131C) channel, a pooled reference containing lysates from LMS, DDLPS, UPS and SS cases was used in all MS

experiments. Samples were incubated with respective TMT labels for 1 h at room temperature, and the reaction quenched with 5% hydroxylamine. Labelled peptides were pooled, dried in a SpeedVac concentrator, and desalted with SepPak C18 Plus cartridges as before.

2.3.3 High-pH reversed-phase fractionation

All samples were fractionated off-line by Dionex UltiMate3000 HPLC system (Thermo Fisher Scientific, Waltham, MA, USA). Each sample was dissolved in 100 µl of solvent A (0.1% NH_4OH in water), sonicated for 5 minutes and centrifuged at 15,000 × g for 2 min. Supernatant was loaded onto a 2.1 × 150 mm, 5 µm Waters (Milford, MA, USA) XBridge C18 column (5µm particles) at a flowrate of 200 µl/min and peptides were separated using gradient of 5-40% of solvent B (0.1% NH_4OH in acetonitrile) for 30 min followed by 40-80% of solvent B in 5 min and held at 80% for additional 5 min. Overall 90 fractions (30 s per fraction) were collected by automatic fraction collector into a 96 well-plate and combined into 10 fractions with a stepwise concatenation strategy. Pooled fractions were dried in SpeedVac concentrator.

2.3.4 Liquid chromatography and mass spectrometry

The LC/MS analysis was performed on a Dionex UltiMate3000 HPLC coupled with the Orbitrap Fusion Lumos Mass Spectrometer (Thermo Scientific, Waltham, MA, USA). Each peptide fraction was dissolved in 40 µL of 0.1% formic acid and 10 µL were loaded to the Acclaim PepMap 100, 100 µm × 2 cm C18, 5 µm, trapping column (Thermo Fisher Scientific, Waltham, MA, USA) with a flow rate 10 µL/min. Peptides were then separated with the EASY-Spray C18 capillary column (75 µm × 50 cm, 2 µm) at 45 °C. Mobile phase A was 0.1% formic acid and mobile phase B was 80% acetonitrile, 0.1% formic acid. The gradient method at flow rate of 300 nL/min included the following steps: for 120 min gradient from 5% to 38% B, for 10 min up to 95% B, for 5 min isocratic at 95% B, re-equilibration to 5% B in 5 min, for 10 min isocratic at 5% B. The precursor ions were selected at 120k mass resolution, with automatic gain control 4×10^5 and ion trap for 50 ms for collision induced dissociation (CID) fragmentation with isolation width 0.7 Th and collision energy at 35% in the top speed mode (3sec). Quantification spectra were obtained at the MS3 level with higher-energy C-trap dissociation (HCD) fragmentation of the top 5 most abundant CID fragments isolated with Synchronous Precursor Selection (SPS) with quadrupole isolation width 0.7 Th, collision energy 65% and 50k resolution. Targeted precursors were dynamically excluded for further isolation and activation for 45 seconds.

2.3.5 MS data processing

The SequestHT search engine in Proteome Discoverer 2.2 or 2.3 (Thermo Scientific, Waltham, MA, USA) was used to search the raw mass spectra against reviewed UniProt human protein entries (v2018_07 or later) for protein identification and quantification. The precursor mass tolerance was set at 20 ppm and the fragment ion mass tolerance was 0.02 Da. Spectra were searched for fully tryptic peptides with maximum 2 missed cleavages. TMT6plex at N-terminus/lysine and Carbamidomethyl at cysteine were selected as static modifications. Dynamic modifications were oxidation of methionine and deamidation of asparagine/glutamine. Peptide confidence was estimated with the Percolator node. Peptide False Discovery Rate (FDR) was set at 0.01 and validation was based on q-value and decoy database search. The reporter ion quantifier node included an integration window tolerance of 15 ppm and integration method based on the most confident centroid peak at the MS3 level. Only unique peptides were used for quantification, considering protein groups for peptide uniqueness. Peptides with average reporter signal-to-noise > 3 were used for protein quantification. Proteins with an FDR < 0.01 and a minimum of two peptides were used for downstream analyses.

2.3.6 Proteomic data imputation and normalisation

All data was processed using custom R scripts in R v3.5.1 or later [497]. Sample data with > 5% missing values relative to associated reference samples were deemed low quality and excluded from analyses (**section 3.2.3.1**). Proteins identified in < 75% of samples were removed, and those remaining imputed using the k-nearest neighbour (k-NN) algorithm in the impute R package [498]. Data was normalised and batch effects removed in a multi-step procedure. Firstly, each sample was divided by the corresponding reference sample, data was then $\log_2$ transformed, median centred across samples, and standardised within samples. For subtype-specific analyses, data was filtered for samples of interest, and protein filtering, imputation, and normalisation performed as before.

2.4 NanoString targeted transcriptomics

2.4.1 RNA extraction

Tumour total RNA was extracted using the All Prep DNA/RNA FFPE kit (Qiagen, Hilden, Germany) following vendor's standard protocol. mRNA concentrations were measured using Qubit fluorometric quantitation (Thermo Fisher Scientific, Waltham, MA, USA). RNA Integrity Number and percentage of total RNA < 200bp in size was measured using

2100 Bioanalzyer system (Agilent, CA, USA). RNA samples were stored at -80C until use in downstream analyses.

2.4.2 Nanostring data processing and analysis

Tumour total RNA was extracted using the All Prep DNA/RNA FFPE kit (Qiagen, Hilden, Germany) following vendor's standard protocol. mRNA concentrations were measured using Qubit fluorometric quantitation (Thermo Fisher Scientific, Waltham, MA, USA). RNA Integrity Number was measured using 2100 Bioanalzyer system (Agilent, CA, USA). RNA samples were stored at -80°C until use. Targeted gene expression profiling was performed using a custom panel of 21 immune-related genes and 3 housekeeper genes with the nCounter PlexSet-96 platform (NanoString Technologies, Seattle, WA, USA; **Table 2.1**). The gene panel was chosen as part of a previous project involving the profiling of STS tumours. It was constructed to select gene analogues of the proteins commonly examined by IHC, genes whose expression indicate T cell function, and genes with a stimulatory or inhibitory immune checkpoint function. Total RNA of 150-450 ng (variable to account for RNA degradation) of tumour samples and calibration samples was input for hybridisation and analysis performed per manufacturer's instructions using the nCounter Max system (NanoString Technologies, Seattle, WA, USA). The expression values of calibration samples were used to adjust for differences between PlexSet plates (i.e., technical variance). The calibrated raw expression data was then normalised using the NanoStringNorm R package by 'CodeCount' = 'geo.mean', 'Background' = 'mean', and 'SampleContent' = 'housekeeping.geo.mean'. Additionally, values < 1 were set to 1, data $\log_2$ transformed and gene-level median centring performed.

2.5 Analysis of The Cancer Genome Atlas data

2.5.1 Reversed-phase protein microarray

The level 4 ($\log_2$ transformed with loading and batch corrected) RPPA dataset from the TCGA sarcoma (TCGA-SARC) study[36] was downloaded from The Cancer Proteome Atlas portal (https://tcpaportal.org/tcpa/) and clinical data downloaded from the TCGA Pan-cancer Clinical Data Resource (TCGA-CDR) within the NCI Genomic Data Commons (https://gdc.cancer.gov/about-data/publications/PanCan-Clinical-2018). The RPPA dataset was restricted to LMS, DDLPS, UPS, and SS cases and feature level (protein) median centred across samples.

2.5.2 RNA sequencing

RNA sequencing (RNAseq) data (fragments per kilobase of exon per million mapped fragments (FPKM)) and corresponding clinical data from the TCGA-SARC study was

Table 2.1 Custom NanoString immune panel

Gene Name	Group
LAMP3	Immune cell markers
CD4	Immune cell markers
KIR3DL1	Immune cell markers
CD68	Immune cell markers
FOXP3	Immune cell markers
CD163	Immune cell markers
NCAM1	Immune cell markers
CD3G	Immune cell markers
CD8A	Immune cell markers
TNFRSF9	Immune checkpoint proteins
CD274	Immune checkpoint proteins
CTLA4	Immune checkpoint proteins
LAAG3	Immune checkpoint proteins
IDO1	Immune checkpoint proteins
PDCD1LG2	Immune checkpoint proteins
PDCD1LG2	Immune checkpoint proteins
STAT6	Other
HLA-A	Other
PRF1	Other
TBX21	Other
VTCN1	Other

downloaded from the TCGA-CDR within the NCI Genomic Data Commons (https://gdc.cancer.gov/about-data/publications/PanCan-Clinical-2018). Samples were restricted those analysed within the TCGA-SARC publication and further restricted to LMS for LMS-specific analyses or LMS, DDLPS, UPS and SS for WGCNA analyses[36]. FPKM data was converted to transcripts per million (TPM) and genes present in < 75% of samples were removed. A value of 1 was added to all measures to address missing values, data was $\log_2$ transformed, median centred across features (genes) and across samples. The TCGA clinical outcome data was censored at 5-years post-surgery.

2.6 Immunohistochemistry

Tissue microarrays (TMA) containing 63 LMS, 50 UPS and 32 DDLPS with at least 2 replicate cores were used for IHC. Consecutive 4µm TMA sections were stained for H&E, CD3, CD4, and CD8 using the DAKO link automated stainer (Agilent, CA, USA). Sections were deparaffinised by xylene and rehydrated by graded ethanol. Antigen retrieval was performed using DAKO FlexEnvision kit (K8002; Agilent, CA, USA) by either pressure cooking in citrate (pH6) for 2 min (CD3) or incubating with pH9 pre-treatment module (PTM) buffer (Agilent, CA, USA) for 20 min at 97 °C (CD4 and CD8). Incubation with primary antibody (CD3 DAKO M0452 at 1:600 dilution; CD4 DAKO 4B12 at 1:80 dilution; CD8 DAKO C8/144B at 1:100 dilution) was for 60 minutes at room temperature. Secondary antibody staining and visualisation was performed using DAKO FlexEnvision (Mouse) Kit, followed by application of DAB and haematoxylin counterstaining. H&E slides were assessed to confirm viable tumour content, and CD3/4/8+ TIL stains counted under direct brightfield microscopy at x400 magnification. For cores with section preservation of 50-100%, cell counts were corrected to 100% area. Data from cases where section preservation was < 50% were excluded. Replicate scores were averaged then multiplied by 1.274 to produce average CD3+, CD4+ or CD8+ TIL/mm^2. Digital microscopy images for all stained TMA sections were captured at x40 resolution using Nanozoomer-XR (Hamamatsu Photonics, Japan).

2.7 Bioinformatics and statistical methods

Unless otherwise specified, all data was analysed using custom R scripts in R v3.5.1 or later [497].

2.7.1 Differential expression analysis

Differentially expressed proteins (DEP) were identified by significance analysis of microarrays (SAM) using samr in R[499]. Normalised and imputed datasets were run using the SAM multiclass test, with 100 permutations. For paired comparisons, two sample unpaired tests (Student's T test statistic) were performed. In each test, delta was selected as the value at which median FDR < 0.01.

2.7.2 Proteomic database representation

The immune component was assessed using the ImmPort database[500]. The matrisome component was assessed using the MatrisomeDB database[501]. The adhesome component was assessed using the consensus integrin adhesome work of Winograd-Katz *et al*[502], the kinome was assessed using the work of Manning *et al*[503].

2.7.3 Overrepresentation analysis, Gene Set Enrichment Analysis and single sample Gene Set Enrichment Analysis

Overrepresentation analysis (OA) and Gene Set Enrichedment Analysis (GSEA) were performed with ClusterProfiler in R using the gene ontology (GO) biological process (BP) and hallmark gene sets with between 9 and 501 genes[504–508]. Proteins were ordered by Log_2 fold change, and for OA were filtered to those identified as uniquely upregulated in histological subtype by differential expression analysis. Single sample GSEA (ssGSEA) was performed using ssGSEA (v10.0.11) on the GenePattern public server[504,509]. Rank normalisation and a weighting exponent of 0.75 were used to assess enrichment of the Hallmark, GO BP, Kyoto encyclopaedia of genes and genomes (KEGG), and Drug Signature database (DSigDB) v1.0 D1 gene sets containing at least 10 genes, and normalised enrichment scores were median centred across gene sets[506–508,510,511]. All gene sets except DSigDB were downloaded from the Molecular Signatures Database (MSigDB) v7.5.1[512]. A background of proteins within the proteomic dataset was used for all analyses.

2.7.4 Clustering

Hierarchical clustering and dimension reduction by principal component analysis (PCA), t stochastic neighbour embedding (tSNE), and uniform manifold approximation and projection (UMAP) were used[513–516]. For hierarchical clustering, a distance measure of Pearson correlation with average linkage was used, unless otherwise specified in the figure legend. For tSNE, perplexity was optimised by running analyses for a range of values (minimum 5 per analysis) and inspecting stability of the results. For UMAP, the same optimisation was performed, but addressed at the number of neighbours used. For all other tSNE and UMAP parameters, the default settings within their respective R packages were used[514,516]. For robust proteome clustering, unsupervised consensus clustering (CC) was performed using ConsensusClusterPlus in R[517]. CC was performed by agglomerative hierarchical clustering using Spearman's rank with average linkage. Protein and item (sample) resampling was set at 80% and CC was run for 1000 iterations for up to 10 clusters (k). Optimal k was determined through inspection of consensus matrices, the cluster tracking plot, the consensus cumulative distribution function (CDF) plot, and the Δ area plot, and by calculating sample silhouette scores in CancerSubtypes[518,519]. Clusters were confirmed as statistically significantly different by SigClust with hard thresholding and 1000 sample simulations ($p < 0.05$)[520].

2.7.5 Weighted gene correlation network analysis

Weighted gene correlation network analysis (WGCNA) was performed using the R WGCNA package[521]. Normalised proteomic data was used to construct a co-expression network. Network type was specified as signed hybrid and constructed with an optimal soft threshold value (β) of 5, determined by graphical inspection of network scale free topology and mean connectivity across a range of β values. Average linkage hierarchical clustering with dynamic cutting was used to identify modules of ≥ 30 proteins, and 1 - Pearson correlation cut height ≥ 0.25.

2.7.6 Protein-protein interaction network analysis

All protein-protein interaction networks were built in Cytoscape v3.9.1[8]. To assess the complement and coagulation cascades, WikiPathway WP558 (63 nodes) was imported, adapted to include the C5 axis, and layout manually applied[522]. To visualise to the WGCNA-identified landscape, a protein co-occurrence matrix was used, with co-occurrence scores between pairs restricted to > 0.05 and an edge-weighted spring embedded layout used. All network measures (degree, betweenness centrality, and closeness centrality) were calculated in Cytoscape.

2.7.7 Survival analyses

Survival analyses assessed 3 clinical outcome measures: 1) local recurrence free survival (LRFS) defined as time from primary disease surgery to radiologically confirmed local recurrence or death, 2) metastasis free survival (MFS) defined as time from primary disease surgery to radiologically confirmed metastatic disease or death, 3) overall survival (OS) defined as time from primary disease surgery to death from any cause. Data was censored at 5 years, and patients who had not experienced a survival event were censored at last follow-up. Kaplan Meier curves were used to visualise clinical outcome over time, and Cox proportional hazard regression was implemented for univariable and multivariable statistical analysis. To maximise statistical power, the group with the large n was selected as the reference for all Cox regression analyses. In all models, the Cox proportional hazard (PH) and linearity assumptions were assessed. Any minor or severe violation of this assumptions was interrogated by use of the Schoenfeld residuals and associated Schoenfeld test, deviance residuals, and martingale residuals. Where necessary, variable categories were grouped and/or transformed, the details of which are reported in the results chapters of this thesis.

2.8 Statistics and reproducibility

No statistical method was used to predetermine cohort size; cases were included based on material and data availability. All statistical tests were two-sided and where required, p values were adjusted to false discovery rate (FDR) using the Benjamini-Hochberg procedure to account for multiple comparisons[65]. Where appropriate, distribution of the data was assessed using Shapiro-Wilk tests for normality, and tests not assuming a normal distribution implemented if $p < 0.05$. Kruskal-Wallis one-way analysis of variance (ANOVA) tests, Dunn's tests, one-way ANOVA tests, Tukey's honestly significant difference (HSD) tests, and chi-squared tests of independence were implemented. Further details of specific statistical tests are listed in figure legends.

2.9 Data Availability

Raw proteomic data are deposited on the ProteomeXchange Consortium via the PRIDE partner repository (dataset identifier PXD036226)[66,67]. At the time of writing, data is password protected whilst this work is under review for publication.

Chapter 3 Profiling the soft tissue sarcoma proteome

3.1 Background and objectives

At present, the STS proteome has not been comprehensively profiled. This represents a gap in the understanding of STS biology, which may hold clinically actionable insights. Standard of care for primary STS is surgical resection, therefore tissue archives surplus to diagnostic requirements contain a rich resource of STS tumour material, which this study seeks to utilise[44]. Tissue is stored as FFPE to increase sample longevity and to enable stability of specimens at room temperature; however, tissue fixation with formalin introduces crosslinks in biomolecules such a proteins. To facilitate protein digestion and analysis, these crosslinks must be reversed[480–482]. Experimental methods have been developed to successfully reverse crosslinking. However the extensive processing required to achieve this means that yields from FFPE tissues are significantly lower than those from the fresh-frozen material utilised by other studies (e.g., CPTAC). Moreover, the input material required for TMT-MS proteomics with offline fractionation far exceeds the amounts required for comparable genomic and transcriptomic profiling. For example, in-house standard operating procedures for our laboratory indicate a minimum requirement of 120 um for protein extraction for TMT analysis, compared to 80 um for combined RNA and DNA extraction. This presents a challenge for rare disease profiling where sample material is scarce.

Our in-house protocol development has established a pipeline for TMT-MS analysis of STS FFPE tissue. However, the cohort herein introduces new challenges. Firstly, this cohort includes samples with lower tumour content, such as heavily pre-treated samples and specimens small in size. Furthermore, histological subtypes of STS that have not previously been profiled are also included. This may translate to differences in the feasibility of sample processing and achievable yields. Secondly, this study underwent a cohort expansion part-way through. As part of the initial phase a pooled reference sample containing peptides from representative tumours was generated. The reference sample was created to monitor and address variation between MS runs. However, it did not include subtypes profiled following cohort expansion, and was therefore not representative of all STS histologies profiled. This may result in poor proteome coverage of some STS subtypes. And finally, this cohort is large and therefore requires multiple TMT sets to be run. This increases the risk of differences arising between runs (i.e., batch effects), and due to stochastic sampling of peptides is recognised to introduce high

amounts of missing data between the TMT sets[466]. There is no formally established pipeline for quality control, and normalisation of TMT data. As a result, appropriate methods for these steps are tailored to individual experiments.

Accordingly, the objectives of this chapter are:

1) To implement the in-house FFPE peptide extraction workflow in a heterogeneous STS cohort
2) To identify appropriate data processing pipelines for large-scale multi-batch STS proteomics

3.2 Results

3.2.1 Patient selection

Patients were selected for inclusion based on the following criteria: 1) histopathologically confirmed diagnosis of AS, ASPS, CCS, DDLPS, DES, DSRCT, EPS, LMS, RT, SS, or UPS, 2) ≥ 18 years of age at the time of sample collection (excluding RT), 3) FFPE surgical resection tumour material available in quantities sufficient for analyses. Where possible, patients were restricted to those where surgery was performed prior to 2014, to ensure sufficient follow up. For rarer subtypes and to facilitate cohort expansion, patients receiving surgery up to 2018 were considered for inclusion on a case-by-case basis. Patients were excluded if the primary tumour specimen was FNCLCC grade 1. RT samples and 2 AS samples were obtained externally. All other samples were retrieved through RMH. Eligible patients were identified by retrospective search of hospital databases, and inclusion was finalised upon inspection of medical and histopathology records. Further details are provided in **section 2.2**.

3.2.2 Peptide extraction from formalin-fixed paraffin-embedded tissue

The implemented peptide extraction protocol (**Figure 3.1A**) was assessed previously in ourlab for the profiling of LMS, UPS, DDLPS, and SS tumours[483]. Briefly, samples were histologically reviewed by H&E to identify viable tumour content. For samples with ≥ 75% tumour area, sections were taken. Where samples contained < 75% tumour content, macrodissection was performed prior to sectioning to enrich for tumour content. This method of enriched sectioning was performed to reduce tumour heterogeneity and ensure tumour cell features were dominant in the final profiling data. Estimation of tumour content and sampling of WD/DDLPS tumours was based on the DD region only. Sections were pooled, deparaffinised, crosslinks reversed by heating, and proteins extracted by

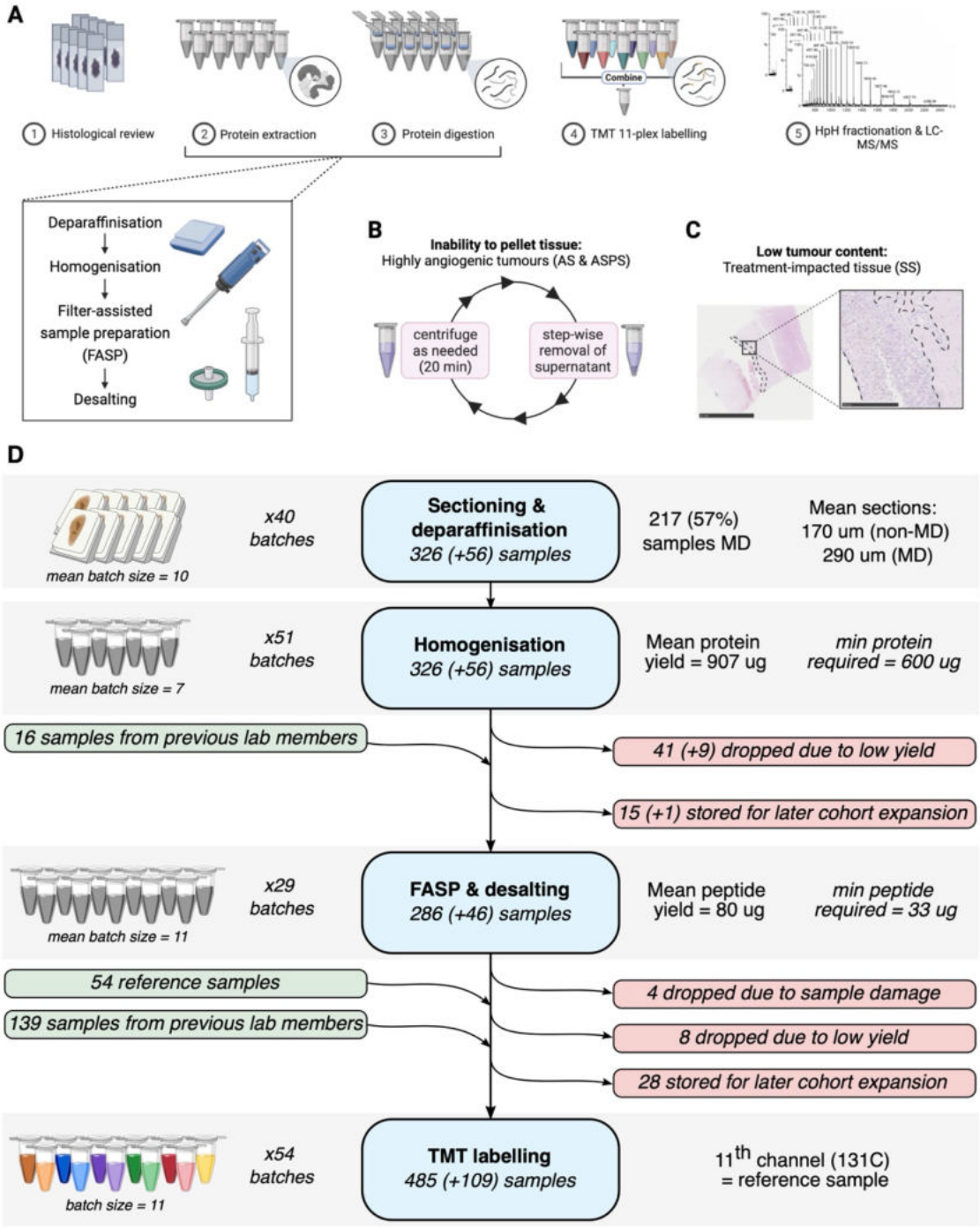


Figure 3.1 Implementation of the STS proteome profiling pipeline
(A) Workflow overview. **B)** Modified deparaffinisation for angiosarcoma (AS) and alveolar soft part sarcoma (ASPS) samples. **C)** Modified sectioning for samples with low viable tumour content such as synovial sarcoma (SS). Scanned image of haematoxylin and eosin (H&E)-stained sample section. Dotted line indicates viable tumour cell area. Scale bars = 10 mm (left), 0.75 mm (right). **D)** Cohort overview through protein extraction, protein digestion and peptide labelling steps. Numbers in brackets indicate replicate samples. Abbreviations: TMT = tandem mass tag; HpH = high pH; LC = liquid chromatography; MS = mass spectrometry; MD = macrodissected.

homogenisation. Protein yields were measured by bicinchoninic acid (BCA) assay. Where yields were < 600 ug and sufficient tumour material remained, the sample was re-extracted. A filter-assisted sample preparation (FASP) protocol was implemented to remove the detergents used during extraction and to digest proteins into peptides[496]. Peptides were desalted and yields were measured by BCA assay. Where yields were < 33 ug and sufficient tumour material remained, the sample was re-extracted. Peptides were then labelled with TMT 11-plex labels in batches of 10 unique tumour samples, and 1 pooled reference sample. Labelled samples were pooled, fractionated by high pH and injected into an LC-coupled MS (full description of methods are described in **section 2.3**. Throughout processing, samples were handled in batches of mixed subtypes. This minimised the impact histological subtype had on inter-batch variation and therefore aided identification of batch-specific effects.

Initial attempts made to profile other subtypes beyond LMS, UPS, DDLPS and SS identified several pitfalls. Firstly, AS and ASPS tumours failed to pellet during the centrifugation steps as part of sample deparaffinisation. The sequential washes required for deparaffinisation therefore resulted in significant losses of tumour material. To reduce tumour loss, the centrifugation time was increased from 3 to 20 minutes and the supernatant removed in a stepwise manner at each wash. If the sample re-suspended during removal of the supernatant, extra centrifugations were performed as needed (**Figure 3.1B**). Whilst the biochemical and biophysical reasons for this difference between subtypes were not formally assessed, it is hypothesised that the high vascular content of AS and ASPS samples contributed to handling difficulties. Secondly, due to the current standard of care for SS patients, a high proportion (58%) of the SS cohort received pre-operative treatment. Treatment effects were evident in many SS tumours upon inspection of H&E sections. These included nuclear and cytoplasmic enlargement and hyalinization (**Figure 3.1C**), as well as cellular necrosis. This reduced the viable tumour area for sampling. For large SS samples with treatment effects located to isolated regions within the tumour, macrodissection was performed and the impact of a reduced viable tumour area counteracted by increasing the sectioning depth. Yet, many SS samples could not be salvaged and thus were excluded from processing and analysis. This biased the SS cohort. Interpretation of downstream SS analyses must consider that this cohort likely underrepresents pre-treated SS patients, and that the full spectrum of the SS patient population is not captured herein. Finally, to ensure enrichment of tumour content in the tissue sample, each section was required to have a tumour cell content ≥ 75%. For comparison, CPTAC proteomic pipelines deem 50 - 80% tumour cell content as acceptable, dependent on the malignancy[523–529]. As with treatment-impacted SS

samples, samples with < 75% tumour content were macrodissected to enrich for tumour cells, and the sectioning depth increased relative to the percentage tumour area sampled. Overall, these modifications to the protocol limited the numbers of repeat extractions required due to insufficient protein or peptide yields. By extension, this limited technical processing variability, and the risk of contamination and human error during peptide extraction.

An overview of the samples extracted, processed, and analysed as part of this study is shown in **Figure 3.1D**. In total, 382 samples were sectioned and deparaffinised across 40 batches. This includes 56 samples of repeat extractions due to low yields during the initial extraction round. On average, 170 um of tissue was sampled from cases with ≥ 75% tumour cell content. Just over half of cases contained < 75% tumour cell content and required macrodissection. For those macrodissected, an average of 290 um tissue was processed. These tissue requirements are reflective of handling surgical resection specimens. The amount of material required for biopsies would be vastly greater. All deparaffinised samples progressed to homogenisation, which was performed over 51 batches. The target protein yield after homogenisation was 600 ug. The achieved average yield, including samples where multiple extractions were merged, surpassed this at 907 ug. Following homogenisation, a total of 66 samples were excluded. This included 50 with low protein yields. Of these, 43 (86%) required macrodissection, indicating the low yields may be resultant of a reduced sampling area. Indeed, within the macrodissected samples, the viable tumour cell area was lower in those with proteins yields < 600 ug compared to those with yields ≥ 600 ug (36.5% vs 44.7%) despite being adjusted for by sectioning depth. This may suggest the implemented protocol performs well for samples of large tumour areas, but that it is challenging to achieve sufficient protein yields when the sampling area is reduced. Of the 7 excluded samples that did not require macrodissection, 5 (71%) were either AS or ASPS samples, reiterating the difficulty in handling these subtypes. The remaining 16 excluded samples (9 DES and 7 LMS) where sufficient protein was extracted were stored for future studies. The resultant 332 samples (including 16 sample processed by previous lab members) underwent FASP and desalting over 29 batches. Input material of 600 ug protein was used for FASP and resulted in an average of 80 ug peptide. For TMT analysis, 33 ug is required; thus, yields of approximately 80 ug permitted samples to be run in duplicate if necessary and left allowances for sample loss/degradation during lyophilisation and freezing. Following desalting, 4 samples were excluded due to damage during handling, 8 had low protein yields, and 28 were stored for future studies. The 28 were comprised of 9 AS (8 recurrence, 1 primary), 8 primary uLMS, 7 DDLPS recurrences, 2 primary CCS, and 2

primary DES. After desalting, 594 samples (including 139 samples prepared by previous lab members and 109 replicates) were labelled with TMT 11-plex labels in 54 batches to be run by MS.

3.2.3 Proteomic data processing

3.2.3.1 Quality control & data exclusion

Proteomic quality control (QC) is an important step in ensuring low quality data is identified and handled appropriately to maintain the robustness of downstream analyses. TMT MS provides a measure of relative abundance, dependent on the assumption that the same amount of labelled peptide from each sample was injected into the mass spectrometer and analysed. The proteins in a sample are assumed to follow a unimodal (normal or gaussian) distribution, with few highly and lowly expressed proteins and many proteins expressed at intermediate abundances. A violation of one or both of these assumptions is indicative of low quality data. If not adjusted or excluded, low quality data can introduce significant skew to a MS dataset. Low quality data in TMT MS experiments can be the result of sample processing errors prior to MS analysis, inefficient TMT labelling, or technical problems with the LC or mass spectrometer. There is no consensus definition of low quality TMT MS data, however it is considered to possess few protein identifications, low protein abundances, high protein-level missing values (MVs), and a bimodal distribution. Notably, most of these measures are relative to each experiment and must be considered in the context of other MS data. The identification and removal of 'extreme outlier' data in this way is central to published MS QC approaches[530].

In this project, QC was performed at the point of data collection for each TMT set. **Figure 3.2A-B** and **Table 3.1** show QC of TMT set 32 as an example. The protein expression profile of each sample was assumed to follow a unimodal distribution. Density plots (**Figure 3.2A**) reveal that while most samples appeared to satisfy this assumption, sample A showed bimodality and significant skew towards low abundance proteins. Distributions were characterised using the bimodality coefficient (BC) and Hartigan's Dip Statistic (HDS), where a bimodal distribution is indicated by BC values > 0.555, a high HDS statistic and a HDS p value < 0.05[531–533]. These methods showed discordance in this dataset; use of BC classified 9/11 samples as bimodal, whilst use of HDS classified no samples as bimodal (**Table 3.1**). In fact, across all TMT sets collected, BC and HDS classified 84% and 0% of samples as bimodal respectively. It was therefore not possible

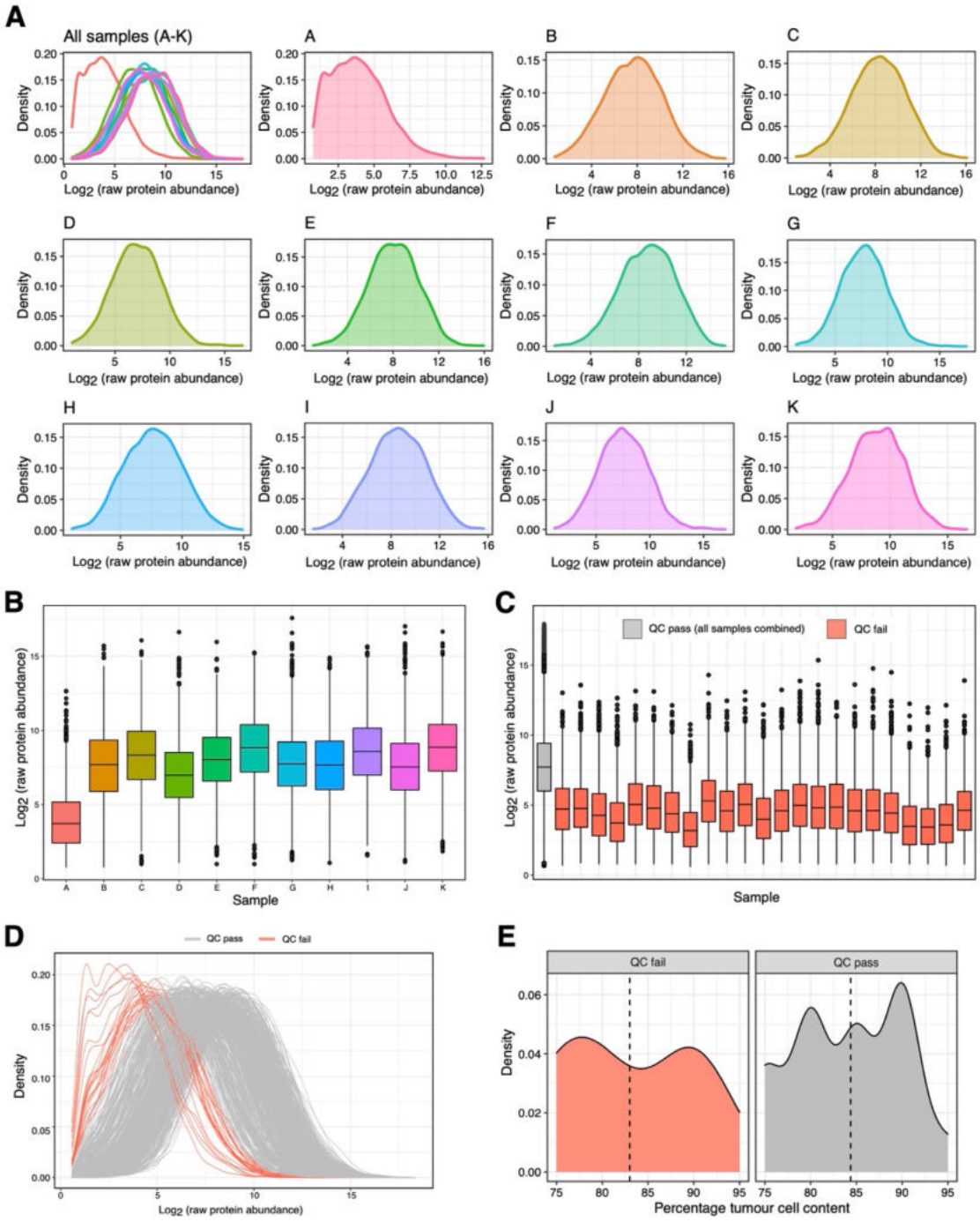


Figure 3.2 Quality control of tandem mass tag (TMT) data.
(A-B) Example TMT set, **C-E)** all data. **A)** Density plots showing protein abundance distribution within each sample. Top left shows all samples overlayed. **B)** Boxplots showing protein abundance across each sample. Boxes indicate 25th, 50th, and 75th percentile, with whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. **C)** Boxplots showing protein abundance across samples that passed QC (combined; grey), and sample that failed QC (individual; red). Boxes drawn as in **B**. **D)** Density plot showing protein abundance distribution for all samples. Samples that passed QC in grey, samples that failed QC in red. **E)** Density plots showing percentage tumour cell content for samples that fail and pass QC. Dashed line indicates mean value.

to use empirical measures of bimodality to objectively identify low quality data. BC and HDS did however highlight sample A as an outlier. Sample A showed the highest BC and HDS statistic, and therefore had the distribution furthest from a perfect unimodal (**Table 3.1**). Sample A also showed a low mean protein abundance (51.1) compared to other samples (435.1 – 1413), fewer protein IDs (3251) compared to other samples (4515 – 4548), and a high percentage of MVs relative to the total number of IDs in set 32 (28.5%; **Figure 3.2B and Table 3.1**). This revealed sample A as an extreme outlier within TMT set 32. Sample A was designated of low quality and excluded from analysis. Repeating this set-based inspection of data for each TMT set revealed the percentage of MVs to be a suitable and innate QC filter, consistently reflective of low protein abundance, few protein IDs, high MVs, and non-unimodal distribution. Maximal acceptable percentage of MVs was set as 5%, resulting in the exclusion of 23 samples from analysis. All 23 samples showed low protein abundances and unexpected distributions (**Figure 3.2C-D**). Whilst specimens were macrodissected to ensure a minimum of 75% tumour cell content was profiled, tumour cell content within the sampling area (i.e., sample purity) ranged from 75% to 100%. The potential impact of sample purity on data quality was therefore assessed. Interestingly, there was no apparent association. Samples that failed or passed QC showed similar average tumour cell contents of 83% and 84% respectively (**Figure 3.2E**). One caveat of this analysis is

Table 3.1 Data metrics collected for each TMT sample.
From left to right: number of proteins IDs, percentage missing values (MV %) relative to all proteins identified within the set, protein abundance mean and standard deviation (SD), bimodality coefficient (BC), Hartigan's Dip Statistic (HSD), and HSD p value.

Sample	Protein IDs		Protein abundance		Data distribution		
	n	MVs (%)	Mean	SD	BC	HDS	HDS p
A	3251	28.5	51.1	235.3	0.788	0.006	0.577
B	4515	0.7	779.4	2166.9	0.616	0.003	1.000
C	4533	0.3	1067.0	2605.2	0.573	0.002	1.000
D	4530	0.4	435.1	1971.8	0.710	0.002	1.000
E	4541	0.2	734.8	1697.1	0.510	0.003	0.995
F	4548	0.0	1247.5	2310.7	0.568	0.003	1.000
G	4543	0.1	772.4	4030.1	0.733	0.003	0.999
H	4537	0.2	685.9	1640.0	0.648	0.002	1.000
I	4544	0.1	1183.5	2782.7	0.658	0.002	1.000
J	4542	0.1	770.8	3620.2	0.713	0.003	0.998
K	4546	0.0	1413.0	3405.8	0.545	0.003	1.000

that data on sample purity was only available for a subset of samples (RMH specimens processed by the candidate). Thus, sample purity measures may not be representative of the full cohort. In addition, the relationship between sample age and data quality was queried. Given the rarity of STS, samples over a large time period were collected to obtain a sufficiently sized cohort. Samples ranged from 3 years old (2018 sample extracted in 2021) to 25 years old (1995 sample extracted in 2020). FFPE storage provides good sample stability, yet some studies have noted DNA and RNA degradation to increase with block age[534,535]. Protein specific studies on FFPE have found age mainly impacted protein yield and did not significantly impact the quality of MS analysis, however reports are inconsistent[536,537]. Herein the age of the FFPE sample did not appear to impact sample quality. Median block year in both the QC fail and pass groups was 2011 and the 15 oldest samples from 1995 to 2002 all passed QC (**Figure 3.3**).

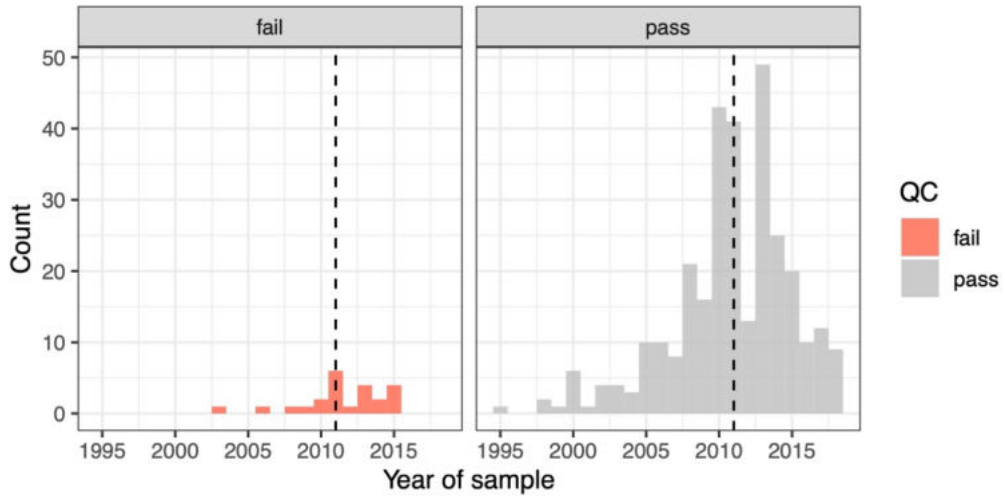


Figure 3.3 Sample age and quality control (QC).
Histograms showing the distribution of sample age (the year of surgery) in QC fail and pass groups. Dashed line indicates median.

In addition to the 23 samples that failed QC, unusable data from 126 other samples was also excluded. Of these, 110 (from 10 TMT sets, including 10 reference samples) were removed due to experimental errors in TMT labelling that resulted in a low labelling efficiency. 15 were removed as replicates, and 1 removed due to contamination. Usable data from 79 samples was also excluded but stored for future analyses following this study. The 79 samples profiled by TMT MS but not analysed herein, comprise 35 recurrences spanning AS, DDLPS, DES, EPS, LMS and SS, 21 primary samples from subtypes excluded (atypical teratoid rhabdoid tumour, myxofibrosarcoma, myxoid liposarcoma, spindle cell sarcoma, chondrosarcoma, and solitary fibrous tumour), 15

UPS samples obtained from external collaborators at UCL, 7 metastasis samples (EPS, LMS), and 2 samples from patients excluded after a cohort audit revealed they did not meet the study inclusion criteria. The final working dataset comprised data for 375 samples including 44 reference samples (i.e., spanning 44 TMT sets).

3.2.3.2 Performance of the reference sample

A reference sample is included in TMT MS analysis to permit multi-batching within studies. The reference is most often a pooled sample comprised of representative material from specimens within the study. It is included in each TMT set to assess inter-batch variations and adjust for batch effects. This study was initiated by previous members of the lab, and was initially designed to profile LMS, DDLPS, UPS, and SS only. It has since undergone an expansion phase to include multiple further histological subtypes into the cohort. The reference sample was designed at the commencement of the study, and thus was created based on the assumption that only LMS, DDLPS, UPS, and SS tumours would be profiled. As a result, the reference sample utilised does not span all subtypes profiled herein. To assess inter-batch variation and perform normalisation, it is vital that the reference sample has minimal MVs at the peptide and by extension protein level. If samples differ significantly from the reference sample, numerous proteins identified in samples will not be present in the reference, and *vice versa.* This is a major introducer of MVs in MS data, which if highly prevalent within a finalised dataset can restrict analysis and interpretation (as discussed in **section 1.6.1.2**). The impact of using a reference sample that doesn't include all subtypes profiled was therefore analysed. All references to MVs made in this chapter refer to protein-level data, not peptide-level data.

Analyses were performed repeatedly during data collection to ensure potential issues were flagged early, however for completeness, **Figure 3.4** presents all data collected. The reference sample was approximately equal part LMS, UPS, DDLPS, and SS (**Figure 3.4A**). Due to sample availability, there were differences in the number of unique samples included. SS had the fewest unique samples (n = 8), then DDLPS (n = 10), then UPS (n = 19), and LMS had the most (n = 30; **Figure 3.4B**). Therefore, representation of the full spectrum of disease heterogeneity of each histological subtypes likely varied. Most TMT sets (n = 29) contained a mix of subtypes that are found in the reference sample as well as not in the reference sample (referred to as 'mixed sets'), 14 sets contained only subtypes found in the reference sample (referred to as 'reference sets'), and 1 set did not contain any subtypes present in the reference sample (referred to as the 'non-reference set'; **Figure 3.4C**).

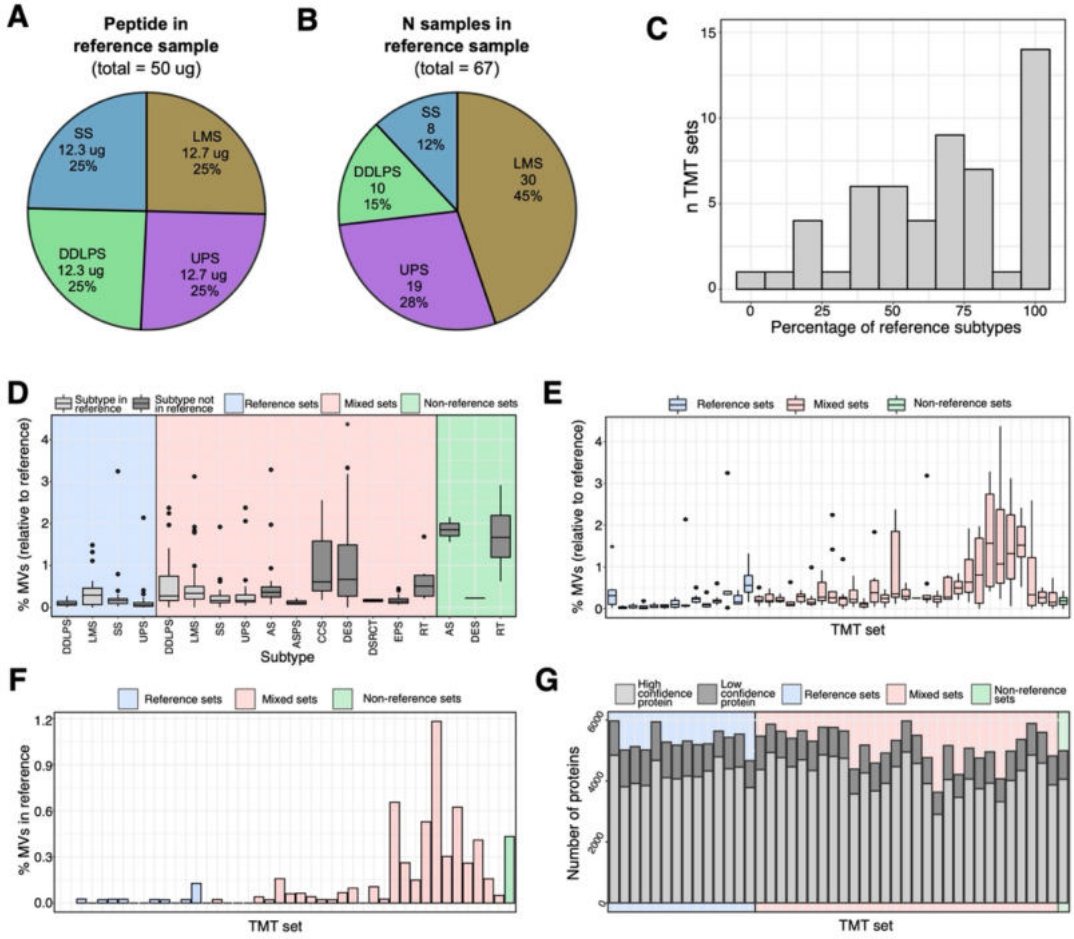


Figure 3.4 Reference sample (REF) composition and use in tandem mass tag (TMT) sets.
(A-B) Pie charts showing the amount of peptide (**A**) and number of unique samples (**B**) per subtype in the REF. **C)** Histogram showing the distribution of TMT sets based on the percentage of refence subtypes in the set**. D)** Boxplots showing the % missing values (MVs) in different subtypes within different TMT set types (indicated by background colour). Boxes indicate 25th, 50th, and 75th percentile, with whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. **E)** Boxplots showing the % MVs in the tumour samples within each TMT set, coloured by set type. Boxes drawn as in **D. F)** Bar plot showing the % MVs in the reference sample within each TMT set, coloured by set type. **G)** Bar plot showing the number of low (dark grey) and high (light grey) confidence protein identifications within each TMT set. Background colour indicates set type. Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DES = desmoid tumour; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; RT = rhabdoid tumour; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma

Across STS subtypes and TMT sets, the percentage of MVs relative to the reference sample varied from approximately 0 - 4 % (**Figure 3.4D-E**). I first focused on determining if disproportionate representation of subtype heterogeneity in the reference resulted different MV levels across subtypes. This revealed no discernible difference; in the reference sets DDLPS, LMS, SS and UPS consistently showed very low levels of MVs

(mean < 1%), and there was no apparent relationship between the number of unique samples in the reference sample and MVs (**Figure 3.4D**).

I then sought to assess the impact of profiling subtypes not included in the reference sample. Mixed sets showed slightly higher MVs compared to reference sets (mean = 0.55% vs 0.25%; **Figure 3.4E**). In addition, within mixed TMT sets, samples from subtypes not present in the reference showed marginally higher MVs on average than samples from subtypes present (mean = 0.54% vs 0.44%; **Figure 3.4D**). This observation however is not consistent. The higher average MVs in subtypes not in the reference sample appears to be driven by CCS and DES samples, which showed mean MVs in the mixed sets of 0.65% and 0.71% respectively. Whilst, other subtypes also not in the reference sample (AS, ASPS, DSRCT, EPS, RT) showed low levels of MVs. This suggests variations in MVs were not due to the lack of subtype inclusion within the reference sample. It may be the case that CCS and DES show the most distinct proteomes compared to the reference subtypes. Yet contradicting this, in the non-reference set DES showed very low MVs (0.23%); the caveat being that this was data from only a single sample in a single set. Further to assessing MVs within tumour samples relative to the reference, the MVs in the reference can also be used a measure of similarity between the reference and profiled specimens. The reference samples themselves in mixed and non-reference sets showed higher MVs relative to the other samples in the set, however this was minimal (mean MVs: reference sets = 0.02%, mixed sets = 0.18%, non-reference set = 0.43%; **Figure 3.4F**).

Reflective of the variations in MVs across sets, the number of proteins identified also differed, although only minor (**Figure 3.4G**). During MS data processing, protein identifications can be classified as low confidence or high confidence, with only high confidence proteins used for downstream analyses. Herein, high confidence proteins are defined as having at least 2 identifiable peptides present for quantification, and an identification FDR ≤ 0.01. The average number of high confidence IDs seen in mixed and non-reference sets (4224 and 4061 proteins respectively) was highly comparable to number of IDs in the reference sets (4240 proteins). Moreover, the proportion of high confidence protein IDs relative to total IDs was also similar across set types (mean percentage of high confidence IDs out of all IDs: reference sets = 80%, mixed sets = 81%, non-reference set = 82%). Therefore, the increased MVs observed was not due to low confidence in protein identification.

Overall, data did vary as a result of the inclusion of subtypes that were not in the reference composition. However, irrespective of set type, MVs were consistently very low in all tumour samples (mean < 2%) and all reference samples (< 1.2%). These levels of MVs are comparable to previous studies[466]. Furthermore, samples failing QC (**section 3.2.3.1**) were distributed throughout both reference sets and mixed sets, indicating high quality data to be obtainable irrespective of the set type. The reference was therefore deemed suitably representative for use but was monitored throughout this study as new subtypes were included and data collected.

3.2.3.3 Handling missing values

One appeal of TMT MS is the low percentage of protein MVs observed within a single TMT set. However, due to the stochastic nature of DDA MS, peptide sampling, and therefore protein identification, can differ extensively between sets[466]. As a result, multi-set TMT experiments often show high MVs (**section 1.6.1.2**)[466]. MS profiling in this study detected 8148 unique proteins across all 44 TMT sets. As anticipated, combining TMT sets introduced MVs between different sets (**Figure 3.5A-B**). MVs increased rapidly upon combination of the first 4 sets, resulting in 24% of the data being MVs, and 3029 proteins identified at a 100% detection rate (i.e., detected in all samples). Combining more than 4 sets increased MVs further, albeit at a much slower rate. It was not until combining all 44 sets, that another 24% increase in MVs was observed. Combination of all 44 sets resulted in just 1786 proteins identified at a 100% detection rate.

In DDA MS, as performed herein, precursor ions are selected in the MS1 scan for further fragmentation and subsequent identification. This selection, and therefore the overall peptide sampling of this procedure is based on abundance. As a result, proteins with MVs across TMT sets (i.e. those with <100% detection rate) were hypothesised to be of low abundance. Indeed, at low detection rate, more low abundance proteins are observed, the majority being in the lowest (1^{st}) expression quartile (**Figure 3.5C**). At ~75% detection there were no 1^{st} quartile proteins, and at ~90% detection there were no 1^{st} or 2^{nd} quartile proteins. By 100% detection, the vast majority of proteins retained were in the highest (4^{th}) expression quartile. To handle the MVs, a conservative approach would be to only assess proteins detected in 100% of samples. However, such a strict requirement would limit the study to 1786 proteins, and result in the exclusion of many useful data points. Therefore, to address the MVs, imputation was used. Imputation was

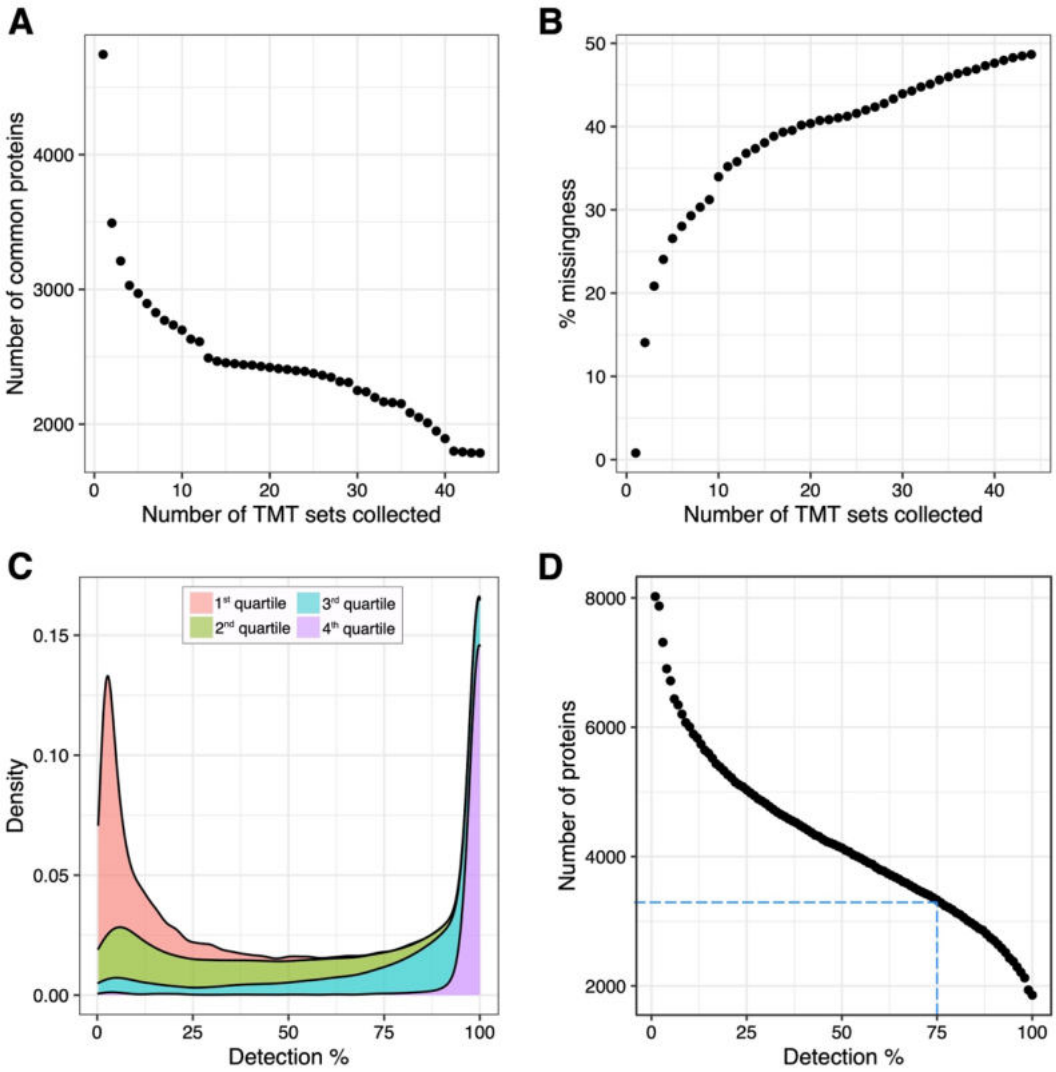


Figure 3.5 The impact of combining tandem mass tag (TMT) sets on protein identification and missing values (MVs) within data
A) The number of common proteins (proteins identified in all samples) as the number of TMT sets increases. **B)** Data-wide MVs (%) as the number of TMT sets increases. **C)** Density plot showing protein distribution across detection %. Proteins split based on average abundance where the 1st quartile indicates the 25% least abundant proteins. **D)** The number of proteins identified at each detection %. Blue dashed line indicates the chosen imputation level of 75%, corresponding to 3290 proteins.

performed using the k-nearest neighbour algorithm (k-NN). k-NN imputation finds the 'k' most similar samples/proteins to a MV based on data that is present[538]. It then averages the nearby data points to assign a value. k-NN is widely implemented in molecular profiling studies and is considered a robust, sensitive, and effective tool to address MVs[538,539]. Given k-NN utilises similar/neighbour data to impute, it relies on confidence in the existing data structure. Accordingly, allowing excessive amounts of MVs leads to poor imputation performance. Proteins were therefore filtered prior to imputation to

increase data confidence. CPTAC studies permit the inclusion of proteins/genes with ≥ 50% non-MVs[523–529]. Herein, proteins present in ≥ 75% of samples were included and considered robustly identified (**Figure 3.5D)**. Filtering and imputation resulted in a final core dataset of 3290 proteins.

3.2.3.4 Normalisation

Acquisition of MS data from FPPE tumour specimens is a lengthy and involved process. This means there are many points at which bias can be introduced into the data. Despite ensuring the same amount of peptide is labelled and injected into the mass spectrometer, sample handling and instrument variation introduces inherent bias. This is amplified in TMT MS where samples are profiled in multiple batches. Normalisation is crucial to addressing this. Correct normalisation enables sample-to-sample comparisons and ensures the reliability of downstream analyses. **Figure 3.6** shows an overview of the normalisation procedure implemented herein and the progressive transformations in data distribution. The first aspect to address was inter-batch variation. Principal Component Analysis (PCA) of the raw data revealed samples to cluster by TMT set, and showed that the reference samples did not cluster together despite being an aliquot of the same sample (**Figure 3.7A** and **Figure 3.8**). Principal component 1 (PC1) accounted for a significant amount of variance within the data (31.18%) compared to other PCs (≤ 10.25%; **Figure 3.8**). Assessment of the PC loadings revealed albumin (ALB) as the most influential feature in PC1, and showed high component loading for ALB in PCs 2, 3, and 5 (**Figure 3.7B**). ALB is the most abundant plasma protein in humans[540]. As DDA MS1 sampling is dictated by abundance, this suggests either inconsistent amounts of sample were labelled, or inconsistent amounts of peptide were injected into the mass spectrometer. As an orthogonal approach to PCA, unsupervised clustering was performed. Clustering reiterated PCA-based observations, showing the references did not cluster together, and that significant batch effects were present (**Figure 3.7C**). To adjust for these differences, samples were normalised within-set relative to each respective reference sample. Next, to support comparative analyses, data was log_2-tranformed and a pan-TMT set adjustment performed by within-protein median centring. This ensured the median value for each protein across all samples was 0, transforming the value ranges without changing the scale of the data. Significant variation in the value range between samples persisted (**Figure 3.6**), and thus within sample standardisation was performed to enable valid downstream sample-to-sample comparative analyses. The resultant normalised data was re-assessed by PCA and unsupervised clustering and showed no evidence of batch effect (**Figure 3.7D-F**). Of note, the data features with highest component loadings in the top 5 PCs were proteins that have been differentially

reported across STS subtypes (**Figure 3.7E**). For example, myosin light chain kinase (MLYK) is an IHC marker for LMS, and cellular retinoic acid binding protein 1 (CRABP1) is expressed in monophasic SS[281,282,541,542]. This reassures that the proteomic data is appropriately normalised to reveal known biological differences within the cohort.

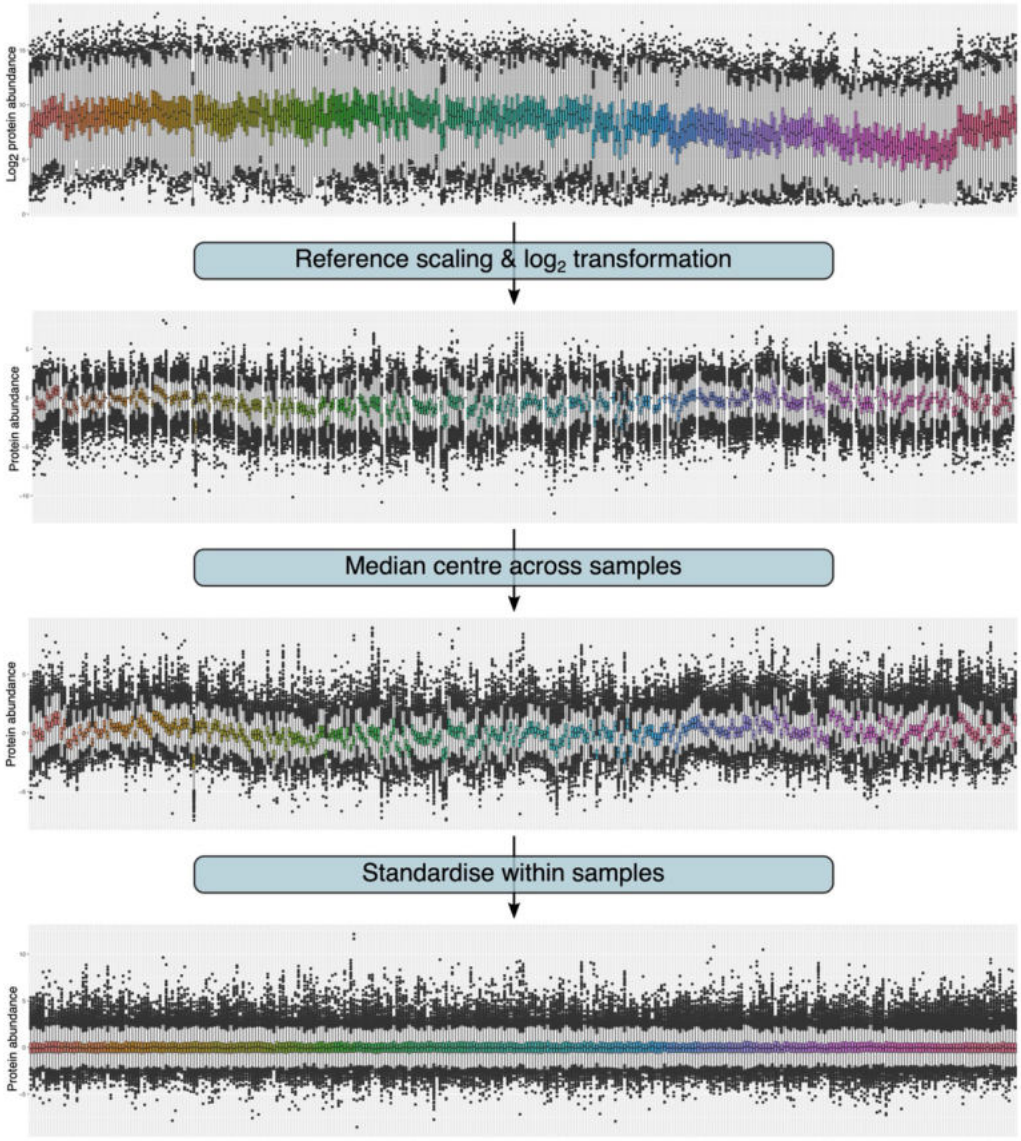


Figure 3.6 Data normalisation overview.
Sequential boxplots showing protein abundance of each sample throughout the normalisation procedure. From top to bottom plots show the imputed unnormalised data (n = 365), reference sample (REF) normalised data (n = 365), median centred data (n = 321), and standardised data (n = 321). Boxes are coloured by TMT set, and indicate 25th, 50th, and 75th percentile, with whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points.

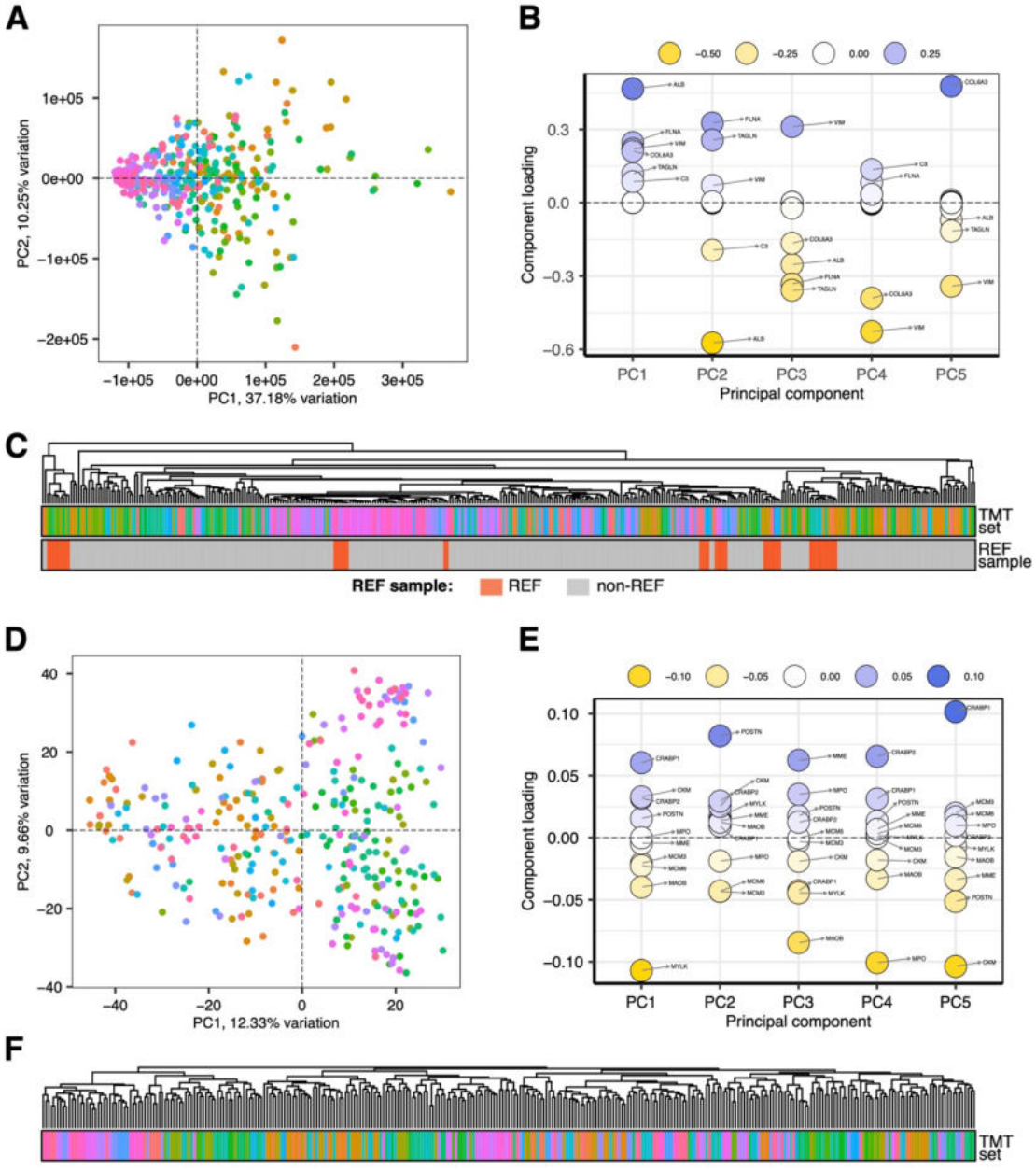


Figure 3.7 Assessment of batch effects in the unnormalised (A-C) and normalised (D-F) dataset. **(A & D)** Principal component analysis (PCA) biplot showing the dataset projected onto principal components (PC) 1 and 2. **B & E)** PC loading plot for PCs 1-5. Circles correspond to individual proteins and colour indicates component loading. Proteins of interest are annotated. **C & F)** Unsupervised hierarchical clustering dendrogram (Pearson's correlation distance) with tandem mass tag (TMT) set **(C & F)** and reference sample (REF; **C** only) annotation.

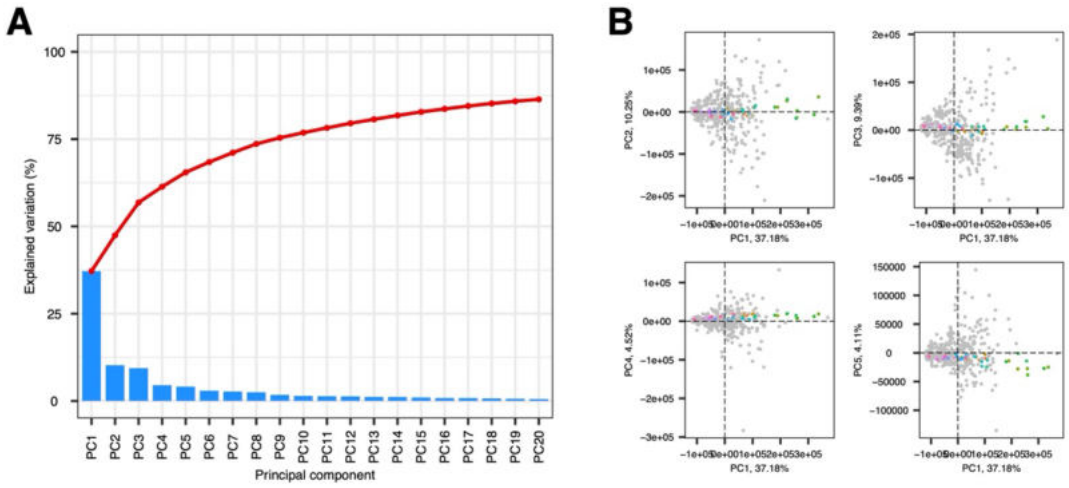


Figure 3.8 Additional principal component analysis (PCA) plots assessing batch effects in the unnormalised data.
(**A**) Scree plot showing the explained variance for the first 20 principal components (PC) in the unnormalised dataset. Red line shows cumulative variance. **B**) PCA biplots for PC 1 with PCs 2-4 in the raw dataset. Non-reference samples (non-REF) are coloured grey, reference samples (REF) coloured individually.

3.3 Discussion and summary

In this chapter, the proteomic data for multiple STS samples was collected and processed. Extraction of peptides from FFPE STS samples highlighted handling differences between different STS histologies, which led to improvements to the in-house peptide extraction protocol. This illustrates the challenges of working with a heterogenous group of diseases, where standardised protocols cannot be applied to different samples and subtypes without consideration. Also illustrated herein are the complexities involved in rare disease research. To achieve a reasonable cohort size, the inclusion of sub-optimal samples (in this case, samples with lower tumour purity) was necessary. This resulted in the macrodissection of many samples and revealed challenges in achieving sufficient yields from such samples. The protocol additions made herein were minor and therefore downstream consequences in the data are not considered a risk. However, if more subtypes, samples with smaller viable tumour areas, or excessively treatment-impacted samples were to be profiled, the suitability of this method would need re-visiting. This is particularly pertinent for studies requiring biopsy profiling, such as those assessing metastatic or recurrent STS disease, which is often not managed by surgical resection. Surgical resection specimens profiled herein which underwent macrodissection required 290 um of tissue on average. Biopsies are vastly smaller than surgical resections and it is likely that under the current protocol

requirements it would be near impossible to obtain sufficient tissue. One limitation in sample preparation was the reliance on a single clinical fellow to obtain tumour purity measures; tumour cell content is not a consistent measure between pathologists. If greater resources were available, it would have been preferable for at least 2 clinicians to estimate tumour purity independently and then establish a consensus score.

Following data collection, metrics were assessed for their use as QC measures, and a high proportion of MVs was identified as a suitable indicator of low-quality data. The use of an MV measure in this study was supported by its association with other measures indicative of poor data, such as protein abundance and bimodality statistics. One caveat of using MV is that it is a relative measure within each TMT set. Therefore, its use, and the specific threshold applied herein was tailored. In addition, the suitability of the MV measure was also identified based on visual inspection of the data distribution. This was feasible herein as data was collected on a rolling basis in batches of 10 tumour samples. If MS data were collected more rapidly, or if retrospective analyses were to be performed on an already collected and complete dataset, visual inspection would be an exceedingly laborious task. Moreover, visual interpretation is reliant on an individual and is vulnerable to bias and human error. This leaves room for QC improvement. Notably, most proteomic QC methods involve relative measures. CPTAC typically perform QC by analysing 2 reference samples within each batch[543]. Combining data from 2 TMT channels enables the creation of a more accurate 'virtual' reference, which acts a measure for overall data quality. However, this utilises double the amount of reference used herein. Given the scarcity of STS tumour tissue, it was not possible to use such large amounts of reference material. The reference samples utilised in most CPTAC studies, irrespective of cancer type, comprise breast xenograft tissue[543]. Similarly, the study herein utilised a non-specific reference sample to profile multiple STS subtypes. This resulted in minor data variations between TMT sets comprising subtypes in the reference sample and those that did not. However, such variations were minimal. The number of protein IDs was highly consistent, and the levels of MVs were consistently low; comparable if not better than previously published studies[466]. The good performance of the reference sample indicates it captures a sufficient level of general STS biology to enable multi-subtype profiling. This may have been aided by the abundance-based sampling of TMT profiling, meaning lowly expressed proteins, were unlikely to be identified. However, if an increased proteome depth was sampled, for example by further fractionation, this may raise issues as the lowly expressed proteins will be identified. The suitability of the reference in this study permitted its use for batch correction. To combine TMT sets, data was successfully normalised using the reference sample, alongside additional inter-

sample variation adjustments. This generated a dataset suitable for sample-to-sample proteomic comparisons.

Overall, this chapter has established a core STS proteomic dataset, confidently quantifying 3290 proteins spanning 321 primary tumour samples of 11 histological subtypes. The applicability of collecting such a rich dataset can be demonstrated by the work of CPTAC and TCGA[156,544]. These consortiums have conducted numerous studies across multiple cancer types, revealing important insights into disease biology, and identifying candidate biomarkers and drug targets. This chapter has therefore established a rich resource for the STS research community. Not only can this dataset be mined for primary analyses as this thesis will go on to evidence, but it can also be used for validation purposes. Validation is central for robust statistical validity in bench to bedside research and is frequently lacking in rare disease studies due to small study populations.

Chapter 4 Overview of the soft tissue sarcoma proteome

4.1 Background and objectives

In **Chapter 3** the collection and processing of proteomic data from primary tumours of patients with multiple histological subtypes of STS was described. In this chapter, the main objective is to provide a high-level, descriptive overview of the STS proteome of the profiled cohort. Multi-subtype profiling of STS has at present been comprehensively performed at the genomic and transcriptomic levels (discussed in **section 1.3)**. Herein, by considering the protein complement in relation to the current literature, we anticipate the revelation of both known and novel biology. To achieve this, the clinicopathological characteristics of the cohort were detailed. Pairwise interactions between clinicopathological variables were assessed, and their relation to clinical outcome measures investigated. Outcome measures of local-recurrence free survival (LRFS), metastasis-free survival (MFS), and overall survival (OS) were used for analysis; all of which were censored at 5-years following surgical resection of the primary disease lesion (i.e., the profiled specimen). Full details as to how these outcome measures were calculated are detailed in **section 2.7.7.** The complete proteomic data was first assessed by unsupervised clustering. Supervised comparisons were then performed to reveal histology-specific proteins. In addition, the representation of several sub-proteome datasets (the adhesome, matrisome, immune component and kinome) in the proteomic data was investigated[500,501,503,545]. To build on proteome-wide observations, descriptive analyses were used to detail the relationship between sub-proteomes and histological subtype. As well as interrogating protein-level data, enrichment analyses were employed to query the overarching biological features within the data. This analysis generated summary enrichment profiles for each tumour, corresponding to specific biological activities, pathways, and drug target profiles. As with sub-proteome investigations, observations of broad biological features and targetable signatures were descriptively noted. By transforming the proteome-wide data into multiple working datasets, this chapter supports the development of a comprehensive proteome-centric understanding of STS. It is hypothesised that through extensive descriptive analysis of this data, several avenues for further research efforts will emerge.

4.2 Results

4.2.1 Baseline cohort characteristics

Baseline clinicopathological characteristics of the profiled cohort are detailed in **Table 4.1**. In total, 321 primary tumours spanning 11 histological subtypes were profiled. Most tumours were either LMS (25%), UPS (17%), SS (13%), or DDLPS (12%). AS comprised a moderate proportion of the cohort (9%), and limited numbers of ultra-rare subtypes (EPS, ASPS, DSRCT, CCS) were included (5%, 1%, 1%, and 1% respectively). In addition, DES, a non-metastatic soft tissue tumour, and RT, a paediatric STS, were profiled and comprised 12% and 4% of the cohort respectively. Overall, 63% were genomically complex STS (LMS, UPS, DDLPS, AS), 16% genomically simple STS driven by fusion events (SS, ASPS, DSRCT, CCS), and 9% genomically simple STS with key recurrent mutations (EPS, DES, RT). There were extensive interactions between the clinicopathological variables. Pairwise associations are summarised in **Supplemental Table 4.1**. Higher order interactions (i.e., interactions between more than 2 variables) were not assessed, but likely exist. Due to the vastly different clinical presentation of DES and RT compared to the typical adult STS population, RT is a paediatric subtype, and DES a locally infiltrative subtype with no metastatic potential; for the purposes of evaluating clinicopathological associations, these diagnoses were included in descriptive analyses but excluded from statistical analyses. Median age of the cohort was 58.4 years (range: 0.1 - 90). The median ages were highest (> 60 years) for UPS, LMS, DDLPS, and AS, and lowest (< 30 years) for DSRCT and ASPS (**Supplemental Figure 4.1A**). Ages of the paediatric RT tumours ranged from 0.1 – 4.7 years. Age also showed an association with grade, tumour depth and PS (**Supplemental Figure 4.1B-D**); with lower grade, deep tumours, and a lower PS seen in younger patients. This cohort comprised more females than males (62.6% vs 37.1%), driven predominantly by the higher incidence of AS in females, which can arise subsequent to radiotherapy for breast carcinoma (**Supplemental Figure 4.1E**). An enrichment of males was seen in CCS, DDLPS and DSCRT. Sex was also associated with anatomical site, likely reflective of subtype differences and the inclusion of uterine tumours (**Supplemental Figure 4.1F**). Furthermore, sex showed a significant association with tumour size; males harboured larger tumours (**Supplemental Figure 4.1G**). Median tumour size was 90 mm (range: 4 – 1090), with larger tumours seen in DDLPS patients (**Supplemental Figure 4.1H**). This was reflective of anatomical site, as most DDLPS were retroperitoneal and large (**Supplemental Figure 4.1I** and **Supplemental Figure 4.1J**). Across the cohort, extremities were the most common sites of disease (38.9%). Most AS were trunk wall, and most DSRCT were intra-abdominal (**Supplemental Figure 4.1I**). The only uterine

Table 4.1 Clinicopathological features of the cohort.
Features of total cohort and individual histological subtypes of soft tissue sarcoma (STS). Continuous variables detailed as median, minimum (min), and maximum (max). Categorical variables detailed as count (n) and percentage. Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DES = desmoid tumour; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; RT = rhabdoid tumour; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma; F = female; M = male; CTX = chemotherapy; RTX = radiotherapy.

		Total	AS	ASPS	CCS	DDLPS	DES	DSRCT	EPS	LMS	RT	SS	UPS
	n	**321**	**30**	**4**	**3**	**39**	**37**	**4**	**16**	**80**	**12**	**43**	**53**
Age at excision (years)	**median**	58.4	68.8	22.3	49.1	63	39.3	28.7	38.5	65.3	1.1	42.3	73.5
	min	0.1	27.3	18.1	25.2	35.1	21.2	16.6	18.3	29.3	0.1	19.6	28.2
	max	90	82.7	33.9	61.9	81.3	78.3	46.1	76.8	86.9	4.7	79.4	90
Tumour size (mm)	**median**	90	58	68	55	190	90	132.5	50	92.5	-	71	80
	min	4	4	30	9	35	25	70	10	5	-	18	15
	max	1090	400	100	95	1090	500	175	240	400	-	760	360
Sex [n (%)]	**F**	201 (62.6)	26 (86.7)	3 (75)	1 (33.3)	15 (38.5)	29 (78.4)	1 (25)	8 (50)	56 (70)	7 (58.3)	27 (62.8)	28 (52.8)
	M	119 (37.1)	4 (13.3)	1 (25)	2 (66.7)	24 (61.5)	8 (21.6)	3 (75)	8 (50)	24 (30)	4 (33.3)	16 (37.2)	25 (47.2)
	unknown	1 (0.3)	-	-	-	-	-	-	-	-	1 (8.3)	-	-
Grade [n (%)]	**2**	115 (35.8)	12 (40)	1 (25)	-	19 (48.7)	-	-	10 (62.5)	47 (58.8)	-	23 (53.5)	3 (5.7)
	3	139 (43.3)	13 (43.3)	-	3 (100)	20 (51.3)	-	3 (75)	5 (31.2)	33 (41.2)	-	13 (30.2)	49 (92.5)
	unknown	67 (20.9)	5 (16.7)	3 (75)	-	-	37 (100)	1 (25)	1 (6.2)	-	12 (100)	7 (16.3)	1 (1.9)
Anatomical site [n (%)]	**Extremity**	125 (38.9)	2 (6.7)	4 (100)	3 (100)	2 (5.1)	9 (24.3)	-	9 (56.2)	31 (38.8)	1 (8.3)	26 (60.5)	38 (71.7)
	Head/neck	13 (4)	4 (13.3)	-	-	-	1 (2.7)	-	-	-	2 (16.7)	2 (4.7)	4 (7.5)
	Intra-abdominal	28 (8.7)	2 (6.7)	-	-	3 (7.7)	4 (10.8)	3 (75)	-	10 (12.5)	3 (25)	2 (4.7)	1 (1.9)
	Retroperitoneal	57 (17.8)	1 (3.3)	-	-	32 (82.1)	-	1 (25)	-	19 (23.8)	2 (16.7)	2 (4.7)	-
	Trunk	65 (20.2)	21 (70)	-	-	2 (5.1)	22 (59.5)	-	1 (6.2)	2 (2.5)	2 (16.7)	7 (16.3)	8 (15.1)
	Pelvic	24 (7.5)	-	-	-	-	1 (2.7)	-	6 (37.5)	9 (11.2)	2 (16.7)	4 (9.3)	2 (3.8)
	Uterine	9 (2.8)	-	-	-	-	-	-	-	9 (11.2)	-	-	-

Continuation of table from previous page

Tumour depth [n (%)]	**Deep**	250 (77.9)	15 (50)	4 (100)	3 (100)	38 (97.4)	30 (81.1)	4 (100)	8 (50)	66 (82.5)	-	39 (90.7)	43 (81.1)
	Superficial	54 (16.8)	15 (50)	-	-	1 (2.6)	2 (5.4)	-	8 (50)	14 (17.5)	-	4 (9.3)	10 (18.9)
	unknown	17 (5.3)	-	-	-	-	5 (13.5)	-	-	-	12 (100)	-	-
Status at excision [n (%)]	**Local**	301 (93.8)	29 (96.7)	3 (75)	2 (66.7)	36 (92.3)	37 (100)	2 (50)	13 (81.2)	78 (97.5)	6 (50)	42 (97.7)	53 (100)
	Metastatic	15 (4.7)	1 (3.3)	1 (25)	1 (33.3)	2 (5.1)	-	2 (50)	-	2 (2.5)	5 (41.7)	1 (2.3)	-
	Locally Metastatic	3 (0.9)	-	-	-	-	-	-	3 (18.8)	-	-	-	-
	Multifocal	1 (0.3)	-	-	-	1 (2.6)	-	-	-	-	-	-	-
	unknown	1 (0.3)	-	-	-	-	-	-	-	-	1 (8.3)	-	-
Radiation associated [n (%)]	**No**	285 (88.8)	14 (46.7)	4 (100)	3 (100)	39 (100)	37 (100)	4 (100)	16 (100)	78 (97.5)	-	42 (97.7)	48 (90.6)
	Yes	24 (7.4)	16 (53.3)	-	-	-	-	-	-	2 (2.5)	-	1 (2.3)	5 (9.4)
	unknown	12 (3.7)	-	-	-	-	-	-	-	-	12 (100)	-	-
Tumour margins [n (%)]	**R0**	133 (41.4)	17 (56.7)	1 (25)	1 (33.3)	9 (23.1)	10 (27)	-	11 (68.8)	42 (52.5)	-	18 (41.9)	26 (49.1)
	R1	151 (47)	11 (36.7)	3 (75)	2 (66.7)	25 (64.1)	20 (54.1)	1 (25)	5 (31.2)	35 (43.8)	-	22 (51.2)	27 (50.9)
	R2	4 (1.2)	-	-	-	-	1 (2.7)	1 (25)	-	1 (1.2)	-	1 (2.3)	-
	unknown	33 (10.3)	2 (6.7)	-	-	5 (12.8)	6(16.2)	2 (50)	-	2 (2.5)	12 (100)	2 (4.7)	-
Performance status [n (%)]	**0**	158 (49.2)	15 (50)	4 (100)	2 (66.7)	17 (43.6)	28 (75.7)	3 (75)	7 (43.8)	40 (50)	-	20 (46.5)	22 (41.5)
	1	82 (25.5)	12 (40)	-	-	12 (30.8)	4 (10.8)	1 (25)	5 (31.2)	16 (20)	-	17 (39.5)	15 (28.3)
	2	16 (5)	-	-	-	2 (5.1)	-	-	-	7 (8.8)	-	3 (7)	4 (7.5)
	3	5 (1.6)	-	-	-	1 (2.6)	-	-	-	1 (1.2)	-	-	3 (5.7)
	unknown	60 (18.7)	3 (10)	-	1 (33.3)	7 (17.9)	5 (13.5)	-	4 (25)	16 (20)	12 (100)	3 (7)	9 (17)
Pre-op treatment [n (%)]	**CTX**	19 (5.9)	5 (16.7)	-	-	1 (2.6)	3 (8.1)	3 (75)	-	-	-	7 (16.3)	-
	RTX	8 (2.5)	-	1 (25)	-	-	-	-	-	1 (1.2)	-	6 (14)	-
	CTX & RTX	13 (4)	-	-	-	-	-	1 (25)	-	-	-	12 (27.9)	-
	None	267 (83.2)	25 (83.3)	3 (75)	3 (100)	38 (97.4)	34 (91.9)	-	15 (93.8)	79 (98.8)	-	17 (39.5)	53 (100)
	unknown	14 (4.4)	-	-	-	-	-	-	1 (6.2)	-	12 (100)	1 (2.3)	-

tumours were uLMS. Most tumours were deep seated (77.9%), with deep tumours unsurprisingly reflective of larger tumours (**Supplemental Figure 4.1K**). Superficial tumours were predominantly seen in AS and EPS patients (**Supplemental Figure 4.1L**). There was a slight enrichment of high grade tumours vs intermediate grade tumours (35.8% grade 2 and 43.3% grade 3), and this reflected histology (**Supplemental Figure 4.1M**). The majority of the cohort were treatment naïve (83.2%). DSRCT and SS were the only subtypes where more patients received preoperative treatment than did not. Surgical margins were typically either R0 and R1, with an approximate equal split between them (41.4% and 47% respectively). At diagnosis, half (49.2%) of patients had a PS of 0, and 25.5% a PS of 1. Few patients had a PS > 1, indicative of high functional impairment of the patient[546], and those that did had either DDLPS, LMS, SS, or UPS. 7.4% (n = 24) of tumours were aetiologically identified as radiation associated. AS accounted for most of the radiation associated lesions overall (66.7%), and most AS were radiation-associated (53.3%) occurring secondary to breast carcinoma.

Missing values in the clinicopathological data were few, although most variables had some level of missingness (range: 0% - 29%; **Table 4.1**). Grade showed the highest missingness, however this was due to the inapplicability of grading to DES and RT diagnoses. PS also showed high missingness due to incomplete clinical records. A large part of the remaining missingness was introduced by the RT cohort, as data was unretrievable through collaborators.

4.2.2 Cohort outcomes and the prognostic significance of clinicopathological variables

Survival data was censored at 5 years and therefore information on longer term outcomes was not available. Median LRFS was not reached (**Figure 4.1A).** Median MFS and OS for the cohort were approximately 48 and 52 months respectively (**Figure 4.1B-C).** At 5-years post-surgery, 37% of patients had experienced a local recurrence event, 48% had experienced a metastatic event, and 47% were deceased.

The Kaplan-Meier curve and univariable Cox regression were used to assess the relationship between individual clinicopathological variables and each outcome measure. Results of the Cox regression analysis are summarised in **Supplemental Table *4.2***. Survival analyses excluded RT and DES as both differ significantly from the typical adult STS population. Due to the heterogeneity of this cohort, many grouped

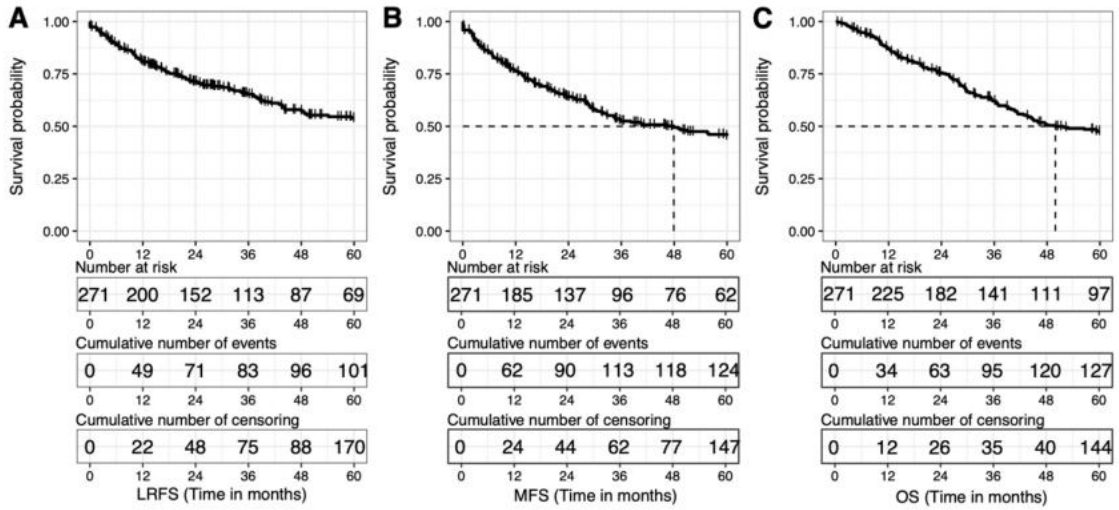


Figure 4.1 Clinical outcome of the proteome-profiled cohort.

Kaplan Meier plots showing local recurrence free survival (LRFS; **A**), metastasis free survival (MFS; **B**), and overall survival (OS; **C**) up to 5-years post-surgery. Dashed line indicates median survival.

variables were unbalanced. Where biologically, clinically, and statistically appropriate, features were combined. The ultra-rare subtypes (ASPS, CCS, DSRCT) were grouped as 'Other', PS of 2 and 3 were grouped, and tumour margins of R1 and R2 were grouped. Of the tumour characteristics, histological subtype was the only variable associated with all 3 outcome measures (**Figure 4.2A**). Compared to the reference population (LMS), DDLPS showed a significantly inferior LRFS (HR = 3.32; 95% CI = 1.87 - 5.87; $p < 0.001$), as did AS (HR = 4.45; 95% CI = 2.33 - 8.51; $p < 0.001$). DDLPS also showed a significantly superior MFS (HR = 0.417; 95% CI = 0.21 – 0.829; p = 0.013), and AS showed a significantly inferior OS (HR = 1.96; 95% CI = 1.09 – 3.53; p = 0.024). These observations are in line with the current literature, which reports DDLPS to have a high local recurrence rate and low metastasis rate compared to other common STS subtypes such as LMS and UPS[341]. AS is reported to have poor overall survival rates and a high propensity for recurrence (both local and distant)[547,548]. Notably, the LMS reference population also has a reported high metastatic potential[245,549–551]. Therefore, whilst the Cox regression did not show a significant difference in MFS between AS and LMS, inspection of the Kaplan-Meier curve did reveal AS to have one of the shortest median MFS (~ 28 months), along with EPS (~ 15 months) and 'Other' (~ 28 months; **Figure *4*.2A**). Anatomical site was significantly associated with LRFS (**Figure *4*.2B**). Both retroperitoneal and trunk wall showed a poorer LRFS compared to extremity (HR = 2.33; 95% CI = 1.42 – 3.82; p = 0.001 and HR = 2.08; 95% CI = 1.16 – 3.73; p = 0.014

respectively). As in previous reports, a higher grade was significantly associated with a shorter MFS (HR = 1.89; 95% CI = 1.3 – 2.75; p < 0.001) and shorter OS (HR = 1.984;

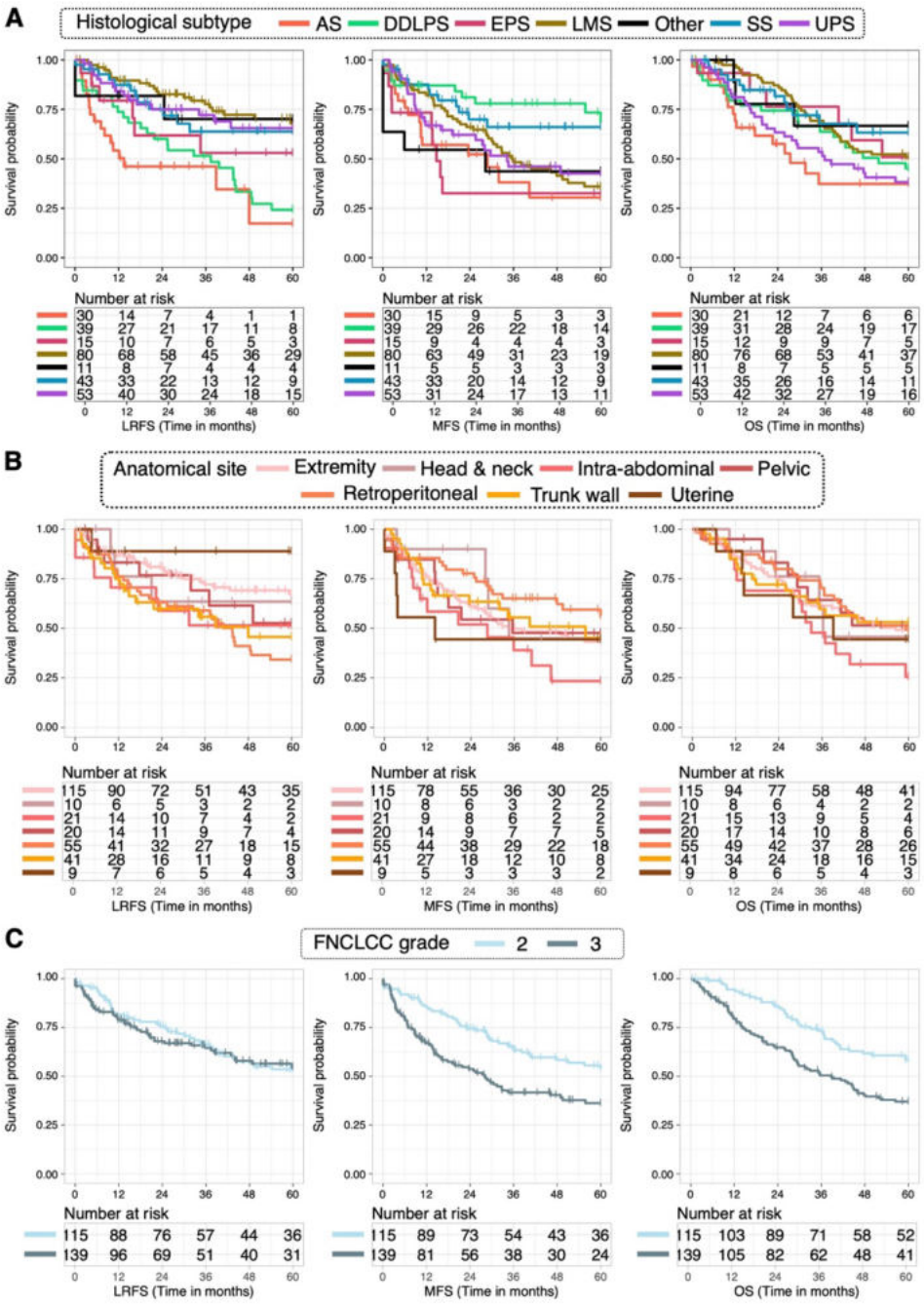


Figure 4.2 Clinical outcome of the proteome-profiled cohort stratified by key tumour characteristics
Kaplan Meier plots showing from left-to-right local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) up to 5-years post-surgery. **(A)** Stratification by histological subtype, where 'other' indicates ASPS, DSRCT, and CCS. **(B)** Stratification by anatomical site. **(C)** Stratification by French Federation of Cancer Center Sarcoma Group (FNCLCC) grade. Abbreviations: AS = angiosarcoma; DDLPS = dedifferentiated liposarcoma; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma. Corresponding univariable Cox regression results are detailed in **Supplemental Table 4.2**.

95% CI = 1.366 – 2.882; p < 0.001; **Figure *4.2*C**), and tumour size was significantly associated with LRFS; although the effect size was negligible (HR = 1; 95% CI = 1 - 1; p = 0.001)[45,46]. Of the patient characteristics, a PS of 1 was significantly associated with a poorer LRFS compared to PS 0 (HR = 1.7; 95% CI = 1.09 – 2.64; p = 0.019; **Figure *4.3*A**), and PS of 1 and 2-3 were significantly associated with a poorer OS compared to PS 0 (HR = 2.18; 95% CI = 1.44 – 3.31; p < 0.001 and HR = 4.04; 95% CI = 2.31 – 7.09; p < 0.001 respectively). Notably, the group of patients for which PS is unknown also showed a significantly poorer OS (HR = 1.76; 95% CI = 1.05 – 2.96; p = 0.031). It is assumed that the true PS of this 'unknown' group, reflect the PS distribution within the rest of the cohort. PS is reported to be a strong predictor for overall patient outcome, which is corroborated in this cohort[45,46,552]. Therefore, is it possible that the patients with a higher PS in the 'unknown' group are driving the significant association with outcome. Also consistent with the literature, males are associated with a significantly shorter OS (HR = 1.63; 95% CI = 1.15 – 2.3; p = 0.006; **Figure *4.3*B**)[45,46]. A higher age was significantly associated with OS, although as with tumour size the effect size was negligible (HR = 1.03; 95% CI = 1.01 – 1.04; p < 0.001). Notably, use of preoperative treatment did not impact outcome (**Supplemental Figure *4.2*A**). Data on adjuvant therapy for patients was unavailable, limiting conclusions as to the impact of surgery alone compared to surgery in combination with chemotherapy or RTX in this cohort. Contrary to some published reports, the surgical margins achieved, and the tumour depth did not show any relation to outcome (**Supplemental Figure *4.2*B-C**)[45,46].

Use of the Cox regression model relies on 2 assumptions[553]. Firstly, that the hazard functions are proportional over time, i.e., for any 2 individuals the ratio of the hazards are constant over time. This is known as the proportional hazards (PH) assumption. Secondly, the Cox model assumes that the log hazard of any continuous covariate is linear. The assumptions for each clinicopathological variable were therefore assessed in null univariable models. All variables satisfied the PH assumption (Schoenfeld test p > 0.01); however, 4 variables did show minor violations (p: 0.01 - 0.05). Namely, subtype in MFS analysis (p = 0.022), preoperative treatment in MFS analysis (p = 0.014), sex in OS analysis (p = 0.039), and PS in both MFS and OS analyses (p = 0.043 and 0.026 respectively). These violations were visually assessed by plotting the deviance and Schoenfeld residuals. The deviance residuals portray the contribution of each sample to the model. Outliers are indicated by deviance residuals with relative extreme values. As such, for model validity deviance residuals are expected to follow an approximate symmetrical distribution around 0. The Schoenfeld residuals directly reflect PH. Where the PH assumption is met, Schoenfeld residuals show time independence, thus a non-

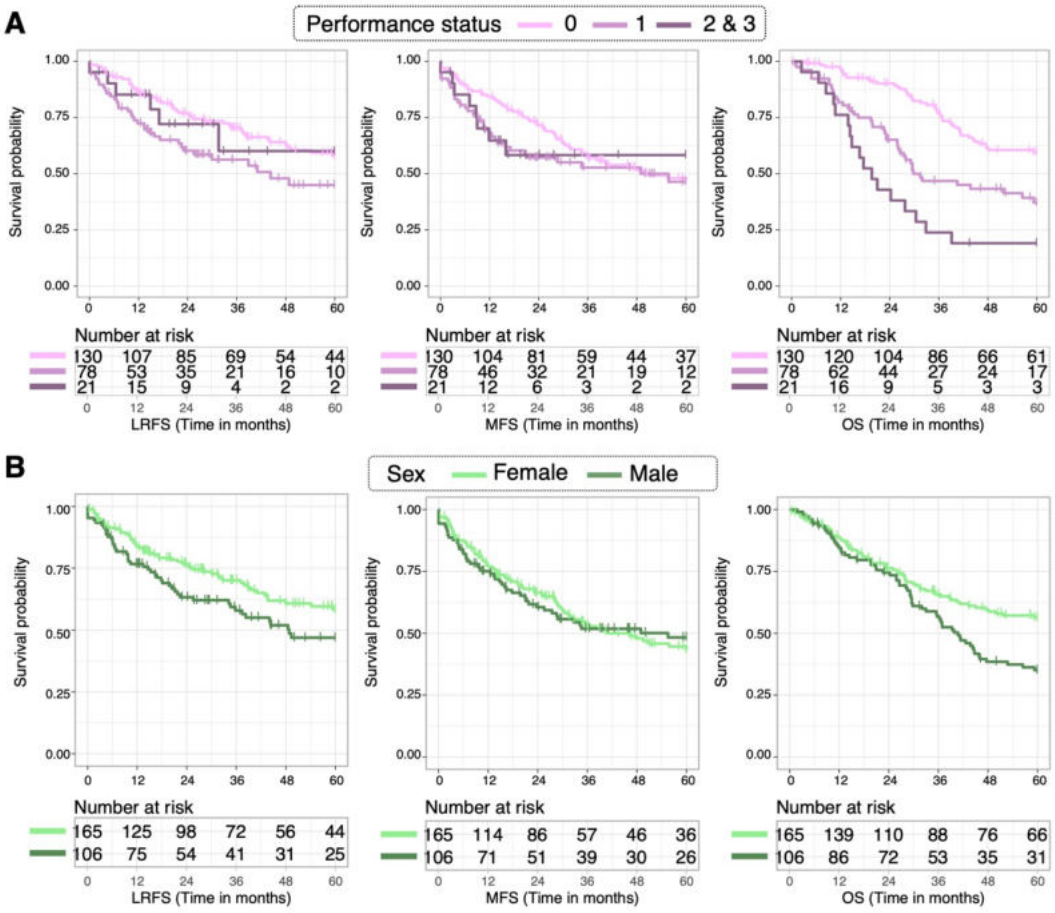


Figure 4.3 Clinical outcome of the proteome-profiled cohort stratified by key patient characteristics Kaplan Meier plots showing from left-to-right local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) up to 5-years post-surgery. **(A)** Stratification by performance status (PS). **(B)** Stratification by sex. Corresponding univariable Cox regression results are detailed in **Supplemental Table 4.2**.

random pattern once plotted indicates PH violation. Herein, the deviance residual plots showed reasonable symmetry around 0 in all cases where a minor PH violation was highlighted (**Supplemental Figure *4.3*A,C,E,G**). This revealed no concerning outliers or overly influential observations. Plotting the Schoenfeld residuals showed minor trends, but no obvious difference was seen between the expected and observed events (**Supplemental Figure *4.3*B,D,F,H**). Given all observations, use of these variables in the Cox regression model was deemed valid. The continuous variables of age and tumour size were also assessed for the presence of non-linearity by plotting the martingale residuals. Martingale residuals range from -∞ to 1, where in the case of OS a low value indicates the patient lived longer than expected based on the model fit, and a value close to 1 indicates the patient died sooner than the model would predict. For patient age, the linearity assumption was met (**Supplemental Figure *4.4***). Only minor deviations from 0 were observed in the highest ages (> 75 years) in the MFS model. Importantly, in the OS model, where age was revealed to be significant, good linearity was seen. For tumour

size, clear non-linearity was seen in all outcome models (**Figure *4.4*A**). Martingale residuals increased in LRFS up to tumours of 300 mm, and to 150 mm for MFS and OS, before decreasing. Strikingly, the case with the lowest martingale residual (LRFS and OS) was the largest tumour, indicative of an indolent nature in this case. By contrast, the second largest tumour (LRFS, MFS, and OS) had a martingale residual near 1. This data point corresponded to a grade 3 SS tumour with metastases present at diagnoses, explaining the high martingale residuals. It was evident that tumour size as an untransformed continuous variable was unsuitable for inclusion in the Cox regression model. To address this, data was log-transformed, and cut-points selected based on visual inspection of the martingale residuals. For continuity, the same stratification was used across all outcome measures. Cut points were selected at 4 and 5; close to the inflection points in LRFS, MFS, and OS, and balancing the number of patients in the smallest (< 4) and largest (> 5) groups (n = 65; **Figure *4.4*B**). Implementation of this stratification revealed tumours > 5 log(mm) to have a significantly poorer LRFS (HR = 1.99; 95% CI = 1.29 – 3.06; p = 0.002), and tumours < 4 log(mm) to have a significantly superior MFS (HR = 0.451; 95% CI = 0.272 – 0.746; p = 0.002) and OS (HR = 0.542; 95% CI = 0.323 – 0.912; p = 0.021; **Figure *4.4*C**). The general increase in risk with increasing size is consistent with previous studies[45,46].

Multivariable analysis was performed to assess the independent significance of each variable, when other clinicopathological variables were adjusted for. Results are summarised in **Supplemental Table *4.3***. Given preoperative treatment is inconsistently reported in the literature as a predictor, and herein was not significant in univariable analysis, was highly unbalanced, and showed minor PH violations, it was excluded[45,46,87,88]. In brief, multivariable LRFS analysis revealed histological subtype, PS, and log(tumour size) to be the only significant variables. The significance of anatomical site was lost upon multivariable adjustment. Multivariable MFS analysis revealed all variables significant in univariable analysis to retain significance (histological subtype, grade, and log(tumour size)). Multivariable OS analysis revealed histological subtype, PS, grade, anatomical site, and log(tumour size) as significant variables. The significance of age was lost. It is important to recognise that variables may not behave identically in univariable and multivariable models. Thus, following multivariable modelling, all assumptions were re-assessed as before. Continuity throughout was desirable and thus where possible, variables were handled as detailed in the univariable analyses. This means that whilst the variables were not optimised to the new model, statistical validity was ensured. As before, PH and linearity (for age) assumptions were found to be met. Use of the transformed and stratified version of tumour size was valid,

and only 2 variables showed minor PH violations (MFS histological subtype Schoenfeld p = 0.014; MFS log(tumour size) Schoenfeld p = 0.04; **Supplemental Figure *4.5***).

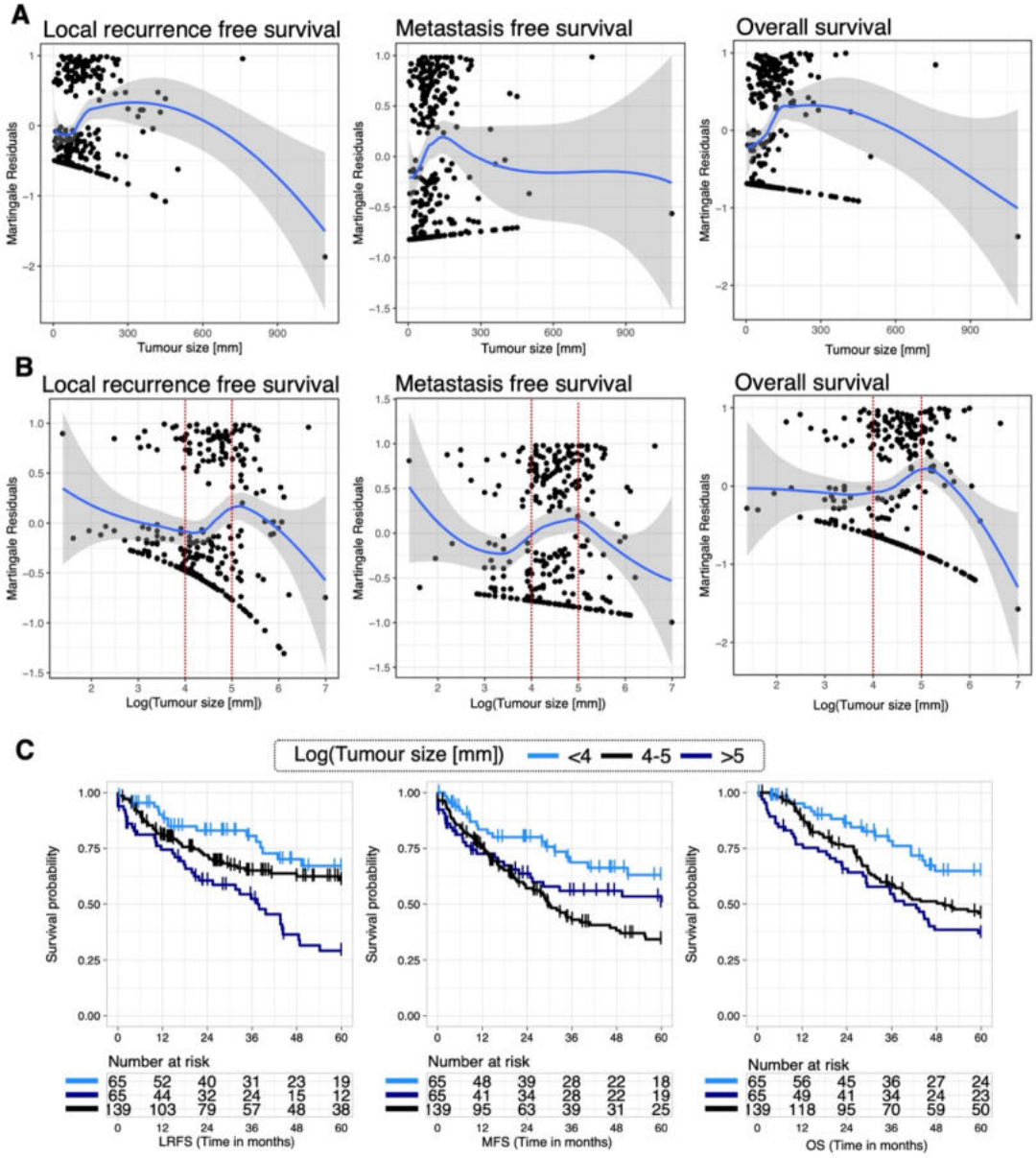


Figure 4.4 Assessing the linearity of tumour size in Cox regression models
(A-B) Left to right: plots for local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) showing martingale residuals against **(A)** tumour size **(B)** and log(tumour size). Blue line indicates a locally weighted smoothed fit and grey shading the coordinate 95% confidence intervals. Red dashed lines **(B)** indicate selected cut points for categorisation of the variable. **(C)** Kaplan Meier plots showing from left to right LRFS, MFS, and OS up to 5-years post-surgery. Stratification by log(tumour size). Corresponding univariable Cox regression results are detailed in **Supplemental Table 4.2**.

4.2.3 The pan-STS proteome landscape

4.2.3.1 An overview of the proteome of STS

Proteomic profiling of primary tumour specimens robustly identified and quantified 3,290 proteins. To visualise the proteome in relation to key clinicopathological variables, unsupervised clustering was performed (**Figure *4.5*A**). This evidenced histological subtype proteome differences, with some subtypes clustering as individual diagnoses, and some clustering in mixed groups. Beyond this, there was no apparent association between other clinicopathological features (anatomical site, grade, sex, age and tumour size) and the unsupervised clustering results. This was true both at the inter- and intra-subtype level. For example, high grade tumours of different subtypes did not cluster together, and samples of the same histological subtype but different anatomical sites did not cluster closely together (e.g., uLMS and stLMS). One major caveat of these interpretations is the close associations noted between some of the clinicopathological variables (**section 4.2.1**). Such associations, coupled with the dominant relationship seen between histology and the proteome in our data, severely limit comparative assessment of other clinicopathological features. Thus, to provide an overview, the proteomic data was further analysed solely in the context of histological subtype.

4.2.3.2 Histological subtype features of the STS proteome

Individual, distinctive clusters of SS, DES, and LMS were observed; with a few tumours of these diagnoses clustering elsewhere (**Figure *4.5*A**). Specifically, only 4 SS, 2 DES and 12 LMS clustered outside of their respective main subtype-specific clusters. DDLPS, RT, and to a lesser extent UPS tumours mostly clustered as diagnosis specific groups, albeit less robustly than SS, DES, and LMS. AS showed the highest level of heterogeneity. AS did not form a clear AS-specific cluster, but instead appeared spread in multiple clusters alongside other subtypes. However, there were exceptions to these clustering patterns, some of which are surprising. For example, SS tumours and LMS tumours in this dataset, and previously published transcriptomic studies, showed distinctive molecular profiles[36,41,165]. However, 1 EPS tumour and 1 LMS tumour were clustered with SS tumours, and 1 UPS tumour and 1 DSRCT tumour were clustered with LMS. These present as ‘outliers’. To better understand whether these are biologically true observations, the sample preparation records for these samples were revisited. In all cases, the ‘outlier’ was not processed alongside any sample that clustered in the vicinity of it (e.g., the DSRCT case was not processed in batches containing LMS tumours). As such, cross-contamination in these samples was highly unlikely to be driving the results.

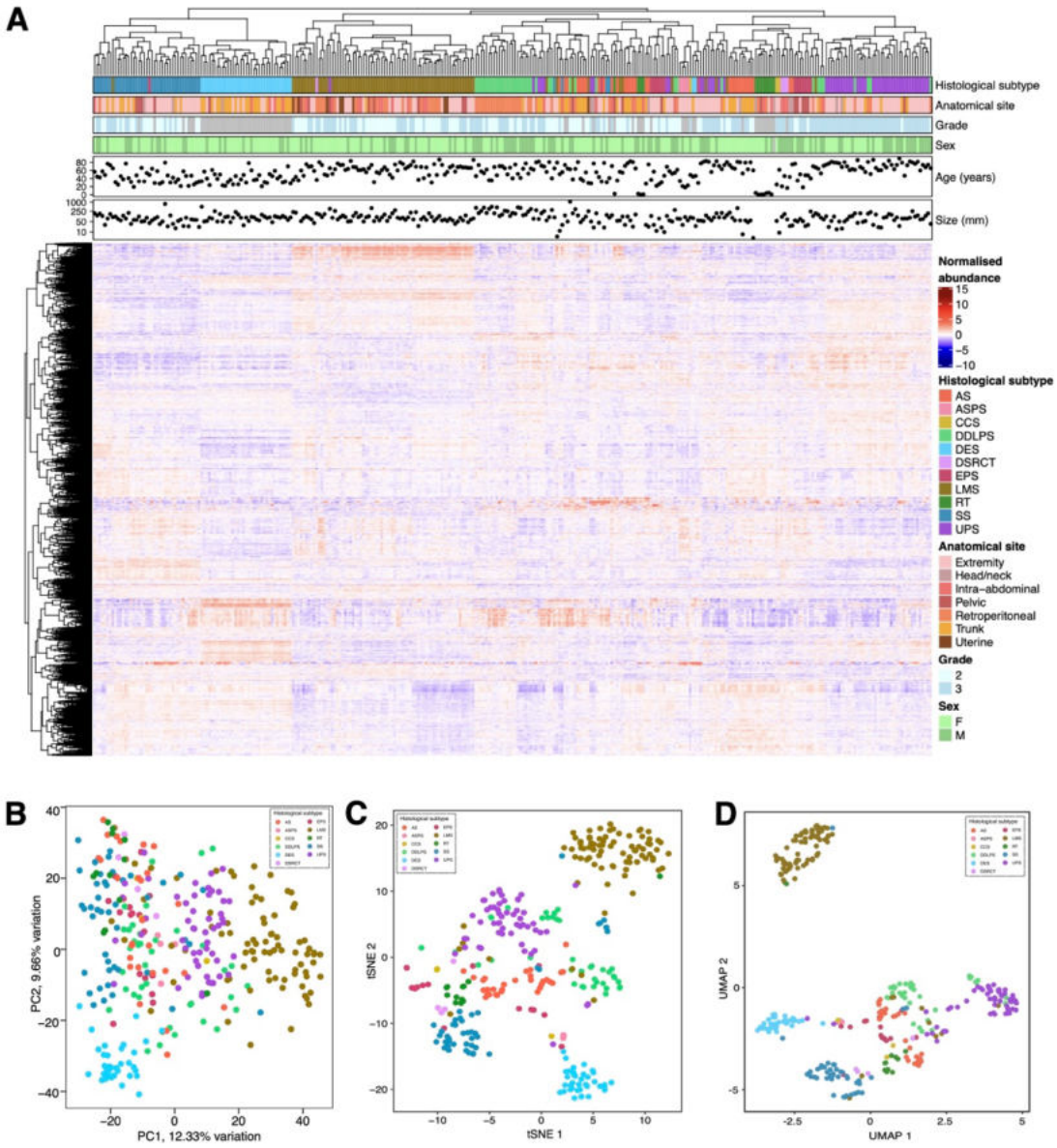


Figure 4.5 The proteome landscape of soft tissue sarcoma (STS).

(A) Annotated heatmap showing the unsupervised clustering (Pearson's distance) of 3290 proteins across the study cohort. From top to bottom, panels indicate histological subtype, anatomical site, tumour grade, patient sex, patient age, and tumour size. **(B-D)** Dimension reduction of the proteomic data with individual cases coloured by histological subtype, using **(B)** principal component analysis (PCA), **(C)** t stochastic neighbour embedding (tSNE), and **(D)** uniform manifold approximation and projection (UMAP). Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DES = desmoid tumour; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; RT = rhabdoid tumour; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma.

Complementary to the unsupervised clustering, PCA, t stochastic neighbour embedding (tSNE), and uniform manifold approximation and projection (UMAP) were also used to visualise data (**Figure *4.5*B-D**). PCA, tSNE and UMAP are dimension reduction approaches which can project high dimensional data in a 2/3-dimension space[513–516].

PCA is a linear method, which assigns equal weights to all pairwise distances. By contrast, tSNE and UMAP are non-linear algorithms which can often achieve better preservation of local data structure distances (i.e., within cluster distances) than PCA. In many cases, UMAP is superior to tSNE as it can also offer improved global structure preservation (i.e., between cluster distances). As a result, the inter-cluster distances in UMAP can have more 'meaning' than in tSNE and PCA. In this dataset, PCA achieved poor clustering with limited separation of samples in PC1 and PC2 (**Figure *4.5*B**). tSNE appeared to approximately mimic the unsupervised clustering results (**Figure *4.5*C**). UMAP however, due to its ability to preserve global structure, illustrated a strikingly prominent LMS signature to be present (**Figure *4.5*D**). In UMAP, most LMS cases clustered separately from the rest of the data, suggesting LMS to have the most distinctive subtype-specific proteome within this cohort.

Given the strong observed relationship between histological subtype and the proteomic data, supervised comparisons were performed to identify the proteins contributing to differences between histologies. To ensure robustness of results, only histological subtypes with ≥ 20 samples were interrogated. Each of these subtypes was compared to the rest of the cohort using Significance Analysis of Microarrays (SAM) 2-class unpaired tests[499]. Significant differentially expressed proteins (DEPs) were defined as those with an FDR < 0.01 and fold change ≥ 1.5. The upregulated DEPs across each comparison were then compared to identify subtype-specific (i.e., unique) upregulated DEPs. To investigate whether these subtype-specific DEPs contained shared biology, the significantly upregulated proteins were queried against the Gene Ontology (GO) and Hallmark gene sets from the Molecular Signature Database (MSigDB) using over-representation analysis[512]. The GO and Hallmark gene sets comprise an expansive set of biological processes and the genes involved in them[506–508]. As suggested by name, gene sets were established at the gene expression level. However, they are applicable to any high dimensional data such as proteomics. The major limitation to the use of gene sets herein is the low proteome coverage in our dataset (3290 proteins) relative to the genome (~25,000 genes). Therefore, to prevent spurious results, a background of only those proteins detected in the dataset was used, instead of the whole genome.

In AS, 386 DEPs were upregulated and 355 downregulated (**Figure *4.6*A**). Of these, 191 were uniquely upregulated in AS relative to all other samples (**Figure *4.6*G**). AS is a disease of the vascular and lymphatic cell lineage[4]. In accordance with this, upregulated AS proteins included those central to angiogenesis (LYN, fold change = 2.841; CD93, fold change = 2.844; and PECAM1 (CD31), fold change = 6.25; **Supplemental Figure**

4.6)[554–556]. Overrepresentation analysis of the AS-specific DEPs highlighted leukocyte activity and cell adhesion as two features significantly enriched. In DDLPS, 252 DEPs were upregulated, and 174 downregulated relative to all other samples (**Figure *4.6*B**). Of these, 47 were uniquely upregulated in DDLPS (**Figure *4.6*G**). These included CDK4 (fold change = 13.14), which is known to be amplified in DDLPS (**Supplemental Figure *4.7***)[4]. Equally, CPM is also reported as amplified in DDLPS by some studies and is uniquely upregulated herein (fold change = 1.619; **Supplemental Figure *4.7***)[216]. Notably, the expression of MDM2, also commonly amplified in DDLPS, was not captured within the proteomic data. Over-representation analysis of the DDLPS-specific DEPs identified no significant findings. Considering the limited number of DEPs this is

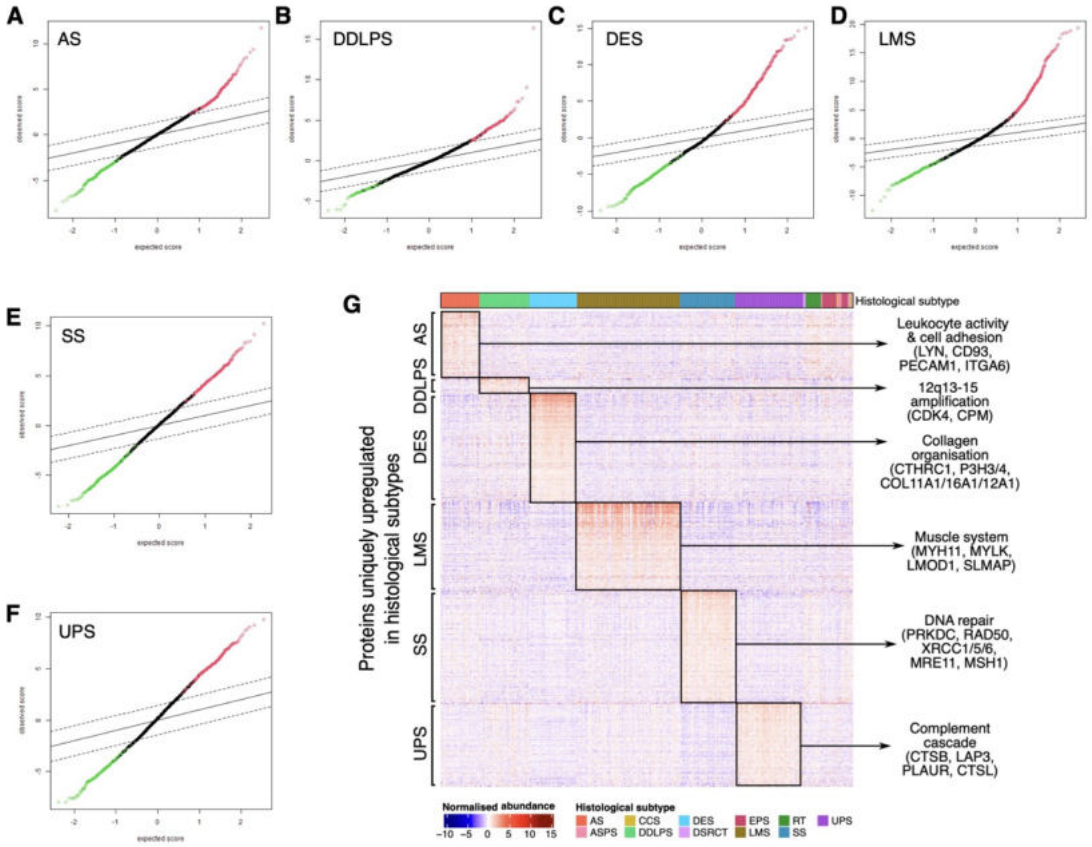


Figure 4.6 Proteomic features of soft tissue sarcoma (STS) histological subtypes

(A-F) Significant analysis of microarray (SAM) 2-class unpaired plots for angiosarcoma (AS; **A**), dedifferentiated liposarcoma (DDLPS; **B**), desmoid tumour (DES; **C**), leiomyosarcoma (LMS; **D**), synovial sarcoma (SS; **E**), and undifferentiated pleomorphic sarcoma (UPS; **F**) compared to the rest of the cohort. Each point is a protein. Proteins within the dashed lines have an FDR ≥ 0.01 and therefore are not significantly differentially expressed proteins (DEPs). Proteins in red are significantly upregulated DEPs (fold change ≥ 1.5) in the subtype, and proteins in green are significantly downregulated DEPs (fold change < 0.667) in the subtype. **(G)** Heatmap showing the proteins (n=1362) uniquely upregulated in histological subtypes (FDR < 0.01, fold change ≥1.5), sorted by histology. Annotations indicate key proteins (DDLPS & SS) and gene sets identified by overrepresentation analysis (AS, DES, LMS, UPS).

unsurprising. In DES, 594 DEPs were upregulated, and 633 downregulated relative to all other samples (**Figure *4.6*C**). Of these, 308 were uniquely upregulated in DES (**Figure *4.6*G**). Consistent with DES being characterised as a highly fibrotic tumour with abundant deposits of ECM, DEPs included several collagen chains (COL11A1/12A1/16A1), as well as P3H3 and P3H4, enzymes responsible for collagen hydroxylation and cross-linking (**Supplemental Figure *4.8***)[557]. In agreement with this, over-representation analysis highlighted 'collagen organisation' as enriched. In LMS, 378 DEPs were upregulated, and 412 downregulated relative to all other samples (**Figure *4.6*D**). Of these, 256 were uniquely upregulated in LMS (**Figure *4.6*G**). In agreement with the LMS cell of origin being smooth muscle, many of the DEPs were muscle-specific proteins[558]. These included MYH11 (fold change = 26.52), MYLK (fold change = 25.847), LMOD1 (fold change = 19.526), and SLAMP (fold change = 18.741; **Supplemental Figure *4.9***). As such, over-representation analysis revealed enrichment of 'muscle system' ontologies. In SS, 475 DEPs were upregulated, and 508 downregulated relative to all other samples (**Figure *4.6*E**). Of these, 322 were uniquely upregulated in SS (**Figure *4.6*G**). Among the DEPs uniquely upregulated were those involved in DNA double strand repair (**Figure 1.4A**), such as PRKDC (fold change = 2.498), RAD50 (fold change = 2.302), XRCC5/6 (fold change = 2.184 and 2.058), and MRE11 (fold change = 1.868; **Supplemental Figure *4.10***). This is in line with previously reported changes in the DNA repair activity of SS tumours[559,560]. Despite the seeming consistent upregulation of several DNA repair proteins in SS, over-representation analysis did not identify any biological pathways as significantly upregulated. This may be due to poor overall coverage of certain gene sets within the proteomic data. Finally, in UPS, 433 DEPs were upregulated, and 455 downregulated relative to all other samples (**Figure *4.6*F**). Of these, 238 were uniquely upregulated in UPS (**Figure *4.6*G**). Notably, cathepsin-B, -D, -L, and -Z (CTSB/D/L/Z), key modulators of immune response, and PLAUR, a promoter of plasmin formation, were upregulated (**Supplemental Figure *4.11***)[561]. In agreement with this, over-representation noted and enrichment of the 'complement cascade', a pathway that amplifies immune response[562]. Overall, the subtype specific proteomic findings confirm expected observations for each subtype queried.

To assess the expression of subtype-specific DEPs in an independent dataset, the TCGA RPPA data was analysed[36]. Of the subtypes profiled herein, TCGA also assessed LMS (n = 80), DDLPS (n = 50), UPS (n = 44), and SS (n = 10) using the RPPA platform. Of the proteins identified by MS-based proteomics as upregulated and unique to these subtypes, 13 (7 SS-specific, 3 LMS-specific, and 3 UPS-specific) were also assessed

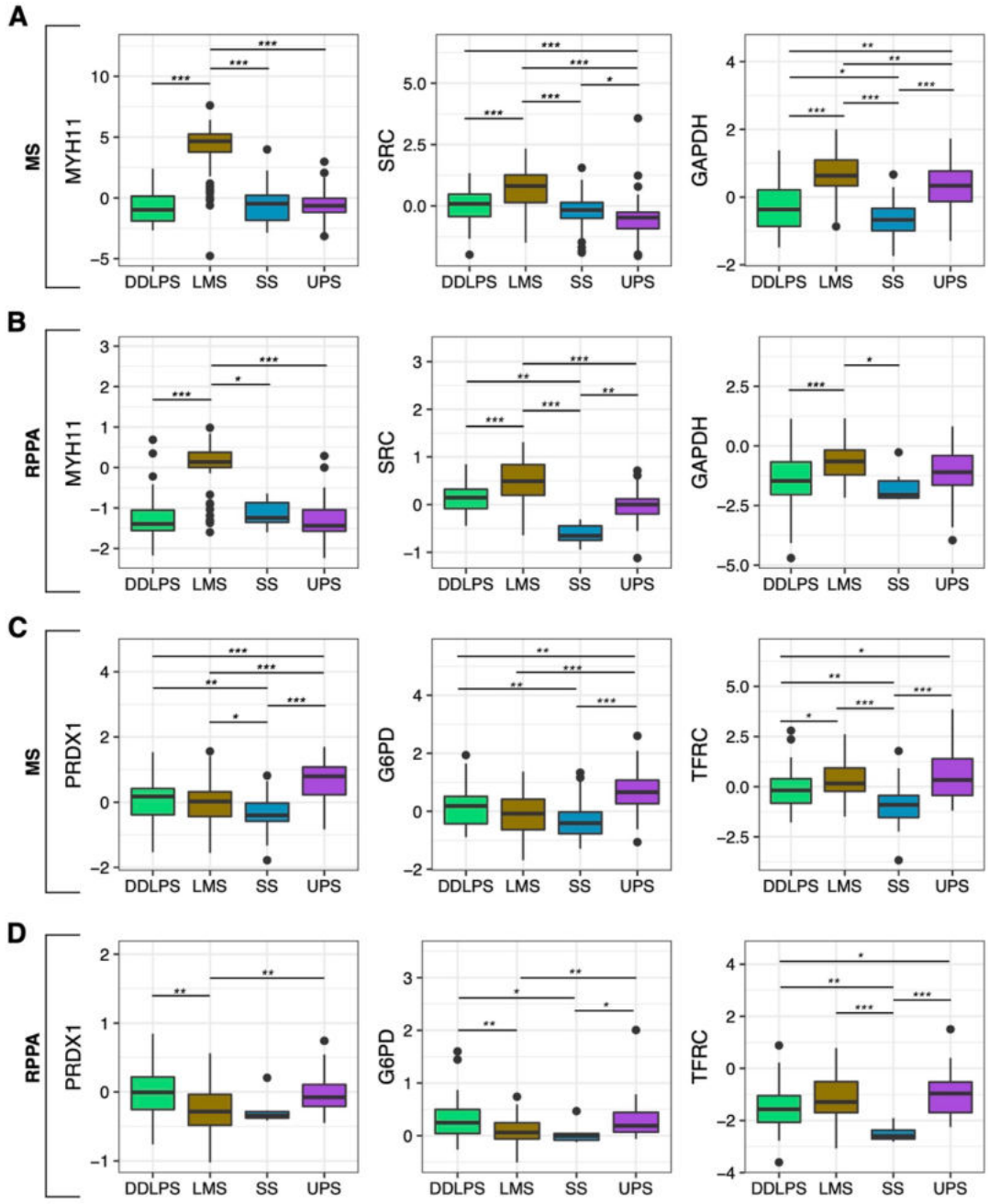


Figure 4.7 Validation of subtype-specific enriched proteins. Boxplots showing the normalised abundance of proteins uniquely upregulated in **(A-B)** leiomyosarcoma (LMS) and **(C-D)** undifferentiated pleomorphic sarcoma (UPS) compared to dedifferentiated liposarcoma (DDLPS), and synovial sarcoma (SS) in two independent cohorts. **(A,C)** Normalised protein abundance based on tandem mass tag (TMT) mass spectrometry (MS) data from the cohort herein. **(B,D)** Normalised protein abundance based on the reverse-phase protein array (RPPA) data from The Cancer Genome Atlas (TCGA) Sarcoma cohort. Boxes indicate 25th and 75th percentile, with median line in the middle, whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. Significance determined by Dunn's tests, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.

within the RRPA data. Given the low number of SS patients profiled as part of the TCGA cohort, the SS-specific proteins were not assessed. The LMS-specific proteins assessed were MYH11, GAPDH, and SRC. The UPS-specific proteins assessed were PRDX1, G6PD, and TFRC. Given, these subtype-specific proteins were derived from proteomic data comparisons performed against all other subtypes profiled herein; the MS-based proteomic data for TCGA-included subtypes was assessed alongside the RPPA data (against LMS, DDLPS, UPS, and SS only). Comparisons were reperformed using Kruskal-Wallis and post-hoc Dunn tests. Within both the MS and RPPA data, Kruskal-Wallis tests identified all proteins to show significantly different expression based on histological subtypes (**Supplemental Table *4.4***, **Supplemental Table *4.5***). Post-hoc Dunn tests in both the MS and RPPA data, revealed MYH11 and SRC as significantly enriched in LMS relative to all other subtypes (**Figure 4.7A-B, Supplemental Table *4.6***, and **Supplemental Table *4.7***). RPPA revealed GAPDH as present at a significantly higher level in LMS compared to DDLPS and SS, but not UPS (**Figure 4.7B** and **Supplemental Table *4.7***); whilst the MS data found high GAPDH in LMS relative to all other subtypes (**Figure 4.7A** and **Supplemental Table *4.6***). Similarly, post-hoc tests illustrated the UPS-specific proteins to be inconsistently observed as enriched in UPS. In the MS data, PRDX1 and G6PD were significantly enriched in UPS compared to all other subtypes, and TRFC was significantly enriched in UPS compared to DDLPS and SS (**Figure 4.7C** and **Supplemental Table *4.6***). However, in the RPPA data: PRDX1 was significantly higher in UPS compared to LMS; TFRC was significantly higher in UPS compared to SS and DDLPS; and G6PD was significantly higher in UPS compared to LMS and SS (**Figure 4.7D** and **Supplemental Table *4.7***). Overall, subtype comparisons based on TCGA RPPA data largely recapitulated the previous SAM-based findings; validating LMS- and UPS-enriched proteins. Minor discrepancies in the comparison of UPS-specific proteins were highlighted and may be resultant of methodological differences between MS measures and antibody measures in RPPA. This illustrates one of the central difficulties in validating MS-based proteomic STS research; where no comparable independent MS datasets are publicly available.

4.2.3.3 An overview of the sub-proteomes of STS

To characterise the composition of the identified proteome, several publicly available databases were queried covering the matrisome (n = 1,027), adhesome (n = 232), immune component (n = 2,483), and kinome (n = 516)[500,501,503,545].

As mesenchymal tumours, STS are hypothesised to deposit excessive matrisomal proteins[563]. Matrisomal proteins encompass ‘core matrisome’ components, which form

the structural basis of the matrisome, as well as 'matrisome-associated' proteins, which cooperate to remodel ECM and modulate matrisome activity. Together, the core and associated matrisome provide architectural support to tissues and facilitate extracellular-intracellular signalling. At present there is little knowledge surrounding the STS matrisome. However, in other cancer types the matrisome is known to modulate disease progression, and dictate response and resistance to treatment[564–567]. The proteomic data was therefore assessed for matrisomal content using the MatrisomeDB database[501]. Coverage of the matrisome was 19% (**Figure *4.8*A**). This encompassed 34% of the core matrisomal proteins and 13% of the associated matrisomal proteins (**Figure *4.8*B**). Unsupervised clustering showed highly similar patterns to proteome-wide data, with the main notable difference being a weaker grouping of UPS tumours (**Figure *4.8*C**). Inspection of the heatmap illustrated generally higher levels of matrix proteins in DES compared to other subtypes. Given the fibrotic nature of DES this is unsurprising. Matrisome proteins strongly expressed in DES included collagen chains (e.g., COL1A1/2, COL2A1, COL3A1, and COL5A1/2/3) and many other core matrisome components (e.g., FBLN1 and PCOLCE). LMS also showed robust high expression of a subset of matrisome proteins. These included glycoproteins such as laminins (LAMA4/A5/B1/B2/C1) and nidogens (NID1/2), as well as 2 type-IV collagen chains (COL4A1 and COL4A2). Interestingly, type-IV collagen, laminins and nidogens are essential constituents of basement membrane (BM)[568,569]. The BM is a specialised and networked type of ECM, which structurally separates tissues[570]. By contrast, the DES-upregulated matrisome components are not BM-specific but rather comprise mostly fibrillar collagen chains which provide tissue strength[571]. This illustrates the presence of structurally and compositionally distinct matrisomes between STS subtypes.

Tumour cells interact with the matrisome through adhesion molecules. The interaction of adhesion receptors with matrisomal proteins triggers intracellular signalling pathways and cellular responses tailored to the external cell environment[572]. Therefore, the representation of adhesome proteins in the proteomic data was assessed using the function atlas of the integrin adhesome[545]. The adhesome showed high coverage (47%; **Figure *4.9*A**). The bulk of these proteins comprised adapter proteins (coverage = 36%), adhesion receptors (coverage = 18%), and actin regulatory proteins (coverage = 14%; **Figure *4.9*B**). Unsupervised clustering of the adhesome revealed similar patterns to whole proteome clustering (**Figure *4.9*C**). LMS, DES, and SS presented with distinct adhesome components, and were grouped in in robust histology-specific clusters. Notably, the proteins most highly expressed were seen as enriched in LMS. In agreement with the LMS matrisome profile, these included several integrin subunits,

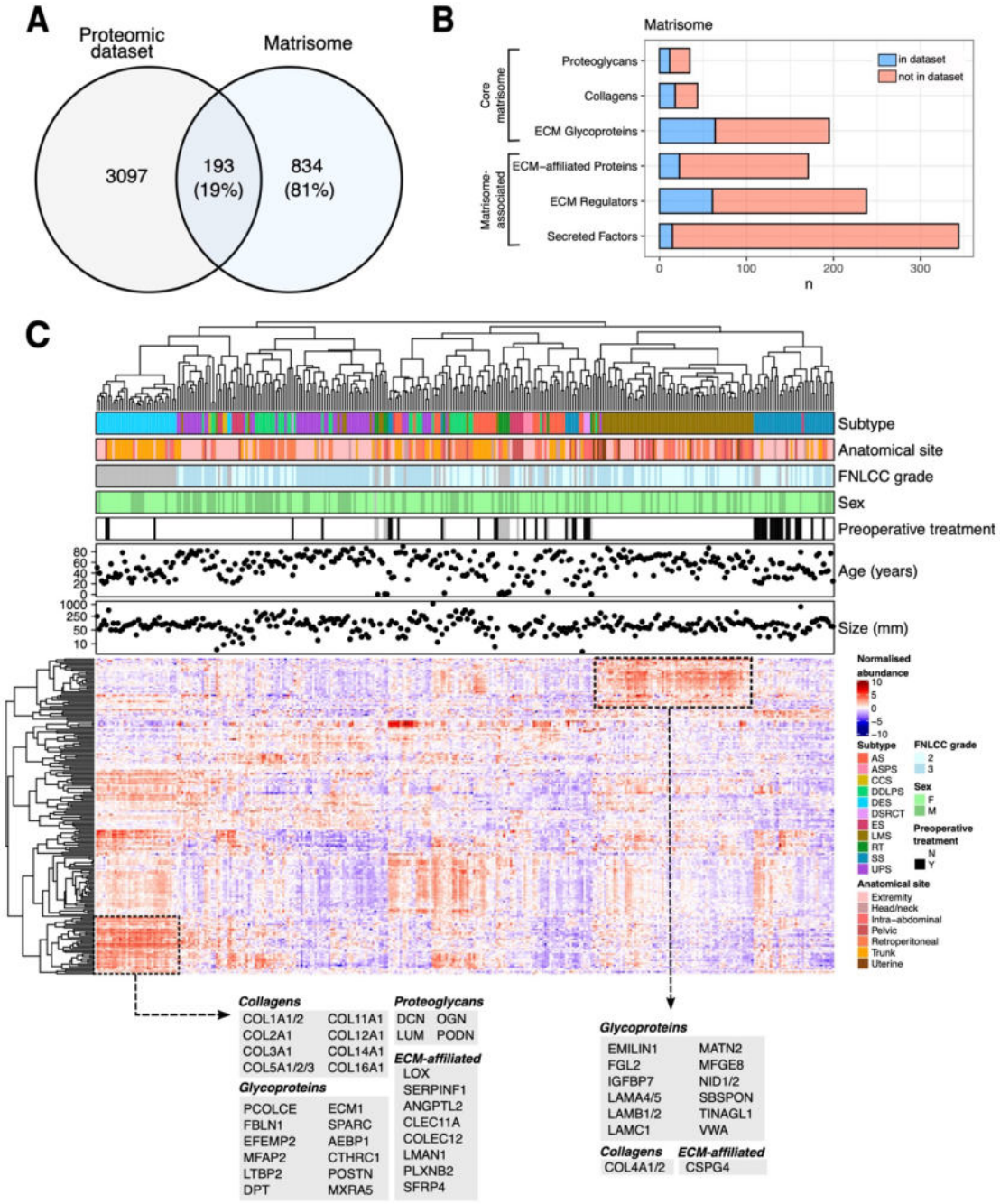


Figure 4.8 The matrisome landscape of soft tissue sarcoma (STS).

(A) Venn diagram showing the overlap between the proteomic dataset and the matrisome database. **(B)** Stacked bar chart showing the representation of each category of matrisome proteins in the proteomic dataset. **(C)** Annotated heatmap showing the unsupervised clustering (Pearson's distance) of 193 matrisome proteins across the study cohort. Regions of interest highlighted with black dashed boxes, and proteins within listed. From top to bottom, panels indicate histological subtype, anatomical site, tumour grade, patient sex, preoperative treatment status, patient age, and tumour size. Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DES = desmoid tumour; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; RT = rhabdoid tumour; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma; ECM = extracellular matrix.

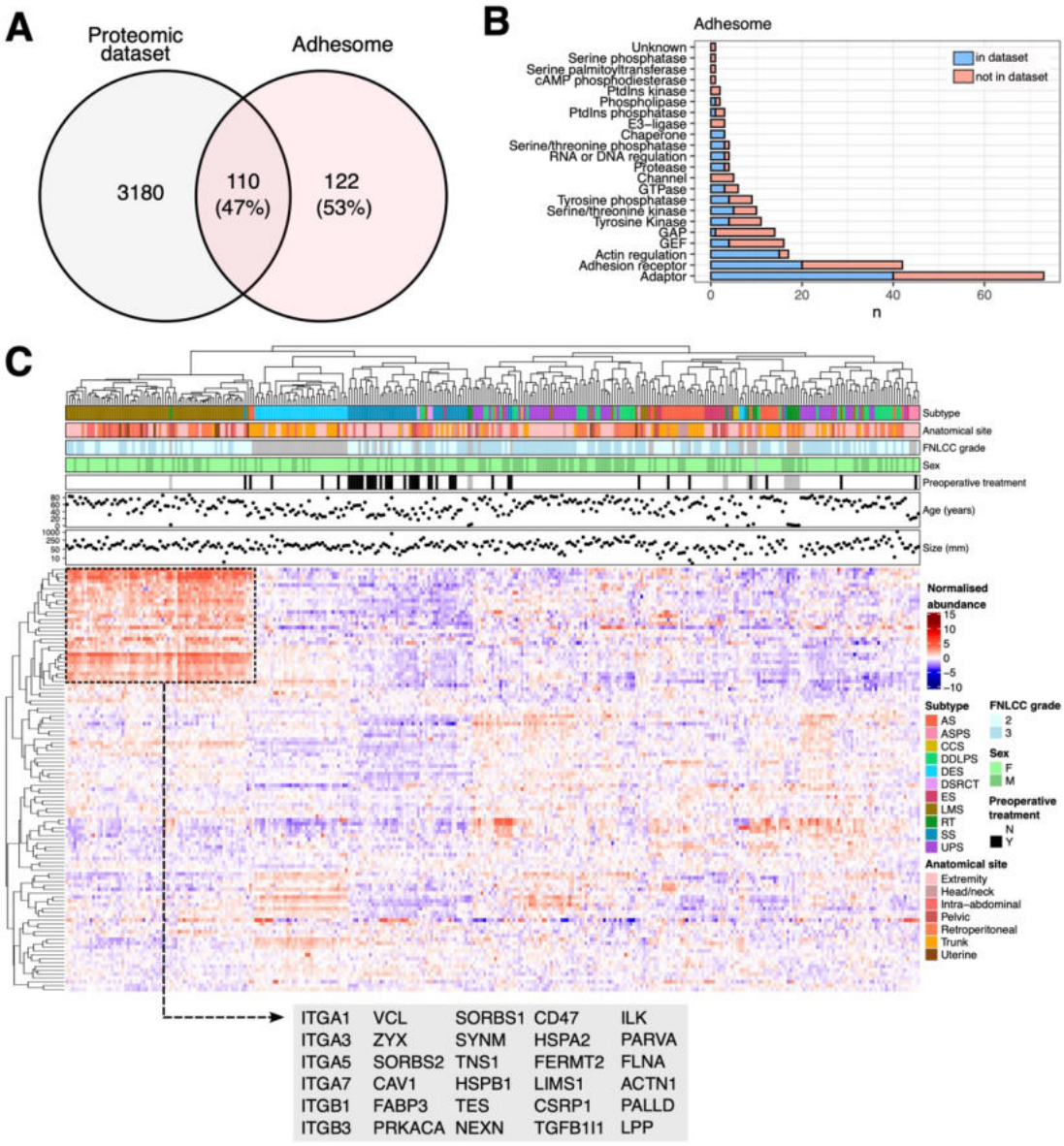


Figure 4.9 The adhesome landscape of soft tissue sarcoma (STS).

(A) Venn diagram showing the overlap between the proteomic dataset and the adhesome database. **(B)** Stacked bar chart showing the representation of each category of adhesome proteins in the proteomic dataset. **(C)** Annotated heatmap showing the unsupervised clustering (Pearson's distance) of 110 adhesome proteins across the study cohort. Regions of interest highlighted with black dashed boxes, and proteins within listed. From top to bottom, panels indicate histological subtype, anatomical site, tumour grade, patient sex, preoperative treatment status, patient age, and tumour size. Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DES = desmoid tumour; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; RT = rhabdoid tumour; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma; GEF = guanine nucleotide exchange factor; GAP = GTPase-activating proteins.

which as dimers (α1β1, α3β1, α7β1) interact with the BM laminins[573]. As well as recapitulating proteome-wide observations, focused interpretation of the adhesome provided a different perspective and thus new insights. Whilst proteome-wide data showed UPS as a relatively homogeneous group clustered together, the adhesome revealed UPS to cluster in 2 spatially distinct regions of the heatmap. Similarly, whilst AS showed extensive heterogeneity in the proteome-wide data, use of the adhesome-level data identifies AS near exclusively grouped as 2 clusters. This suggests subsets of AS tumours share adhesome biology. Indeed, over-representation analysis of AS DEPs did identify cell adhesion as a feature uniquely upregulated in AS (**Section 4.2.3.2**).

In addition to the matrisome, another key constituent of the tumour microenvironment (TME) is the immune component. Given the central role of immune cells in STS biology (discussed in **section 1.3.1.2**), and the promise ICB have shown in a subset of STS patients (discussed in **section 1.2.3.4**), the proteome was assessed for immune specific proteins using the ImmPort database[500]. The proteomic data covered 13% of the immune component, with most immune proteins mapping to classifications of antimicrobial (coverage = 29%), antigen procession and presentation (coverage = 10%), or cytokine (coverage = 64%; **Figure *4.10*A-B**). Unsupervised clustering showed similar profiles for LMS, DES, and UPS to the proteome-wide data (**Figure *4.10*C**). Although, more mixing of DDLPS and UPS within clusters was observed here, and an increase in the number of LMS tumours clustering cluster away from the main LMS-specific cluster was seen. As in the adhesome level data, nearly all AS separated into 2 very distinctive groups. Both AS groups were within multi-subtype clusters, one mostly with subsets of DDLPS and LMS, and the other mostly with subsets of EPS, UPS, and SS. Immune protein abundance appeared higher in the AS, DDLPS, LMS mixed cluster, which specifically showed high expression of immunoglobulins and complement proteins. This alludes to the identification of an immune high population that spans multiple subtypes. Moreover, this reinforces LMS specific studies noting high immune activity in a subset of patients and agrees with the favourable responses to ICB seen in subsets of DDLPS patients[139,140].

Kinases are central to many oncogenic pathways, and thus are the targets of many anti-cancer drugs (discussed in **section 1.2.3.3**). Accordingly, kinase representation in the proteomic data was assessed using the protein kinase complement characterised by Manning *et al*[503]. Of all databases queried, the kinome was the poorest represented within the proteomic data (7%; **Figure *4.11*A**). Best coverage was seen for the serine/threonine-specific protein kinases (STE; 18%) and CMGC (cyclin-dependent

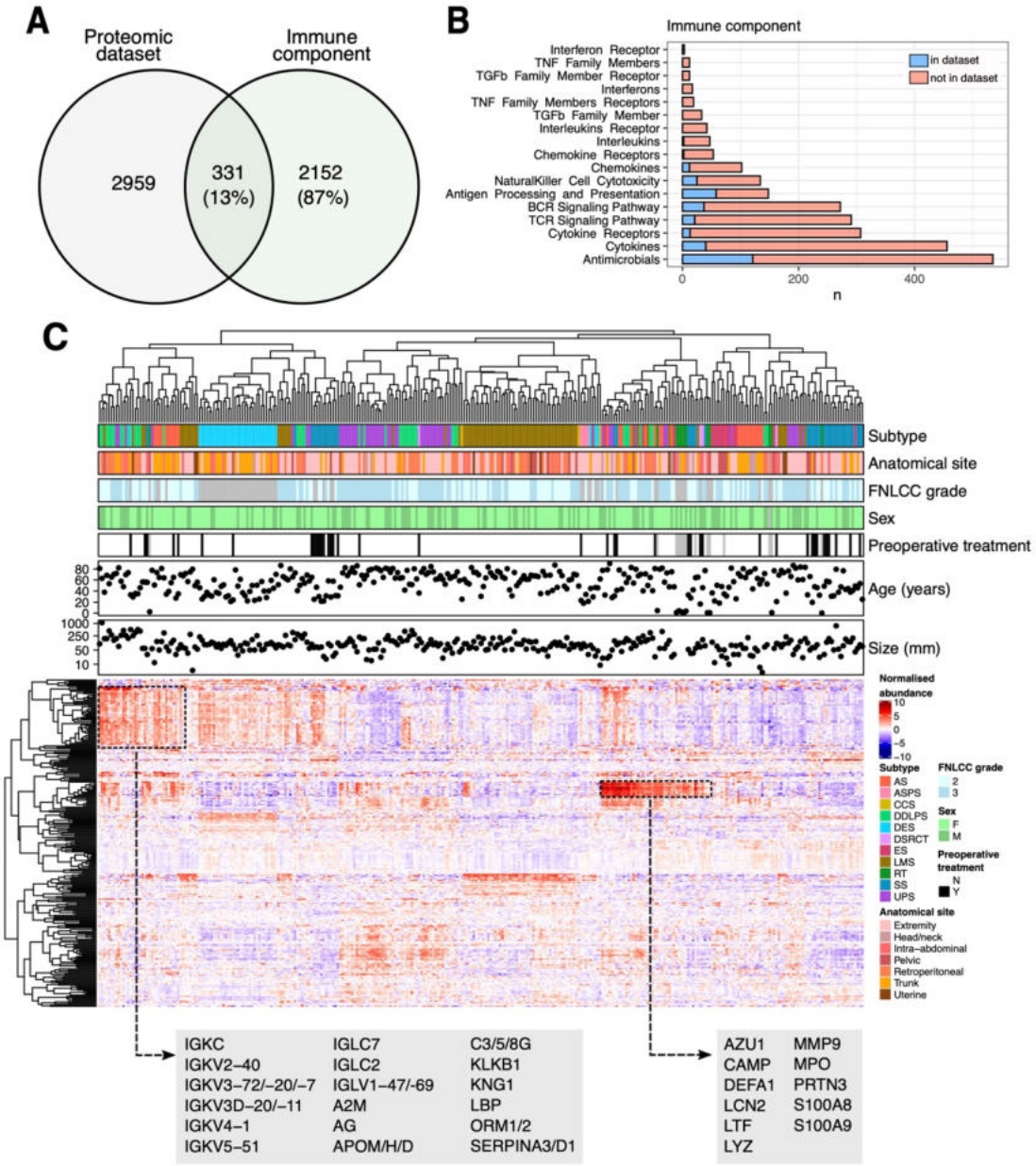


Figure 4.10 The immune landscape of soft tissue sarcoma (STS).
(A) Venn diagram showing the overlap between the proteomic dataset and the immune database. **(B)** Stacked bar chart showing the representation of each category of immune proteins in the proteomic dataset. **(C)** Annotated heatmap showing the unsupervised clustering (Pearson's distance) of 331 immune proteins across the study cohort. Regions of interest highlighted with black dashed boxes, and proteins within listed. From top to bottom, panels indicate histological subtype, anatomical site, tumour grade, patient sex, preoperative treatment status, patient age, and tumour size. Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DES = desmoid tumour; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; RT = rhabdoid tumour; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma; TNF = tumour necrosis factor; TGFβ = transforming growth factor β; TCR = T cell receptor; BCR = B cell receptor.

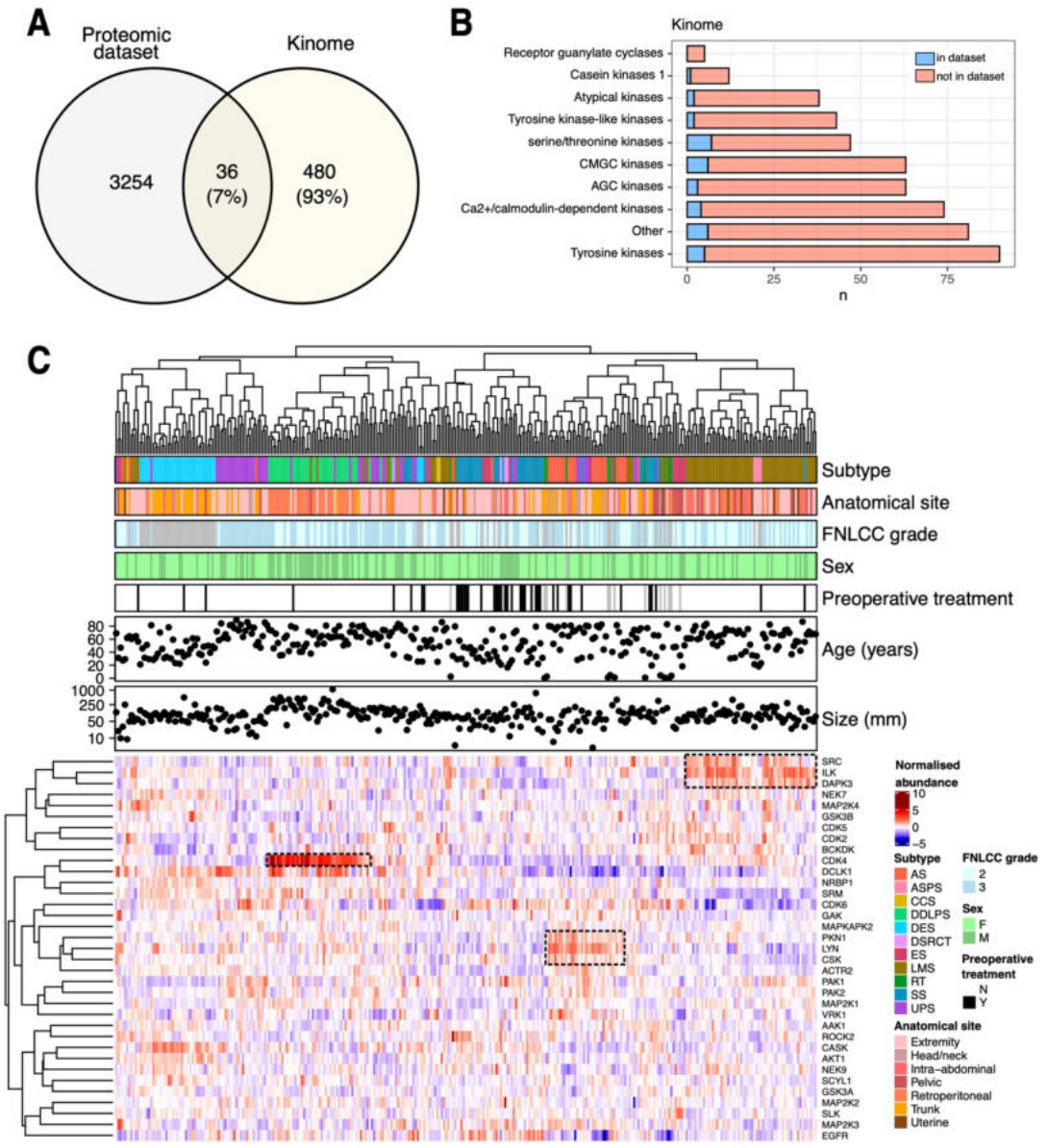


Figure 4.11 The kinome landscape of soft tissue sarcoma (STS).
(A) Venn diagram showing the overlap between the proteomic dataset and the kinome database. **(B)** Stacked bar chart showing the representation of each category of kinome proteins in the proteomic dataset. **(C)** Annotated heatmap showing the unsupervised clustering (Pearson's distance) of 36 kinases across the study cohort. Regions of interest highlighted with black dashed boxes. From top to bottom, panels indicate histological subtype, anatomical site, tumour grade, patient sex, preoperative treatment status, patient age, and tumour size. Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DES = desmoid tumour; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; RT = rhabdoid tumour; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma.

kinases (CDK), mitogen-activated protein kinases (MAPK), glycogen synthase kinases (GSK) and CDK-like kinases) kinase group (10%; **Figure *4.11*B**). As in the proteome-wide data, unsupervised clustering showed distinct DES and LMS profiles (**Figure *4.11*C**). Proteins of note illustrated as highly expressed in LMS included integrin-linked kinase (ILK) and SRC. ILK is a β1 integrin signal transducer, notable due to the strong β1 integrin and BM signature observed in LMS, whilst SRC has previously been reported as a marker to discriminate LMS from UPS[574]. The kinome also revealed common AS biology. In all previous assessments, AS showed a consistently heterogeneous profile. Yet using the kinome, AS clustered together, and appeared to show enrichment of protein kinase N1 (PKN1), c-terminal Src kinase (CSK), and LYN. Therefore, despite low coverage of the complete kinome, this dataset was able to identify candidate kinase biology characteristic of the highly heterogeneous AS subtype. Similarly, robust clustering of DDLPS was seen; unsurprisingly driven by abundant CDK4.

4.2.3.4 An overview of the biological features of STS

To explore overarching biological features within the proteomic dataset, single sample GSEA (ssGSEA) was performed. ssGSEA queries pre-defined gene sets such as those available within the Molecular Signature Database (MsigDB) to calculate an enrichment score for each sample corresponding to each gene set[504,509,512]. This enrichment score is a measure of the coordinated expression of genes within a gene set. To achieve a comprehensive understanding of the proteome, ssGSEA was applied to several gene set databases: the GO biological processes (BP), the Hallmarks, and the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [506–508,510].

GO BP is comprised of 17,949 genes and 7,481 gene sets representing a wide array of biological activity[506,507]. GO BP is a comprehensive dataset that is structured hierarchically with broad ‘parent’ ontologies and more specialised ‘child’ ontologies. Unsupervised clustering of the ssGSEA GO BP normalised enrichment scores (NES) highlighted several notable insights. Consistent with protein-level interpretations, DES and SS each showed defined clusters enriched in developmental and ECM processes, and DNA processing and cell cycle activity respectively (**Figure *4.12***). The enrichment of DNA activity and cell cycle processes did not however appear restricted to SS. Enrichment of these gene sets was also observed in a mixed subtype group comprising mostly AS. Consistent with this, recent transcriptomic profiling of AS has noted the enrichment of cell cycle genes in a subset of patients[575]. In contrast to protein-level data, GO BP also revealed 2 distinct clusters of LMS. Whilst not formally interrogated it appeared that each cluster showed differential immune activity. Broad enrichment of

immune gene sets (inflammatory and cellular responses) was notable in 1 group of LMS, whilst absent in the other. Indeed, LMS transcriptomic subtypes have been reported to show immune variance[36,43,274,283]. Differences in immune activity were observed across the cohort. Yet interestingly this is not simply a presence and absence. There appear 2 groups of immune related gene sets showing discordant expression across the cohort: 1 mapping to inflammatory responses, and 1 to cellular response. This supports the concept of differential immune activation across STS and is in agreement with literature characterising STS into immune subtypes with different features[136,231] (**section 1.3.1.2**).

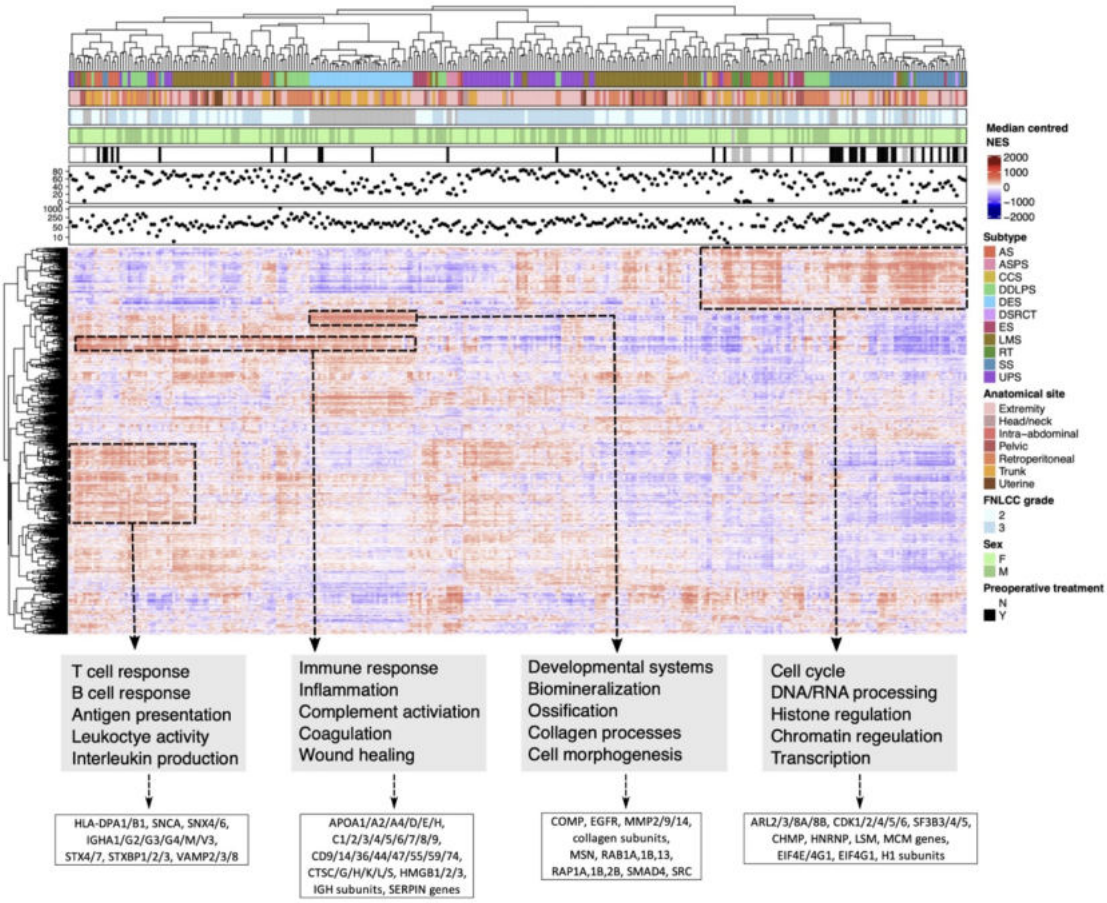


Figure 4.12 Gene ontology biological processes (GO BP) landscape of soft tissue sarcoma (STS). **(A)** Annotated heatmap showing the unsupervised clustering (Pearson's distance) of 2267 GO BP gene sets across the study cohort. Regions of interest highlighted with black dashed boxes and annotated to provide an overview of the gene sets within, and the proteins within those gene sets. From top to bottom, panels indicate histological subtype, anatomical site, tumour grade, patient sex, preoperative treatment status, patient age, and tumour size. Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DES = desmoid tumour; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; RT = rhabdoid tumour; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma.

Within the GO BP gene sets, attempts to cover biology as comprehensively as possible have introduced significant redundancy. This often results in repetitive annotations of datasets, from which interesting findings can be challenging to decipher. Complementary

to GO BP, and more streamlined, are the Hallmarks. The Hallmarks comprise only 4,383 genes and 50 gene sets[508]. They concisely describe key biological activity and were constructed through the refinement of multiple other gene sets ('founder' sets). Application of the Hallmarks against the proteomic dataset and subsequent unsupervised hierarchical clustering revealed a starkly different profile to the protein-level data and GO BP features (**Figure *4.13***). The only consistent finding was a robust separation of DES from the rest of the cohort; seemingly driven by an enrichment of

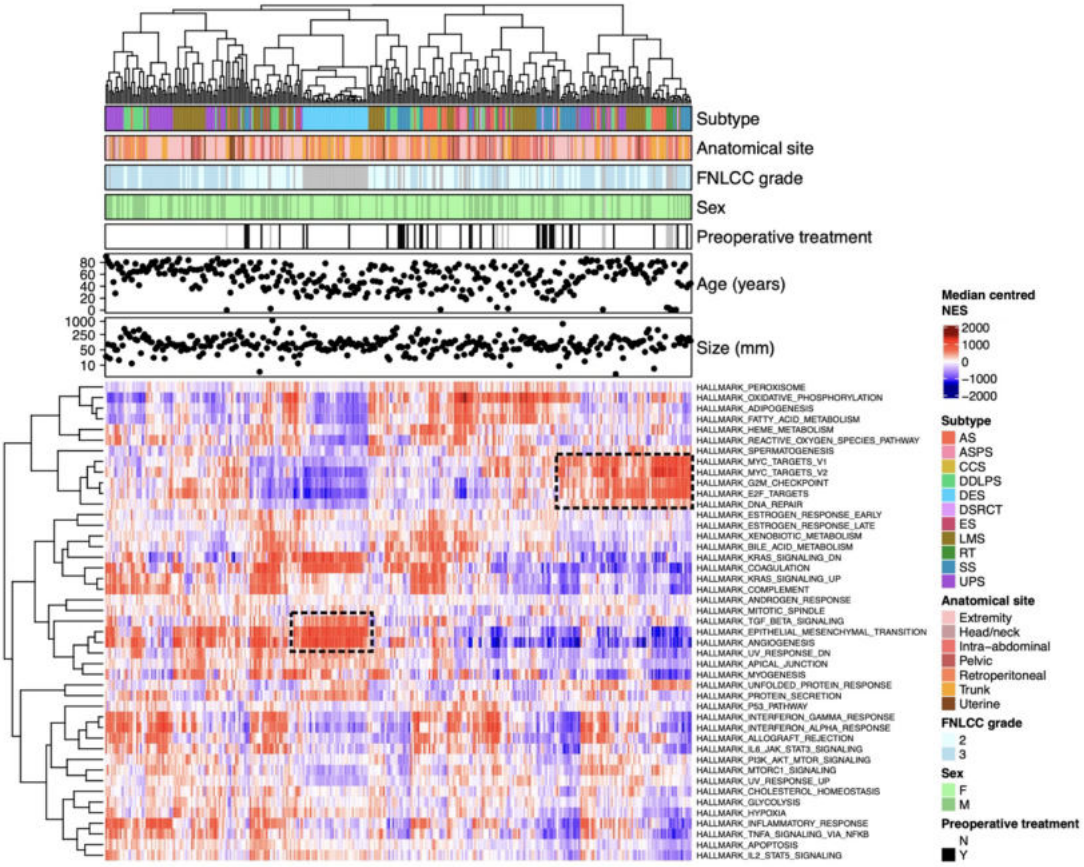


Figure 4.13 Hallmark landscape of soft tissue sarcoma (STS).
(A) Annotated heatmap showing the unsupervised clustering (Pearson's distance) of 45 Hallmark gene sets across the study cohort. Regions of interest highlighted with black dashed boxes. From top to bottom, panels indicate histological subtype, anatomical site, tumour grade, patient sex, preoperative treatment status, patient age, and tumour size. Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DES = desmoid tumour; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; RT = rhabdoid tumour; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma.

TGFβ signalling, EMT, and angiogenesis. Beyond this, huge inter- and intra-subtype heterogeneity was revealed; even for the typically well-defined LMS and SS subtypes. For example, one group of Hallmarks which showed high pan-subtype (SS, RT, AS, DDLPS, LMS, UPS) enrichment were the MYC and E2 factor (E2F) targets, and G2/M

checkpoint. These hallmarks are proliferative signatures associated with cell cycle activity. The subtype-independent enrichment of these scores highlights common biology within a subset of STS patients. Excessive proliferation supports rapid tumour growth which can confer aggressive tumour behaviour. Indeed, MYC deregulation and enrichment of the G2/M hallmark have been found to be associated with poor clinical outcome in other cancer types[576–578]. In STS, grade is a clear and consistent prognosticator for outcome (**section 1.2.2.1** and **section 4.2.2**), thus it was hypothesised that tumours showing high MYC and G2/M activity were of higher grade. However, inspection of the clustering showed no apparent relationship between high enrichment of these hallmarks and high grade tumours.

KEGG comprises 5,245 genes within 186 gene sets[510]. In contrast to GO BP and Hallmarks which are collections of gene sets describing broad biological activities, KEGG is a pathway database. Each KEGG gene set details a group of genes that exist within the same pathway or process. Unsupervised hierarchical clustering of the ssGSEA KEGG NES highlighted UPS as high in immune activity, in agreement with previous findings herein and in published literature[36,220] (**Figure *4.14***). Analyses also illustrated robust clustering of LMS and DES as individual subtypes. However, within the LMS cluster evident heterogeneity was revealed. This heterogeneity was particularly seen in the enrichment level of metabolic pathways, consistent with a reported LMS subtype with metabolic enrichment[281].

4.2.3.5 An overview of drug target profiles in STS

Current treatment for advanced STS is predominantly structured as a “one size fits all” approach, whereby molecular heterogeneity is not considered (**section 1.2.3)**. Low response rates across the STS population illustrate a need for targeted approaches to treatment. Therefore, to assess whether the STS proteome can reveal candidate drugs for personalised treatment, the Drug Signature Database (DSigDB) was queried and ssGSEA performed as before[511]. DSigDB is a collection of gene sets which correspond to drug targeting profiles. It is categorised into 4 levels: D1 approved drugs; D2 kinase inhibitors; D3 perturbagen signatures; and D4 computational drug signatures. To reveal candidates with high clinical applicability, analysis was restricted to D1. Unsupervised clustering of the D1 NES revealed extensive heterogeneity within histological subtypes (**Figure *4.15***). Notably, there was no apparent correlation between pre-operative treatment status and the drug target profile of these tumours. Overall, clustering appeared influenced by vincristine, podophyllotoxin, paclitaxel, and vinblastine. Except for the tubulin inhibitor podophyllotoxin, these are anti-neoplastic drugs[579–581]. The target

profiles of these drugs showed high enrichment in DES and a subset of LMS, and strikingly low enrichment in a mixed group of mostly AS and DDLPS. Paclitaxel is a first-line treatment of choice for AS, thus it is highly surprising that AS tumours show low abundance of the proteins targeted by paclitaxel[104]. Heterogeneity of AS may explain this, as responses to paclitaxel are not universally observed in all AS patients[582,583]. Alternatively, the restricted proteome coverage compared to the genome may be limiting interpretation. Indeed, of the 11 genes within the paclitaxel target profile, only 3 are captured within the proteomic data. Notably, TKIs (gefitinib, bosutinib, sunitinib, crizotinib, nilotinib, dasatinib, vandetanib, axtinib, and sorafenib) cluster together and all show similar levels of heterogeneity and subtype-independence across the cohort **Figure *4.15***). In TKI clinical trials in STS, poor ORR are frequently observed (as discussed in **section 1.2.3.3**). It has been hypothesised that heterogeneity in mixed

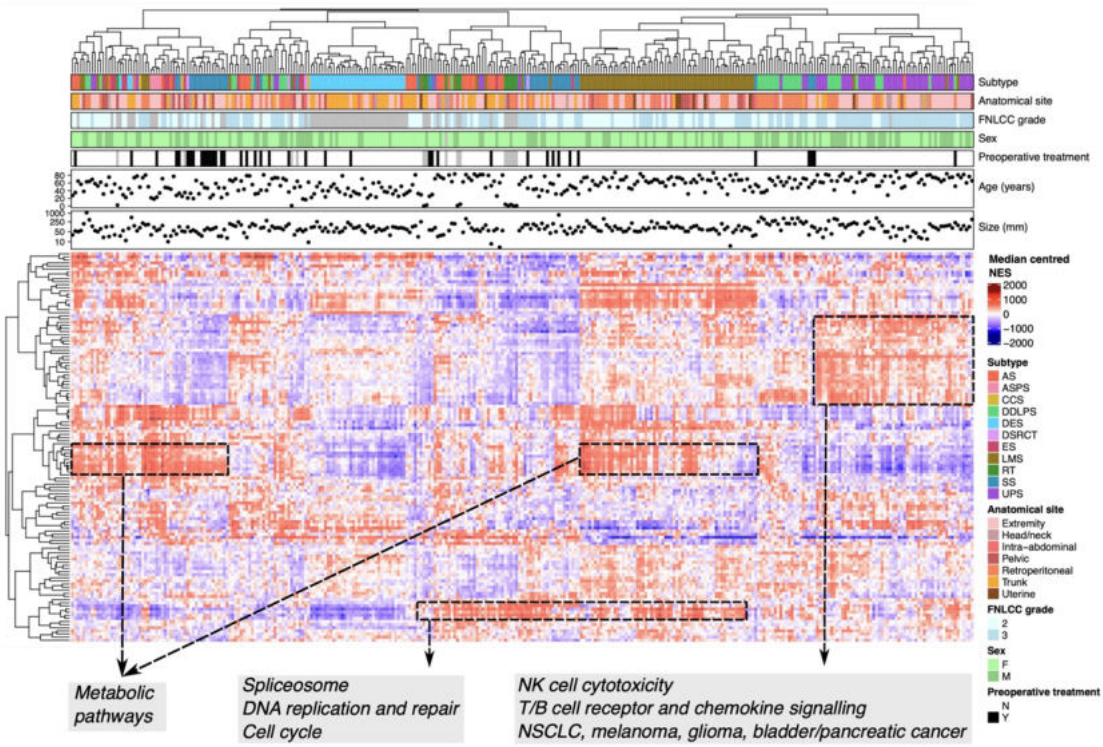


Figure 4.14 Kyoto encyclopaedia of genes and genomes (KEGG) landscape of soft tissue sarcoma. **(A)** Annotated heatmap showing the unsupervised clustering (Pearson's distance) of 125 KEGG gene sets across the study cohort. Regions of interest highlighted with black dashed boxes and annotated to provide an overview of the gene sets within. From top to bottom, panels indicate histological subtype, anatomical site, tumour grade, patient sex, preoperative treatment status, patient age, and tumour size. Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DES = desmoid tumour; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; RT = rhabdoid tumour; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma; NSCLC = non-small cell lung cancer; NK = natural killer.

subtype STS clinical trials masks any ORR benefit that may be observable in a subset of patients. The enrichment of TKIs targeting profiles herein supports this clinically observed heterogeneity in responses across subtypes.

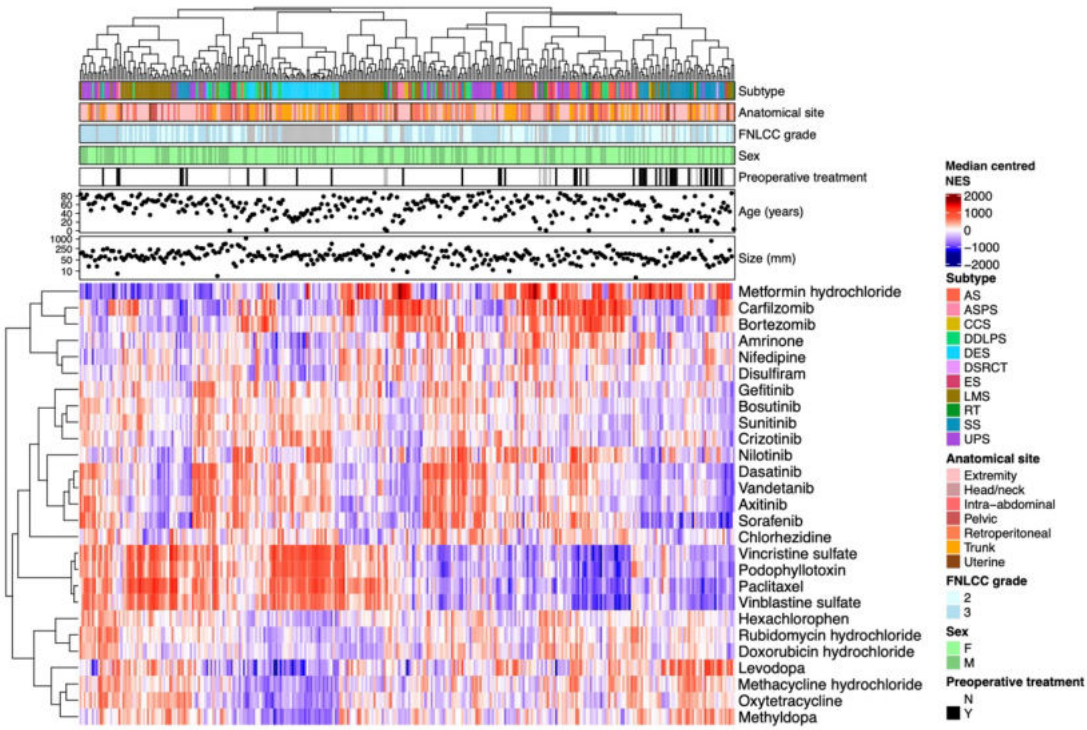


Figure 4.15 Drug target profile expression in soft tissue sarcoma (STS).
(A) Annotated heatmap showing the unsupervised clustering (Pearson's distance) of 27 Drug Signature database (DSigDB) D1 profiles across the study cohort. From top to bottom, panels indicate histological subtype, anatomical site, tumour grade, patient sex, preoperative treatment status, patient age, and tumour size. Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DES = desmoid tumour; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; RT = rhabdoid tumour; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma.

4.3 Discussion and summary

This chapter has presented a proteomic overview of multiple histological subtypes of STS. To date, this represents the largest proteomic profiling effort in STS by far, making the dataset a hugely rich resource.

The profiled cohort had specific and deliberate features. Firstly, to establish the baseline proteomic profile of STS, this cohort was restricted to primary tumours only. This prevented the introduction of heterogeneity resultant of disease stage. Secondly, the cohort was designed to include multiple histological subtypes, reflective of disease incidence[234,235]. The inclusion of more prevalent subtypes (LMS, DDLPS, UPS, SS) in higher numbers enables data for these diagnoses to be utilised for in-depth and

statistically robust assessments. This facilitates analyses that may be applicable to a non-trivial proportion of the STS population. At the other extreme, including ultra-rare tumours (ASPS, CCS, DSRCT, EPS) provides invaluable data for patients with huge unmet need. Research and clinical practice in ultra-rare STS are often founded in data from limited case studies[44]. Thus, providing comprehensive molecular profiling for as few as 3-4 patients with ultra-rare diagnoses is vitally important. In line with standard of care for most primary STS in the UK being surgical resection alone, most of the cohort were treatment naïve[44]. The exceptions to this (DSRCT and SS) are known to routinely receive neoadjuvant therapy[44,584,585]. Also in agreement with literature reports, age was associated with subtype, and retroperitoneal tumours tended to be large at diagnosis[586,587]. This is due to the retroperitoneum having a large potential space which permits tumours to grow undetected. Given the study inclusion criteria herein, the cohort was therefore largely representative of the STS disease population. Features of note that may deviate from some other STS studies were: 1) a underrepresentation of uLMS tumours relative to incidence (9% of the LMS cohort vs 25% of all LMS diagnoses), and 2) an enrichment of high grade tumours, likely reflective of the complex patient caseload seen at RMH, where most samples were sourced, and the inclusion of putative high-grade subtypes (e.g., ASPS, CCS, EPS)[303]. As anticipated, clinicopathological variables showed extensive interactions, particularly with histological subtype. In addition, key clinicopathological features such as tumour grade, histological subtype, and anatomical site were associated with clinical outcome measures, consistent with current literature[45,46]. Thus, as a clinically annotated, representative cohort, the proteomic data generated is of wide-reaching relevance to the STS research and clinical communities.

As well as providing an overview of the profiled cohort, this chapter also covered a top-level interpretation of the STS proteome landscape. The comprehensive dataset was assessed by both unsupervised and supervised methods, revealing subtype-specific proteome features. By leveraging the richness of this dataset, sub-proteomes mapping to key biological entities were also characterised. Further to considering protein-level information, the expression of broader biological features from MSigDB and the targetable profiles of drugs from DSigDB were assessed. Specifically, these were descriptively detailed with reference to histological subtype. Taken together these analyses revealed both known and novel biological features and identified research avenues that warrant further investigation.

The proteome-wide data was shown to strongly associate with histological subtype. Reiterating results of large-scale transcriptomic studies of STS, distinctive molecular

profiles of LMS, DES, and SS were seen, with clustering illustrating each subtype to harbour specific proteome features[36,41,165]. However, there were some 'outlier' cases unexpectedly clustering within the robust subtype-specific clusters of SS, DES, and LMS. One reason for this may be the significant heterogeneity within histological subtypes of STS. Alternatively, these may represent misdiagnoses. Although these patients were diagnosed by experienced histopathologists at a specialist sarcoma centre, STS diagnosis is challenging, and misdiagnosis can still occur. Indeed, other molecular profiling studies, such as the sarcoma methylation classifier and TCGA study have reported reclassification of STS following analysis[36,367].

AS showed the most heterogeneous profile of all subtypes assessed. This may be resultant of the inclusion of both secondary radiation-associated AS and primary sporadic AS. Despite this heterogeneity, supervised analysis did reveal an enrichment of cell adhesion and leukocyte-related activity to be specific to AS tumours. The relevance of cell adhesion to AS tumours is unknown, however an enrichment of an immune process is pertinent, as a subset of AS patients have been shown to respond to ICB intervention[575,588–593]. Also consistent with the known biology of each subtype: DDLPS, characterised by amplification of the *CDK4*-containing genomic *locus*, showed high CDK4 expression; DES, a fibrotic tumour, showed enrichment of ECM processes; LMS, a smooth muscle derived subtype, showed enrichment of muscle related processes; SS, a tumour with increased replication stress, showed enrichment of DNA repair proteins; and UPS showed enrichment of the complement cascade[4,36,220,348,557,559,560]. Interestingly, PLAUR a promoter of plasmin formation was also upregulated in UPS. Plasmin is central to the coagulation pathway, a process highly interconnected with the complement cascade[594–596]. These observations highlight immune activity in UPS and are consistent with previous molecular profiling studies showing UPS as immune-enriched[36,220]. Moreover, this is also in line with clinical trials showing favourable ICB responses in UPS patients[139,140]. This recapitulation of known tumour biology offers reassurances as to the ability of the MS dataset to accurately capture tumour profiles.

The noted matrisomal enrichment in DES is expected given known tumour characteristics[557]. Yet, a surprising enrichment of matrisome components was also observed in LMS. Specifically, LMS were abundant in BM proteins. Furthermore, adhesome data highlighted an enrichment of BM-specific integrins in LMS. In cancer, the BM plays many important roles, such as relaying extracellular signals intracellularly and structurally encapsulating the tumour[569,570,597]. Structurally, the BM prevents local tumour

invasion of adjacent tissues. Therefore, it is hypothesised that the BM-dominant matrisome and adhesome of LMS underlies the relatively low likelihood of local recurrence in LMS. Equally, the lower expression of BM proteins in DES may explain the high locally invasive nature of this subtype[557]. Notably, the BM also functions to prevent metastasis[569,598]. It would therefore be of interest to investigate ECM, and specifically BM, changes in metastatic LMS disease. At present the biological roles of matrisome in LMS has not been explored, thus this illustrates an example of novel biology revealed through MS profiling.

Further to deciphering contrasting features between subtypes of STS, this chapter also highlighted subtype-specific heterogeneity. In proteome-wide data, LMS show the most distinctive proteome with a strong smooth muscle phenotype. Yet, when the data was focused to the immune component, and GO BP and KEGG enrichment profiles generated, LMS heterogeneity was observed. This is consistent with multiple studies suggesting transcriptomic subtypes of LMS exist[36,43,274,281–283]. Notably, the transcriptomic subtypes have been revealed in cohorts restricted to LMS samples. It is striking that herein, where LMS data is relative to other STS, subtypes are also seen. LMS heterogeneity is observed in the context of immune features (immune component, and GO BP), and metabolic features (KEGG). In agreement, the reported transcriptomic subtypes of LMS exhibit differential immune and metabolic activity[36,43,274,281,283]. AS also showed proteomic heterogeneity. In the proteome-wide, adhesome, matrisome, and immune component data, AS cluster poorly and fall broadly into 2-3 heatmap regions. This is also observed in the GO BP, KEGG, and hallmark measures. Inspection suggested the level of immune activity, cell cycle activity, and DNA repair activity may underlie the observed heterogeneity. Supervised analyses noted leukocyte activity as enriched in AS, thus whilst this immune process may be dominant relative to other subtypes, immune activity within AS appears more nuanced. This may offer an explanation as to why ICB responses are only observed in a subset of AS patients[575,588–593].

The inclusion of multiple histological subtypes and co-ordinate analysis in this study supported discovery of pan-subtype biology. For example, immune features (immune, GO BP, KEGG) are repeatedly identified to show differing expression across subtype. This expression showed limited relation to histology and is in line with previous literature[136,231]. In addition, pan-subtype analysis of DSigDB profiles also showed extensive heterogeneity with relation to histological subtype. This is in agreement with many clinical trials in STS and current clinical practice, where highly varied treatment

responses are achieved (as discussion in **section 1.2.3**). Whilst patients with certain subtype diagnoses do respond more favourably to certain interventions (eg. GEM+DOC in uLMS; **section 1.2.3.2)**, trends are not consistent, and responses are seen across histological subtypes.

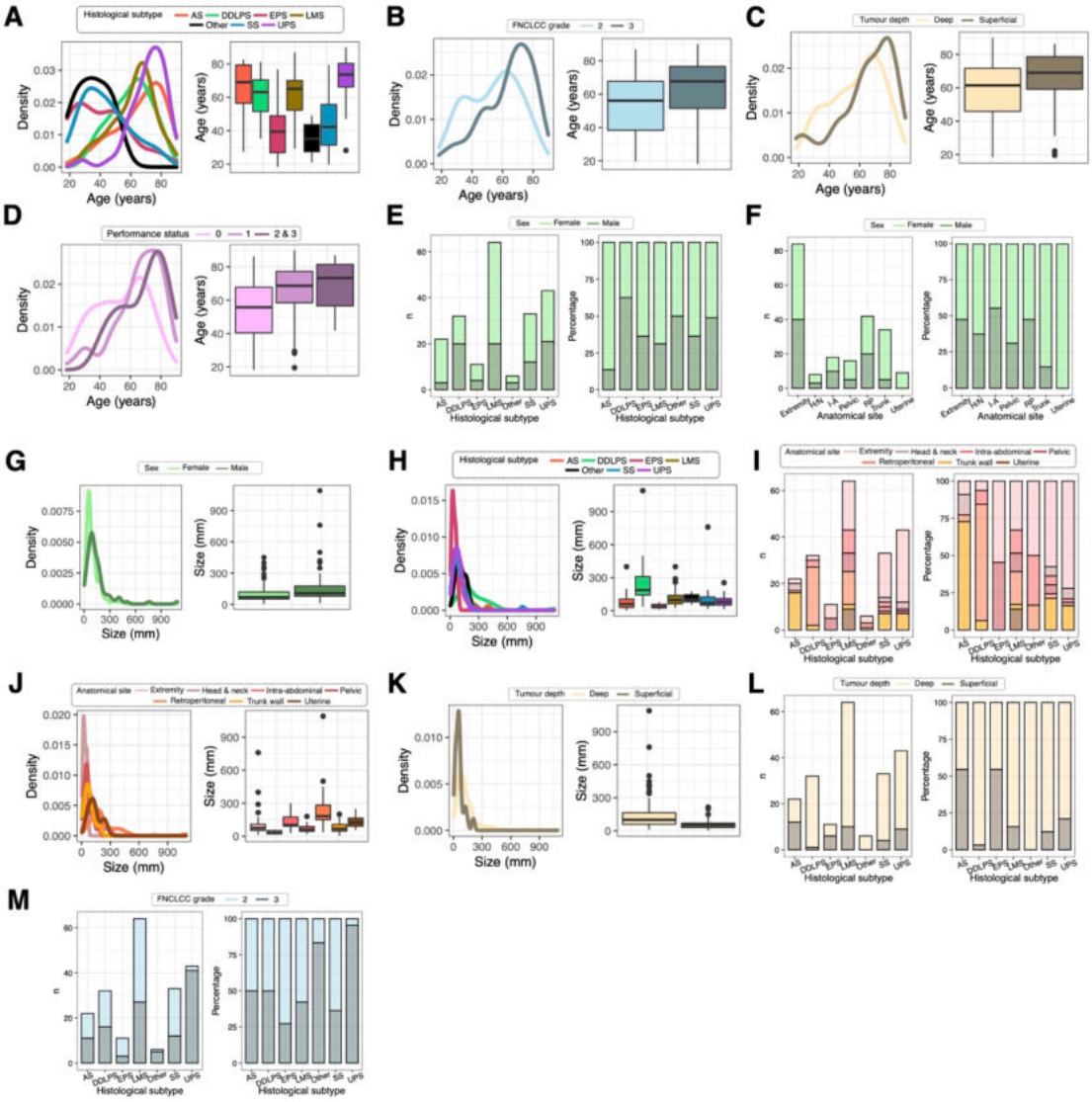
Within this chapter there were limitations. In order to characterise the baseline proteomic profile of STS, the cohort comprised solely primary tumours. Yet this was at the cost of limiting the applicability of any findings. The extent to which primary and recurrent/metastatic STS proteomes share biology is unknown. Therefore, biological insights revealed in primary tumours may not be translatable to advanced STS patients. Given the baseline proteome has now been characterised as a result of this project, future directions of interest include the profiling of matched recurrence and/or metastasis samples. Indeed, several candidate samples for such analyses have been processed and data has begun to be collected (**Chapter 3**). Additionally, no 'normal' specimens were profiled. All data is therefore relative to the other STS samples analysed, and does not facilitate the differentiation between malignant STS and normal tissue. As a result, despite observed enrichment of a certain protein or biological process, such biology may not be specific to the malignancy. It is crucial to interpret any findings as relative to the full cohort. The absence of 'normal' tissue herein is due to a lacking definition of what 'normal' represents in STS. For some subtypes, such as LMS, a clear cell of origin is known; yet in most cases the identification of a suitable 'normal' tissue is not possible[4]. Furthermore, even where a suitable 'normal' tissue is identified, availability is often limited. Practically, 'normal' tissue entails use of adjacent/margin tissue. Yet tissue adjacent to a tumour can vary extensively, is influenced by the tumour itself, and thus is not truly 'normal'.

In addition to cohort limitations, methodological limitations were also present. The gene sets and databases queried herein are rooted in transcriptomic data, and therefore subsequent analyses produce more robust insights when genome coverage is high. This is challenging to achieve with MS data and difficulties are heightened where the database or gene set profile is small itself. In most analyses herein, known biological features were observed suggesting use of these approaches is valid. However, the appropriateness of DSigDB use, given drug target profiles can be small, is unclear. Another limitation of this chapter is the reliance on descriptive analysis. This restricts interpretation and the robustness of observations and claims. Yet, descriptive assessment was sufficient in achieving the objectives of this chapter. A wide range of proteomic-derived information, both known and novel, has been captured and this has

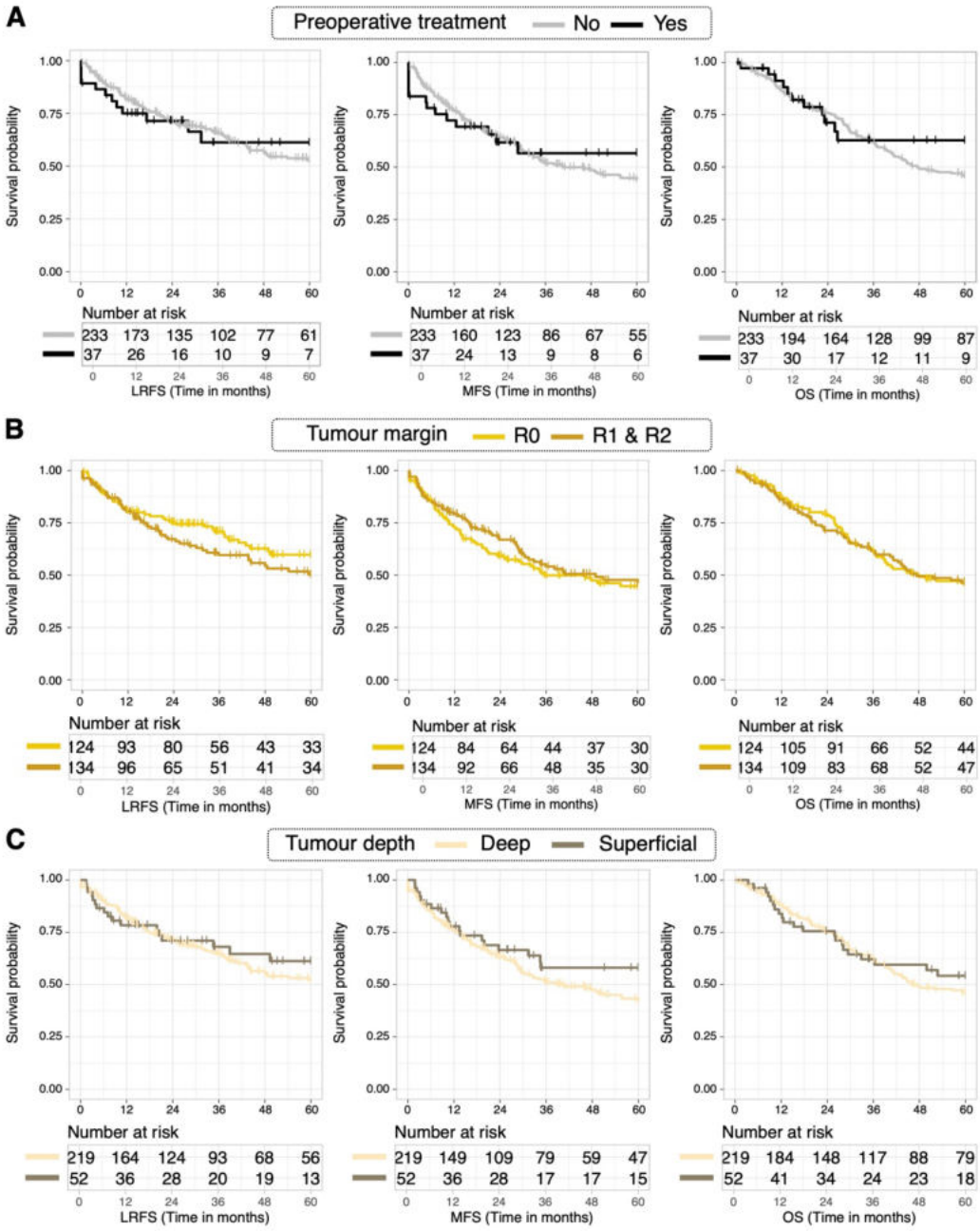
led to promising avenues for future research being identified. Despite the lack of formal statistical assessments, this chapter forms the first step toward establishing a much-needed proteomic understanding of STS.

4.4 Supplemental material

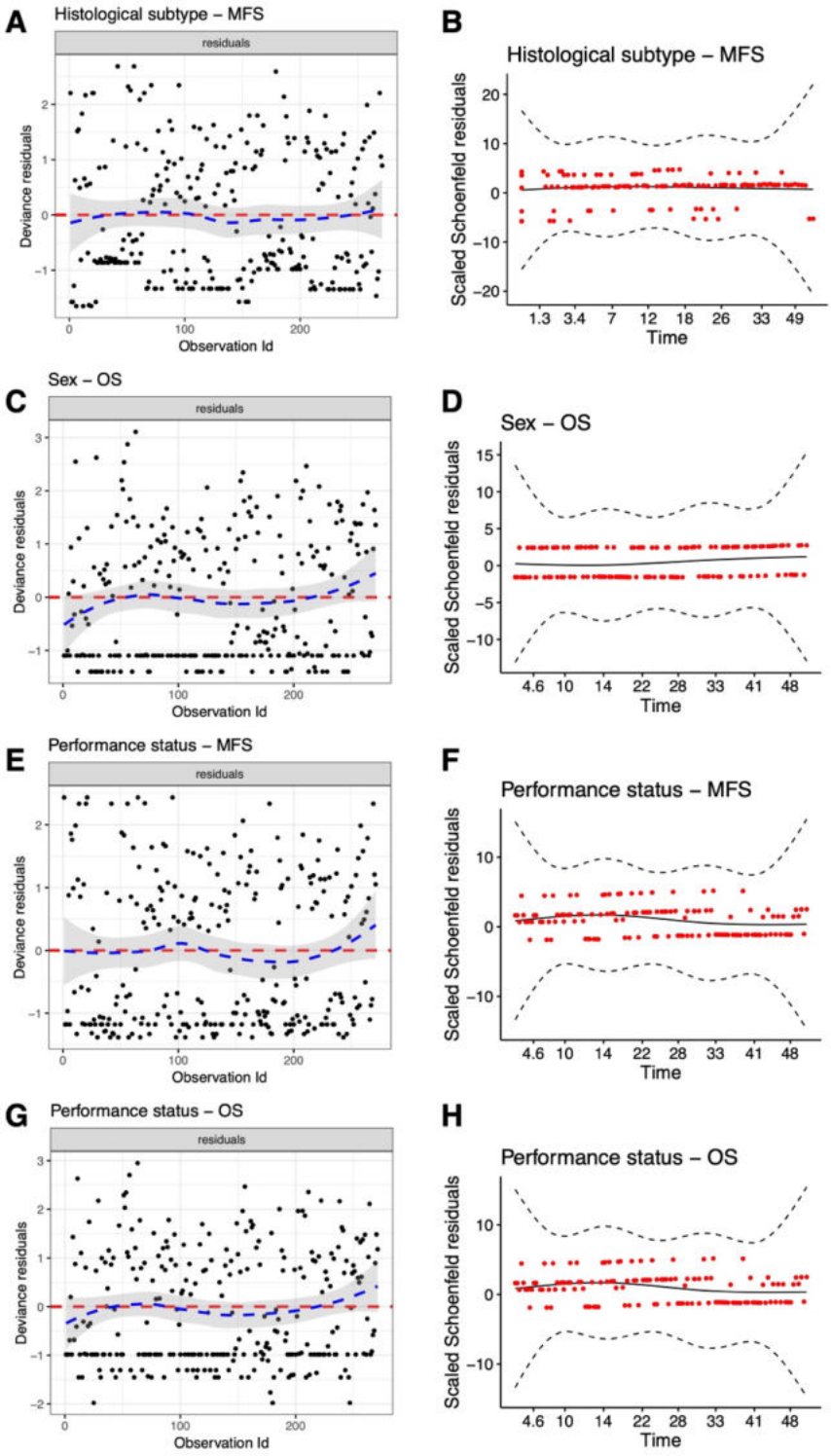
4.4.1 Supplemental Figures



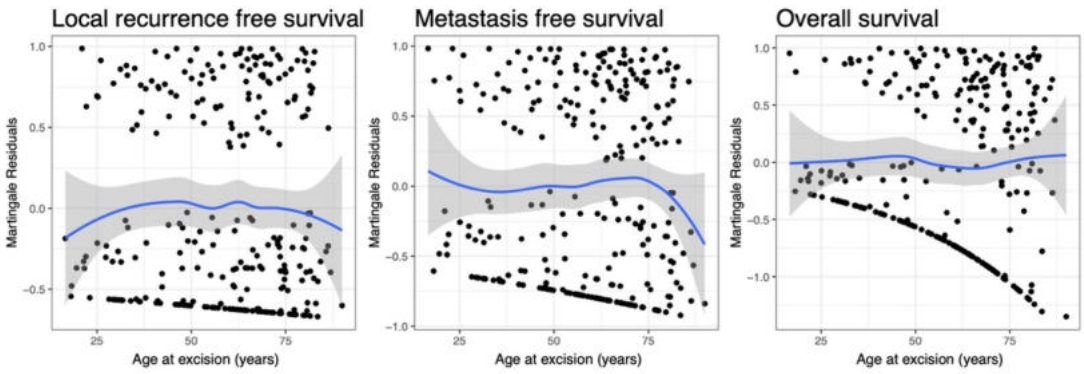
Supplemental Figure 4.1 Associations between clinicopathological variables.
(A-M) Density plots and box plots are shown for associations between continuous and categorical variables. Boxes indicate 25th and 75th percentile, with median line in the middle, whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. Stacked bar plots for number and percentage are shown for associations between 2 categorical variables. Plots illustrate the relationship between **(A)** histological subtype and age, **(B)** grade and age, **(C)** tumour depth and age, **(D)** performance status and age, **(E)** histological subtype and sex, **(F)** anatomical site and sex, **(G)** tumour size and sex, **(H)** tumour size and histological subtype, **(I)** histological subtype and anatomical site, **(J)** tumour size and anatomical site, **(K)** tumour size and tumour depth, **(L)** histological subtype and tumour depth, **(M)** histological subtype and grade. Abbreviations: FNCLCC = French Federation of Cancer Center Sarcoma Group; AS = angiosarcoma; DDLPS = dedifferentiated liposarcoma; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma. Corresponding statistical tests are detailed in **Supplemental Table 4.1**



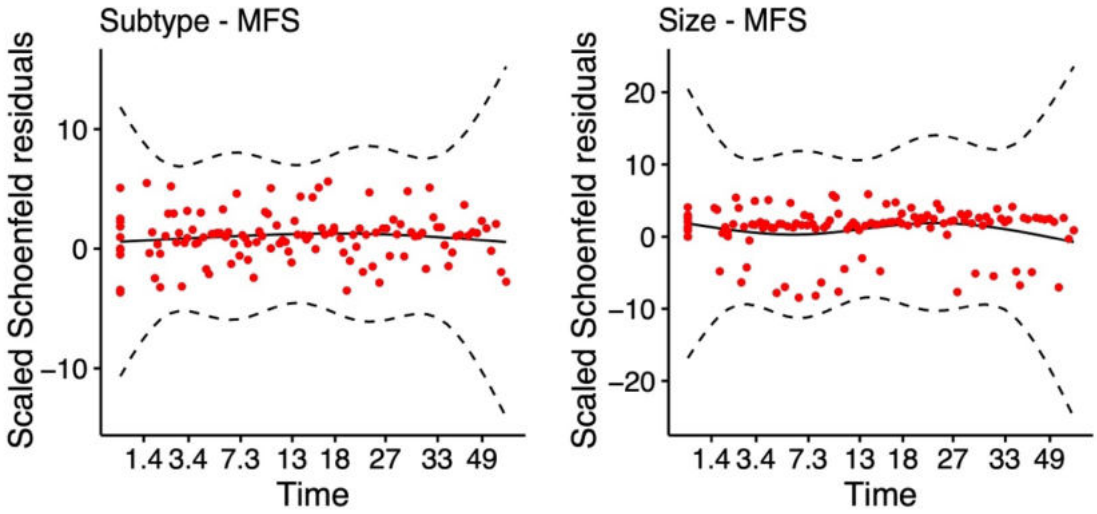
Supplemental Figure 4.2 Clinical outcome of the proteome-profiled cohort stratified by non-significant tumour and patient characteristics.
Kaplan Meier plots showing from left to right, local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) up to 5-years post-surgery. **(A)** Stratification by preoperative treatment status, where 'Yes' indicates patients that received either chemotherapy, radiotherapy, or chemotherapy and radiotherapy in the neoadjuvant setting. **(B)** Stratification by tumour margin. **(C)** Stratification by tumour depth. Corresponding univariable Cox regression results are detailed in **Supplemental Table 4.2**.



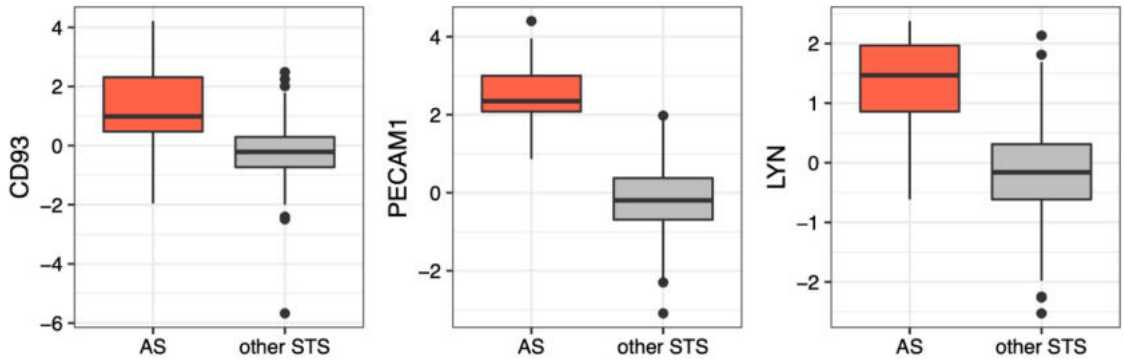
Supplemental Figure 4.3 Assessment of the proportional hazards (PH) assumption in null univariable Cox models.
Plots shown for variable-model combinations where a minor violation of the PH assumption was identified: **(A-B)** histological subtype and metastasis free survival (MFS); **(C-D)** sex and overall survival (OS); **(E-F)** performance status and MFS; **(G-H)** performance status and OS. Deviance residuals **(A,C,E,G)** plotted for each observation. Red dashed line at 0, blue line indicates a locally weighted smoothed fit and grey shading the coordinate 95% confidence intervals. Scaled Schoenfeld residuals **(B,D,F,H)** plotted over time for each observation. Solid black line indicates a smoothed spline fit of residuals and dashed black lines indicate +/- 2-standard error.



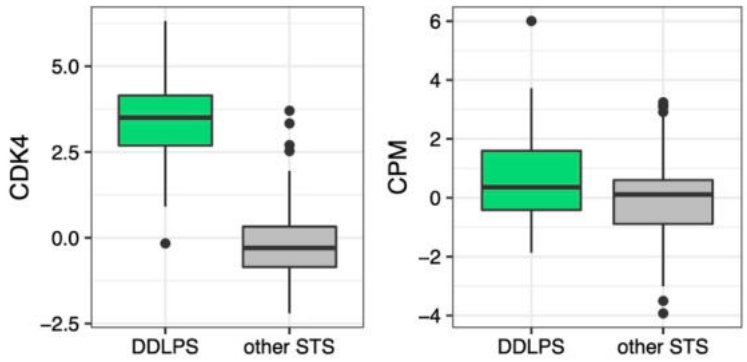
Supplemental Figure 4.4 Assessing the linearity of age in Cox regression models.
Left to right: plots for local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) showing martingale residuals against age. Blue line indicates a locally weighted smoothed fit and grey shading the coordinate 95% confidence intervals.



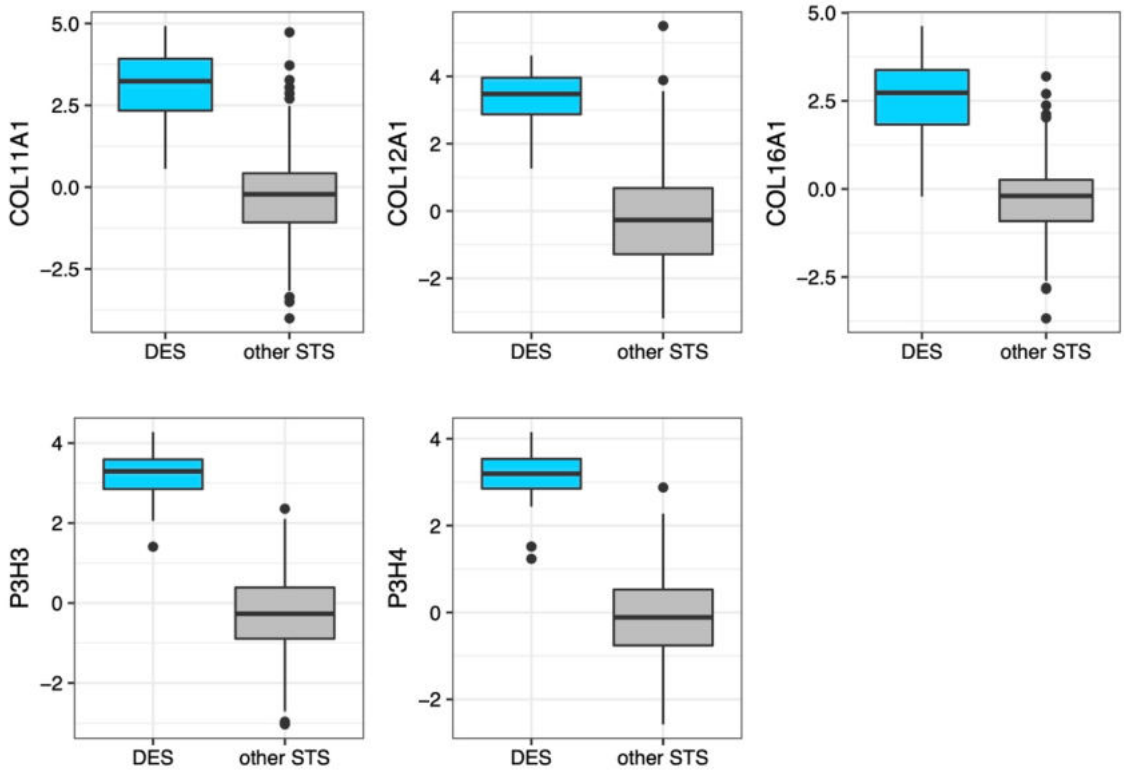
Supplemental Figure 4.5 Assessment of the proportional hazards (PH) assumption in multivariable Cox models.
Plots shown for variable-model combinations where a minor violation of the PH assumption was identified (subtype and metastasis free survival (MFS) and size and MFS). Scaled Schoenfeld residuals plotted over time for each observation. Solid black line indicates a smoothed spline fit of residuals and dashed black lines indicate +/- 2-standard error.



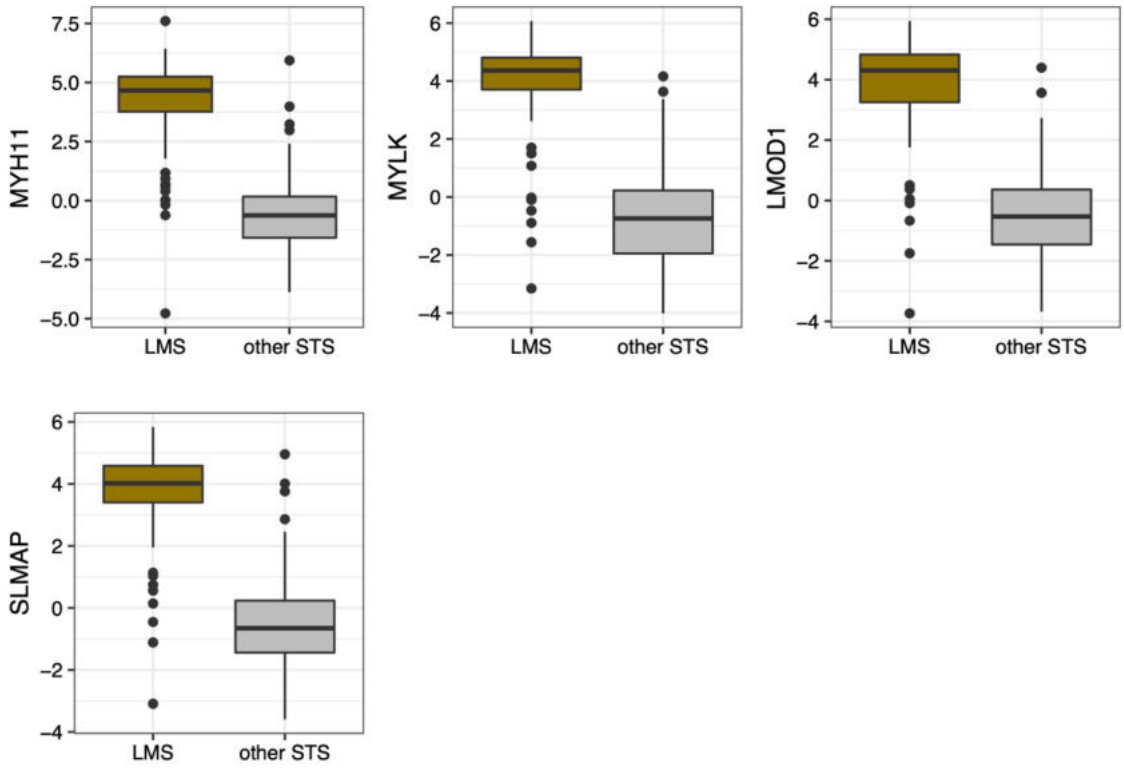
Supplemental Figure 4.6 Angiosarcoma (AS)-specific enriched proteins.
Box plots showing normalised abundance of select significant differentially expressed proteins (DEPs; fold change ≥ 1.5; FDR < 0.01) in AS compared to the rest of the cohort. Boxes indicate 25th and 75th percentile, with median line in the middle, whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. Abbreviations: CD93 = cluster of differentiation 93; PECAM1 = platelet and endothelial cell adhesion molecule 1; LYN = lyn proto-oncogene.



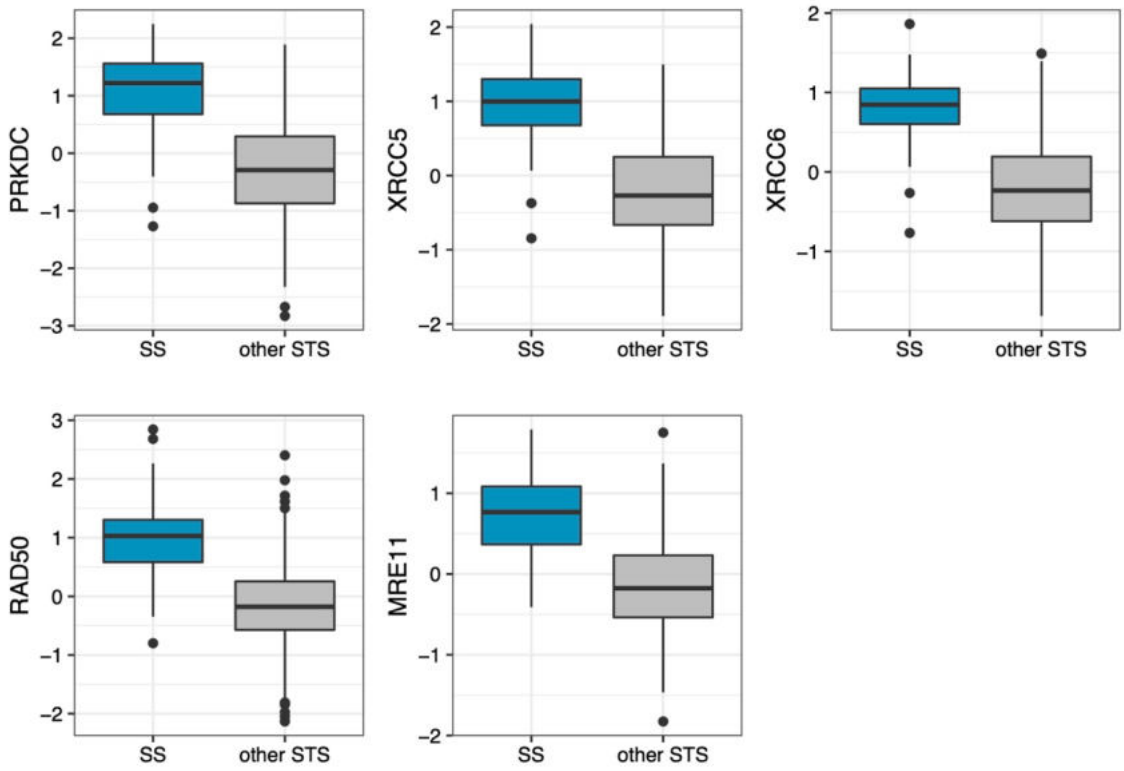
Supplemental Figure 4.7 Dedifferentiated liposarcoma (DDLPS)-specific enriched proteins.
Box plots showing normalised abundance of select significant differentially expressed proteins (DEPs; fold change ≥ 1.5; FDR < 0.01) in DDLPS compared to the rest of the cohort. Boxes indicate 25th and 75th percentile, with median line in the middle, whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. Abbreviations: CDK4 = cyclin dependent kinase 4; CPM = carboxypeptidase M.



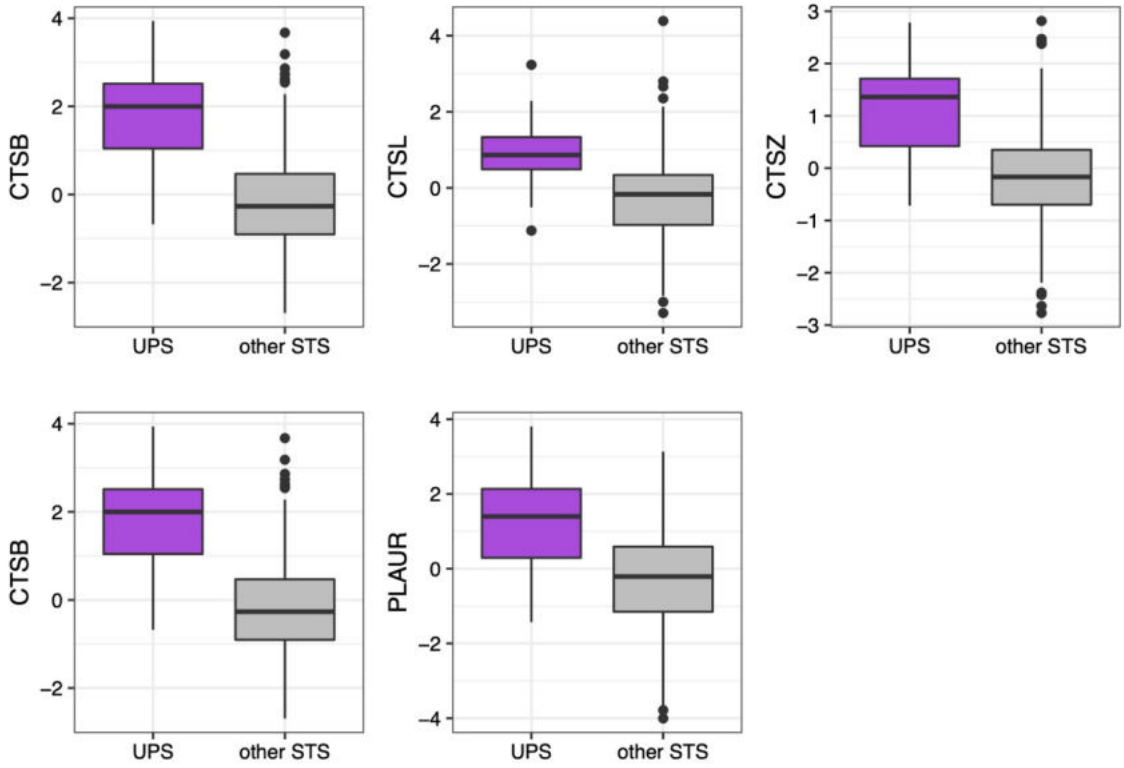
Supplemental Figure 4.8 Desmoid tumour (DES)-specific enriched proteins.
Box plots showing normalised abundance of select significant differentially expressed proteins (DEPs; fold change ≥ 1.5; FDR < 0.01) in DES compared to the rest of the cohort. Boxes indicate 25th and 75th percentile, with median line in the middle, whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. Abbreviations COL11A1/12A1/16A1 = collagen type 11/12/16 alpha 1 chain; P3H3/4 = prolyl 3-hydroxylase 3/4.



Supplemental Figure 4.9 Leiomyosarcoma (LMS)-specific enriched proteins.
Box plots showing normalised abundance of select significant differentially expressed proteins (DEPs; fold change ≥ 1.5; FDR < 0.01) in LMS compared to the rest of the cohort. Boxes indicate 25th and 75th percentile, with median line in the middle, whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. Abbreviations: MYH11 = myosin heavy chain 11; MYLK = myosin light chain kinase; LMOD1 = leiomodin 1; SLMAP = sarcolemma associated protein.



Supplemental Figure 4.10 Synovial sarcoma (SS)-specific enriched proteins.
Box plots showing normalised abundance of select significant differentially expressed proteins (DEPs; fold change ≥ 1.5; FDR < 0.01) in SS compared to the rest of the cohort. Boxes indicate 25th and 75th percentile, with median line in the middle, whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. Abbreviations: PRKDC = protein kinase, DNA-activated, catalytic subunit; XRCC5/6 = X-ray repair cross complementing 5/6; RAD50 = RAD50 double strand break repair protein; MRE11 = MRE11 homolog, double strand break repair nuclease.



Supplemental Figure 4.11 Undifferentiated pleomorphic sarcoma (UPS)-specific enriched proteins. Box plots showing normalised abundance of select significant differentially expressed proteins (DEPs; fold change ≥ 1.5; FDR < 0.01) in UPS compared to the rest of the cohort. Boxes indicate 25th and 75th percentile, with median line in the middle, whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. Abbreviations: CTSB/L/Z/D = cathepsin B/L/Z/D; PLAUR = plasminogen activator, urokinase receptor.

4.4.2 Supplemental Tables

Supplemental Table 4.1 Statistical associations between clinicopathological features.
Significant results in bold.

Variable 1	Variable 2	Test performed	Test statistic	Degrees of freedom	p	FDR
Anatomical site	Grade	Chi-squared	8.847	6	0.182	0.274
Anatomical site	**Performance status**	**Chi-squared**	**25.094**	**12**	**0.014**	**0.038**
Anatomical site	**Tumour depth**	**Chi-squared**	**29.272**	**6**	**< 0.001**	**< 0.001**
Anatomical site	Tumour margin	Chi-squared	20.942	12	0.051	0.098
Histological subtype	**Anatomical site**	**Chi-squared**	**230.782**	**36**	**< 0.001**	**< 0.001**
Histological subtype	**Grade**	**Chi-squared**	**41.690**	**6**	**< 0.001**	**< 0.001**
Histological subtype	Performance status	Chi-squared	12.928	12	0.374	0.462
Histological subtype	**Sex**	**Chi-squared**	**17.220**	**6**	**0.009**	**0.026**
Histological subtype	**Tumour depth**	**Chi-squared**	**35.450**	**6**	**< 0.001**	**< 0.001**
Histological subtype	Tumour margin	Chi-squared	8.174	4	0.085	0.149
Performance status	Grade	Chi-squared	1.231	2	0.540	0.597
Sex	**Anatomical site**	**Chi-squared**	**19.526**	**6**	**0.003**	**0.012**
Sex	Grade	Chi-squared	0.012	1	0.914	0.914
Sex	Performance status	Chi-squared	1.650	2	0.438	0.511
Sex	Tumour depth	Chi-squared	4.467	1	0.035	0.081
Sex	Tumour margin	Chi-squared	3.013	2	0.222	0.310
Tumour depth	Grade	Chi-squared	0.018	1	0.892	0.914
Tumour depth	**Performance status**	**Chi-squared**	**13.071**	**2**	**0.001**	**0.006**
Tumour depth	Tumour margin	Chi-squared	6.393	2	0.041	0.086
Tumour margin	Grade	Chi-squared	4.379	2	0.112	0.181
Tumour margin	Performance status	Chi-squared	4.941	4	0.293	0.385
Age	Anatomical site	Kruskal-Wallis	10.154	6	0.118	0.164
Age	**Grade**	**Kruskal-Wallis**	**18.347**	**1**	**< 0.001**	**< 0.001**
Age	**Histological subtype**	**Kruskal-Wallis**	**68.073**	**6**	**< 0.001**	**< 0.001**
Age	**Performance status**	**Kruskal-Wallis**	**28.290**	**2**	**< 0.001**	**< 0.001**
Age	Sex	Kruskal-Wallis	1.936	1	0.164	0.177
Age	**Tumour depth**	**Kruskal-Wallis**	**6.526**	**1**	**0.011**	**0.019**
Age	Tumour margin	Kruskal-Wallis	4.451	2	0.108	0.164
Tumour size	**Anatomical site**	**Kruskal-Wallis**	**73.934**	**6**	**< 0.001**	**< 0.001**
Tumour size	Grade	Kruskal-Wallis	0.199	1	0.655	0.655
Tumour size	**Histological subtype**	**Kruskal-Wallis**	**60.947**	**6**	**< 0.001**	**< 0.001**
Tumour size	Performance status	Kruskal-Wallis	3.989	2	0.136	0.164
Tumour size	**Sex**	**Kruskal-Wallis**	**12.030**	**1**	**< 0.001**	**0.001**
Tumour size	**Tumour depth**	**Kruskal-Wallis**	**26.937**	**1**	**< 0.001**	**< 0.001**
Tumour size	Tumour margin	Kruskal-Wallis	3.928	2	0.140	0.164

Supplemental Table 4.2 Univariable Cox regression assessing clinicopathological features.
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval.

		LRFS		MFS		OS	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
Age at excision (years)		1 (0.992-1.01)	0.581	1.01 (0.996-1.02)	0.226	**1.03 (1.01-1.04)**	**<0.001**
Tumour size (mm)		**1 (1-1)**	**0.001**	1 (0.998-1)	0.662	1 (0.999-1)	0.307
Sex	*F (ref)*	-	-	-	-	-	-
	M	1.47 (0.994-2.17)	0.054	1.01 (0.702-1.45)	0.966	**1.63 (1.15-2.3)**	**0.006**
Histological subtype	*LMS (ref)*	-	-	-	-	-	-
	UPS	1.26 (0.648-2.44)	0.499	1.02 (0.628-1.65)	0.942	1.56 (0.968-2.5)	0.068
	SS	1.36 (0.669-2.77)	0.395	0.54 (0.286-1.02)	0.058	0.793 (0.414-1.52)	0.483
	DDLPS	**3.32 (1.87-5.87)**	**<0.001**	**0.417 (0.21-0.829)**	**0.013**	1.25 (0.731-2.12)	0.420
	AS	**4.45 (2.33-8.51)**	**<0.001**	1.39 (0.783-2.46)	0.261	**1.96 (1.09-3.53)**	**0.024**
	EPS	1.96 (0.789-4.85)	0.147	1.53 (0.747-3.13)	0.245	0.996 (0.421-2.36)	0.993
	Other	1.27 (0.38-4.28)	0.694	1.39 (0.594-3.27)	0.445	0.719 (0.222-2.33)	0.582
Anatomical site	*Extremity (ref)*	-	-	-	-	-	-
	Head & neck	1.3 (0.397-4.27)	0.662	0.766 (0.277-2.12)	0.607	1 (0.363-2.78)	0.994
	Intra-abdominal	2.06 (0.977-4.34)	0.058	1.56 (0.849-2.86)	0.152	1.68 (0.93-3.03)	0.086
	Pelvis	1.44 (0.631-3.27)	0.389	0.954 (0.471-1.93)	0.896	0.881 (0.418-1.85)	0.738
	Retroperitoneal	**2.33 (1.42-3.82)**	**0.001**	0.655 (0.395-1.09)	0.101	0.925 (0.581-1.47)	0.742
	Trunk	**2.08 (1.16-3.73)**	**0.014**	0.924 (0.542-1.58)	0.771	0.971 (0.562-1.68)	0.917
	Uterine	0.404 (0.055-2.97)	0.373	1.42 (0.569-3.56)	0.451	1.2 (0.48-3.01)	0.693
FNCLCC grade	*2 (ref)*	-	-	-	-	-	-
	3	1.07 (0.718-1.608)	0.728	**1.89 (1.3-2.75)**	**<0.001**	**1.984 (1.366-2.882)**	**<0.001**
	unknown	0.993 (0.423 - 2.333)	0.988	0.858 (0.34 - 2.166)	0.746	0.735 (0.263 - 2.05)	0.556
Preoperative treatment	*No (ref)*	-	-	-	-	-	-
	Yes	1 (0.547-1.83)	1.000	0.945 (0.542-1.65)	0.843	0.769 (0.414-1.43)	0.407
Performance status	*0 (ref)*	-	-	-	-	-	-
	1	**1.7 (1.09-2.64)**	**0.019**	1.27 (0.838-1.93)	0.258	**2.18 (1.44-3.31)**	**<0.001**
	2-3	1.16 (0.495-2.73)	0.730	1.25 (0.596-2.62)	0.555	**4.04 (2.31-7.09)**	**<0.001**
	unknown	1.11 (0.609-2.01)	0.738	1.38 (0.837-2.26)	0.209	**1.76 (1.05-2.96)**	**0.031**
Tumour depth	*Deep (ref)*	-	-	-	-	-	-
	Superficial	0.867 (0.515-1.46)	0.592	0.716 (0.439-1.17)	0.181	0.875 (0.548-1.4)	0.575
Tumour margin	*R0 (ref)*	-	-	-	-	-	-
	R1 & R2	1.31 (0.871-1.97)	0.194	0.888 (0.62-1.27)	0.516	1.01 (0.708-1.43)	0.971
	unknown	1.75 (0.786-3.91)	0.170	0.764 (0.307-1.9)	0.563	0.456 (0.143-1.45)	0.185
Log(Tumour size [mm])	4-5 (ref)						
	< 4	0.697 (0.394-1.23)	0.213	**0.451 (0.272-0.746)**	**0.002**	**0.542 (0.323-0.912)**	**0.021**
	> 5	**1.99 (1.29-3.06)**	**0.002**	0.746 (0.483-1.15)	0.186	1.32 (0.893-1.95)	0.163

Supplemental Table 4.3 Multivariable Cox regression assessing clinicopathological features.
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval.

		LRFS		MFS		OS	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
Age at excision (years)		0.999 (0.982-1.02)	0.919	1 (0.983-1.02)	0.97	1.01 (0.997-1.03)	0.111
Sex	*F (ref)*	-	-	-	-	-	-
	M	1.46 (0.907-2.36)	0.119	1.22 (0.787-1.89)	0.373	1.49 (0.968-2.28)	0.07
Histological subtype	*LMS (ref)*	-	-	-	-	-	-
	UPS	1.51 (0.684-3.31)	0.309	0.905 (0.505-1.62)	0.738	1.4 (0.762-2.56)	0.279
	SS	1.58 (0.681-3.65)	0.288	0.591 (0.277-1.26)	0.173	1.1 (0.522-2.33)	0.797
	DDLPS	1.58 (0.772-3.24)	0.211	**0.333 (0.148-0.75)**	**0.008**	1.03 (0.508-2.08)	0.94
	AS	**7.98 (3.15-20.2)**	**<0.001**	**2.94 (1.27-6.78)**	**0.012**	**4.77 (2.12-10.7)**	**<0.001**
	EPS	2.54 (0.769-8.36)	0.126	**3.05 (1.17-7.98)**	**0.023**	1.97 (0.646-6.03)	0.233
	Other	1.39 (0.337-5.72)	0.649	1.41 (0.464-4.3)	0.543	1.87 (0.481-7.3)	0.366
Anatomical site	*Extremity (ref)*	-	-	-	-	-	-
	Head & neck	0.935 (0.238-3.67)	0.924	1.02 (0.295-3.54)	0.972	1.1 (0.331-3.66)	0.877
	Intra-abdominal	2 (0.849-4.72)	0.113	1.56 (0.792-3.05)	0.199	**2.08 (1.03-4.21)**	**0.042**
	Pelvis	1.75 (0.689-4.47)	0.239	1.14 (0.526-2.47)	0.742	1.64 (0.719-3.72)	0.24
	Retroperitoneal	1.45 (0.643-3.29)	0.369	0.826 (0.409-1.67)	0.594	0.902 (0.427-1.9)	0.786
	Trunk	1 (0.431-2.33)	0.998	0.768 (0.365-1.61)	0.485	0.773 (0.361-1.65)	0.507
	Uterine	0.737 (0.087-6.21)	0.779	1.43 (0.457-4.49)	0.537	2.03 (0.668-6.18)	0.212
FNCLCC grade	*2 (ref)*	-	-	-	-	-	-
	3	1.23 (0.76-1.99)	0.401	**1.94 (1.26-3)**	**0.003**	**1.88 (1.2-2.94)**	**0.006**
	unknown	0.842 (0.319-2.23)	0.729	0.711 (0.264-1.91)	0.498	0.808 (0.265-2.46)	0.708
Performance status	*0 (ref)*	-	-	-	-	-	-
	1	**1.7 (1.02-2.81)**	**0.041**	1.56 (0.973-2.51)	0.065	**2.07 (1.31-3.29)**	**0.002**
	2-3	1.16 (0.463-2.89)	0.754	1.24 (0.543-2.82)	0.612	**3.01 (1.55-5.84)**	**0.001**
	unknown	1.07 (0.561-2.02)	0.845	1.4 (0.804-2.44)	0.234	1.55 (0.875-2.76)	0.132
Tumour depth	*Deep (ref)*	-	-	-	-	-	-
	Superficial	1.1 (0.567-2.13)	0.78	0.571 (0.316-1.03)	0.064	0.925 (0.512-1.67)	0.797
Tumour margin	*R1 & R2 (ref)*	-	-	-	-	-	-
	R0	0.773 (0.492-1.21)	0.262	1.1 (0.734-1.65)	0.644	1.04 (0.698-1.54)	0.859
	unknown	1.2 (0.497-2.88)	0.688	1.5 (0.519-4.33)	0.454	0.775 (0.224-2.68)	0.687
Log(Tumour size [mm])	4-5 (ref)						
	< 4	**0.445 (0.231-0.859)**	**0.016**	**0.39 (0.212-0.716)**	**0.002**	**0.486 (0.266-0.887)**	**0.019**
	> 5	1.69 (0.897-3.18)	0.105	1.03 (0.599-1.78)	0.908	1.64 (0.987-2.72)	0.056

Supplemental Table 4.4 Associations between histological subtype and subtype-specific proteins in the proteomics data.
Abbreviations: MYH11 = myosin heavy chain 11; GAPDH = glyceraldehyde-3-phoshphate dehydrogenase; SRC = proto-oncogene tyrosine-protein kinase Src; PRDX1 = peroxiredoxin 1; G6PD = glucose-6-phoshphate dehydrogenase; TFRC = transferrin receptor. Corresponding post-hoc analysis results are detailed in **Supplemental Table 4.6**.

Variable 1 (continuous)	Variable 2 (categorical)	Kruskal-Wallis test		
		Test statistic (X^2)	Degrees of freedom	p
MYH11	Histological subtype	126.81	3	< 0.001
GAPDH	Histological subtype	89.303	3	< 0.001
SRC	Histological subtype	71.992	3	< 0.001
PRDX1	Histological subtype	54.589	3	< 0.001
G6PD	Histological subtype	44.221	3	< 0.001
TFRC	Histological subtype	45.695	3	< 0.001

Supplemental Table 4.5 Associations between histological subtype and subtype-specific proteins in the reverse-phase protein array (RPPA) data from The Cancer Genome Atlas (TCGA).
Abbreviations: MYH11 = myosin heavy chain 11; GAPDH = glyceraldehyde-3-phoshphate dehydrogenase; SRC = proto-oncogene tyrosine-protein kinase Src; PRDX1 = peroxiredoxin 1; G6PD = glucose-6-phoshphate dehydrogenase; TFRC = transferrin receptor. Corresponding post-hoc analysis results are detailed in **Supplemental Table 4.7**.

Variable 1 (continuous)	Variable 2 (categorical)	Kruskal-Wallis test		
		Test statistic (X^2)	Degrees of freedom	p
MYH11	Histological subtype	74.199	3	< 0.001
GAPDH	Histological subtype	20.322	3	< 0.001
SRC	Histological subtype	51.369	3	< 0.001
PRDX1	Histological subtype	18.378	3	< 0.001
G6PD	Histological subtype	18.917	3	< 0.001
TFRC	Histological subtype	21.026	3	< 0.001

Supplemental Table 4.6 Post-hoc test associations between histological subtype and subtype-specific proteins in the proteomic data.
Abbreviations: MYH11 = myosin heavy chain 11; GAPDH = glyceraldehyde-3-phoshphate dehydrogenase; SRC = proto-oncogene tyrosine-protein kinase Src; PRDX1 = peroxiredoxin 1; G6PD = glucose-6-phoshphate dehydrogenase; TFRC = transferrin receptor; DDLPS = dedifferentiated liposarcoma; LMS = leiomyosarcoma; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma

Protein	Comparison	Dunn's test		
		Test statistic (Z)	p	p adjusted
MYH11	**DDLPS - LMS**	**-8.529**	**<0.001**	**<0.001**
	DDLPS - SS	-0.361	0.718	0.861
	LMS - SS	**8.386**	**<0.001**	**<0.001**
	DDLPS - UPS	-0.504	0.614	0.921
	LMS - UPS	**8.804**	**<0.001**	**<0.001**
	SS - UPS	-0.129	0.898	0.898
GAPDH	**DDLPS - LMS**	**-6.196**	**<0.001**	**<0.001**
	DDLPS - SS	**1.999**	**0.046**	**0.046**
	LMS - SS	**8.736**	**<0.001**	**<0.001**
	DDLPS - UPS	**-3.076**	**0.002**	**0.003**
	LMS - UPS	**3.168**	**0.002**	**0.002**
	SS - UPS	**-5.315**	**<0.001**	**<0.001**
SRC	**DDLPS - LMS**	**-3.672**	**<0.001**	**<0.001**
	DDLPS - SS	1.245	0.213	0.213
	LMS - SS	**5.249**	**<0.001**	**<0.001**
	DDLPS - UPS	**3.430**	**<0.001**	**<0.001**
	LMS - UPS	**8.135**	**<0.001**	**<0.001**
	SS - UPS	**2.184**	**0.029**	**0.035**
PRDX1	DDLPS - LMS	0.726	0.468	0.468
	DDLPS - SS	**2.717**	**0.007**	**0.010**
	LMS - SS	**2.428**	**0.015**	**0.018**
	DDLPS - UPS	**-3.999**	**<0.001**	**<0.001**
	LMS - UPS	**-5.564**	**<0.001**	**<0.001**
	SS - UPS	**-7.038**	**<0.001**	**<0.001**
G6PD	DDLPS - LMS	1.397	0.162	0.162
	DDLPS - SS	**2.901**	**0.004**	**0.006**
	LMS - SS	1.949	0.051	0.062
	DDLPS - UPS	**-3.032**	**0.002**	**0.005**
	LMS - UPS	**-5.153**	**<0.001**	**<0.001**
	SS - UPS	**-6.242**	**<0.001**	**<0.001**
TFRC	**DDLPS - LMS**	**-2.225**	**0.026**	**0.031**
	DDLPS - SS	**3.182**	**0.001**	**0.003**
	LMS - SS	**6.019**	**<0.001**	**<0.001**
	DDLPS - UPS	**-2.435**	**0.015**	**0.022**
	LMS - UPS	-0.447	0.655	0.655
	SS - UPS	**-5.931**	**<0.001**	**<0.001**

Supplemental Table 4.7 Post-hoc test associations between histological subtype and subtype-specific proteins in the reverse-phase protein array (RPPA) data from The Cancer Genome Atlas (TCGA).
Abbreviations: MYH11 = myosin heavy chain 11; GAPDH = glyceraldehyde-3-phoshphate dehydrogenase; SRC = proto-oncogene tyrosine-protein kinase Src; PRDX1 = peroxiredoxin 1; G6PD = glucose-6-phoshphate dehydrogenase; TFRC = transferrin receptor; DDLPS = dedifferentiated liposarcoma; LMS = leiomyosarcoma; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma

Protein	Comparison	Dunn's test		
		Test statistic (Z)	p	p adjusted
MYH11	**DDLPS - LMS**	**-7.277**	**< 0.001**	**< 0.001**
	DDLPS - SS	-0.618	0.536	0.644
	LMS - SS	**2.704**	**0.007**	**0.014**
	DDLPS - UPS	0.138	0.890	0.890
	LMS - UPS	**7.185**	**< 0.001**	**< 0.001**
	SS - UPS	0.682	0.495	0.743
GAPDH	**DDLPS - LMS**	**-4.098**	**< 0.001**	**< 0.001**
	DDLPS - SS	0.746	0.456	0.456
	LMS - SS	**2.632**	**0.008**	**0.025**
	DDLPS - UPS	-1.670	0.095	0.142
	LMS - UPS	2.193	0.028	0.057
	SS - UPS	-1.562	0.118	0.142
SRC	**DDLPS - LMS**	**-4.161**	**< 0.001**	**< 0.001**
	DDLPS - SS	**3.233**	**0.001**	**0.002**
	LMS - SS	**5.182**	**< 0.001**	**< 0.001**
	DDLPS - UPS	1.531	0.126	0.126
	LMS - UPS	**5.647**	**< 0.001**	**< 0.001**
	SS - UPS	**-2.459**	**0.014**	**0.017**
PRDX1	**DDLPS - LMS**	**3.392**	**< 0.001**	**0.002**
	DDLPS - SS	1.833	0.067	0.100
	LMS - SS	0.306	0.760	0.912
	DDLPS - UPS	-0.203	0.839	0.839
	LMS - UPS	**-3.496**	**< 0.001**	**0.003**
	SS - UPS	-1.920	0.055	0.110
G6PD	**DDLPS - LMS**	**3.239**	**0.001**	**0.004**
	DDLPS - SS	**2.223**	**0.026**	**0.039**
	LMS - SS	0.771	0.441	0.529
	DDLPS - UPS	-0.317	0.752	0.752
	LMS - UPS	**-3.468**	**< 0.001**	**0.003**
	SS - UPS	**-2.363**	**0.018**	**0.036**
TFRC	DDLPS - LMS	-1.904	0.057	0.068
	DDLPS - SS	**2.730**	**0.006**	**0.010**
	LMS - SS	**3.639**	**< 0.001**	**0.001**
	DDLPS - UPS	**-2.774**	**0.006**	**0.011**
	LMS - UPS	-1.099	0.272	0.272
	SS - UPS	**-4.074**	**< 0.001**	**< 0.001**

Chapter 5 Proteomic heterogeneity in LMS, UPS, and DDLPS

5.1 Background and objectives

Chapter 4 descriptively characterised the profiled cohort and proteomic landscape of multiple histological subtypes of STS. This alluded to the presence of proteomic heterogeneity within subtypes ('intra-subtype'). Intra-subtype heterogeneity is clinically seen in STS as demonstrated by differing patient outcomes and responses to treatment intervention across patients (discussed in **section 1.2.3**). We hypothesis that this clinical heterogeneity is underscored by molecular biology, and in particular, proteome biology. This chapter investigates the intra-subtype biological heterogeneity of LMS and the immune intra-subtype heterogeneity of DDLPS and UPS, and discuses proteomic findings in relation to clinical applications.

In **Chapter 3,** LMS showed a distinctive proteome relative to other STS subtypes. Yet, when analyses were focused on specific biological entities and broad measures of biological activity (e.g., the immune component, and GO BP and hallmarks of MSigDB), proteomic subtypes of LMS emerged. At present, there has been extensive transcriptomic work characterising 3 molecular subtypes of LMS (as discussed in **section 1.4.1.2**)[36,43,274,281–283]. Yet the clinical implications of LMS molecular subtypes are unclear, and it is unknown whether these transcriptomic findings are present at the proteome level. By design, this cohort profiled many LMS samples (n = 80). Such rich data could be leveraged to facilitate intra-subtype analyses. Herein, proteomic subtypes of LMS were discovered using unbiased methods and were characterised both biologically and clinically. Furthermore, indirect comparisons between the proteome-derived subtypes and transcriptome-derived subtypes of LMS were performed by use of the TCGA RNAseq data.

In **Chapter 4**, UPS and DDLPS were shown to harbour variable immune activity, with some tumours showing exceptionally high immune levels. Current STS literature notes high immune activity to exist in a minority of STS tumours across subtypes[220,231]. However, UPS is often highlighted as the subtype with the highest immune infiltrate[36,220]. Clinical trials evaluating immunotherapies in STS, and particularly ICBs, report favourable responses to be more prevalent in UPS and DDLPS populations compared to other subtypes[139,140]. As such, this chapter investigates the immune composition of DDLPS and UPS tumours. To increase to statistical power of such analyses and given

the similar ICB response rates seen in these 2 subtypes, samples were combined as 1 cohort (n = 92). Proteomics data collected herein (**Chapter 3**), as well as targeted immune transcriptomic data and IHC data were utilised to explore immune-associated heterogeneity. The findings of which were discussed in the context of clinical applications.

In line with this, the objectives of this chapter are:

1) To investigate and characterise proteomic heterogeneity of LMS.
2) To assess immune-based heterogeneity within a mixed DDLPS and UPS cohort.

5.2 Results

5.2.1 Intra-subtype heterogeneity in LMS

5.2.1.1 Clinicopathological features of the LMS cohort

Tumour specimens from 80 primary LMS tumours were profiled. Clinicopathological features are summarised in **Table 5.1**. Briefly, patients had a median age of 65.3 years at the time of surgery and a median tumour size of 90 mm. There were more females than males (70% vs 30%) and more high grade tumours than intermediate grade (58.8% vs 41.2%). Most tumours were deep (82.5%) and located in either the extremities (38.8%) or retroperitoneum (23.8%). Low numbers of uterine tumours were present (9%). 2 patients had metastatic disease at surgery, 2 had radiation-associated disease, and 1 received preoperative treatment (RTX). Surgical margins were most often R0 (52.5%), and most patients had a PS of either 0 (50%) or 1 (20%). Most missing data was due to the PS variable, which was comparably missing in the LMS cohort (20%) as it was in the full cohort (18.7%). In addition, tumour margin information was not available for 2 patients. Interactions between clinicopathological features were assessed and revealed a significant association to exist between tumour depth and anatomical site (FDR < 0.001), and tumour depth and tumour size status (FDR < 0.001; **Supplemental Figure 5.1** and **Supplemental Table 5.1**). Superficial tumours were smaller than deep-seated tumours and were exclusively located in either the extremities or pelvis.

Table 5.1 Clinicopathological features of the leiomyosarcoma (LMS) cohort.
Continuous variables detailed as median, minimum (min), and maximum (max). Categorical variables detailed as count (n) and percentage. Abbreviations: F = female; M = male; RTX = radiotherapy.

		LMS
	n	**80**
Age at excision (years)	**median**	65.3
	min	29.3
	max	86.9
Tumour size (mm)	**median**	92.5
	min	5
	max	400
Sex [n (%)]	**F**	56 (70.0)
	M	24 (30.0)
Grade [n (%)]	**2**	47 (58.8)
	3	33 (41.2)
Anatomical site [n (%)]	**Extremity**	31 (38.8)
	Intra-abdominal	10 (12.5)
	Retroperitoneal	19 (23.8)
	Trunk	2 (2.5)
	Pelvic	9 (11.2)
	Uterine	9 (11.2)
Tumour depth [n (%)]	**Deep**	66 (82.5)
	Superficial	14 (17.5)
Status at excision [n (%)]	**Local**	78 (97.5)
	Metastatic	2 (2.5)
Radiation associated [n (%)]	**No**	78 (97.5)
	Yes	2 (2.5)
Tumour margins [n (%)]	**R0**	42 (52.5)
	R1	35 (43.8)
	R2	1 (1.2)
	unknown	2 (2.5)
Performance status [n (%)]	**0**	40 (50.0)
	1	16 (20.0)
	2	7 (8.8)
	3	1 (1.2)
	unknown	16 (20.0)
Pre-op treatment [n (%)]	**RTX**	1 (1.2)
	None	79 (98.8)

5.2.1.2 Cohort outcomes and the prognostic significance of clinicopathological variables

Survival data was censored at 5 years and therefore information on longer term outcomes was not available. Median LRFS and OS for the cohort were not reached (**Figure *5.1*A,C**). Median MFS was ~ 36 months (**Figure *5.1*B**). At 5-years post-surgery, 26% of patients had experienced a local recurrence event, 56% had experienced a metastatic event, and 48% were deceased.

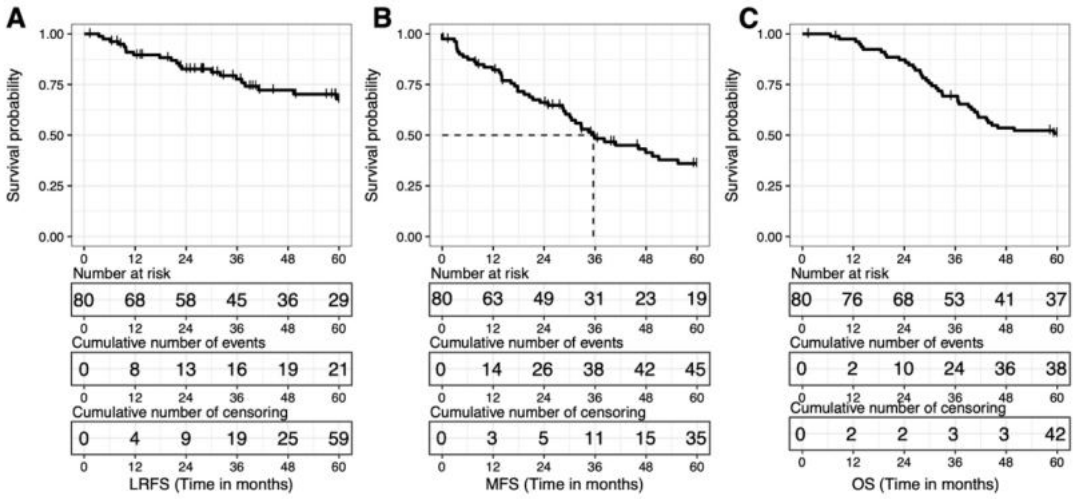


Figure 5.1 Clinical outcome of the leiomyosarcoma (LMS) cohort.
Kaplan Meier plots showing local recurrence free survival (LRFS; **A**), metastasis free survival (MFS; **B**), and overall survival (OS; **C**) up to 5-years post-surgery. Dashed line indicates median survival.

To assess whether clinicopathological variables were associated with LMS patient outcomes, Kaplan Meier curves were plotted and univariable Cox regressions performed, the results of which are summarised in **Supplemental Table 5.2**. Due to small numbers within some of the categories, anatomical sites were merged. Trunk wall and extremity tumours were grouped, and intra-abdominal, retroperitoneal, and pelvic tumours were grouped. These are representative of extra-cavity and intra-cavity lesions respectively, and are the anatomical sites differentiated between by the TCGA sarcoma study[36]. Given the clinical differences between uLMS and stLMS (**section 1.4.1**), uterine tumours were kept as a distinct category despite n being small. All other variables were handled as in **section 4.2.2**. Univariable Cox regression revealed anatomical site to be a significant prognosticator across all clinical outcomes measured (**Supplemental Figure 5.2A**). Specifically, trunk wall and extremity tumours showed a significantly superior LRFS (HR = 0.27; 95% CI = 0.111 – 0.657; p = 0.003), MFS (HR = 0.392; 95%

CI = 0.207 – 0.742; p = 0.004), and OS (HR = 0.421; 95% CI = 0.21 – 0.844; p = 0.015). FNCLCC grade, tumour depth, and log(tumour size) were significantly associated with MFS (**Supplemental Figure 5.2B-D**) Grade 3 tumours showed a poorer MFS (HR = 2.46; 95% CI = 1.35 – 4.46; p = 0.003), whilst superficial tumours and the smallest tumours (log(size) < 4 mm) showed superior MFS (HR = 0.291; 95% CI = 0.09 – 0.94; p = 0.039, and HR = 0.226; 95% CI = 0.069 – 0.74; p = 0.014 respectively). A performance status of 2 - 3 was significantly associated with a poorer OS (HR = 7.82; 95% CI = 3.15 – 19.4; p < 0.001; **Supplemental Figure 5.3**).

Following multivariable adjustment (summarised in **Supplemental Table 5.3**), anatomical site remained a significant prognosticator for LRFS, MFS, and OS. PS remained significant in the OS model, and FNCLCC grade remained significant in the MFS model. Notably, unlike in the univariable analyses, tumour depth and size were not significant for MFS, and thus do not hold significant independent prognostic value in this cohort. Additionally, grade gained significance for OS, and PS gained significance for LRFS. A gain of significance in multivariable analyses is notable. One potential reason could be the influence of missing clinicopathological data, which resulted in the exclusion of 2 patients from multivariable analysis. However, neither of these patients presented with an extreme LRFS; 1 was censored at approximately 5 years, and 1 experienced an LR event at approximately 3 years. Similarly, OS for these patients was unremarkable; both censored at approximately 5 years. It is therefore unlikely that the exclusion of such patients in the multivariable model is driving a gain of significance in other clinicopathological variables. A more probable explanation is the presence of statistical suppression. In regression models variables often 'interact' with each other[599]. One type of 'interaction' is suppression[600]. A suppressor variable is a weak predictor of the dependent variable (DV) itself, but when included in a model increases the predictive ability of other independent variables (IV). A suppressor can be conceptually thought of as a mediator (facilitating the effect of another IV to the DV) or a moderator (managing the strength of another IV to the DV). In the multivariable models herein, 7 variables are included. It is possible that multiple suppressors are present within the data, each conveying positive, negative, and/or reciprocal suppression. Grade and PS did not show any pairwise association with other clinicopathological variables in this cohort (**section 5.2.1.1**), and thus, if indeed present, the mechanism of suppression in these models is unclear.

As described previously (**section 4.2.2**), the use of a transformed and ordinal tumour size variable meant all assumptions of the Cox model were met. A minor PH violation

was noted in the univariable and multivariable MFS regressions for anatomical site (Schoenfeld p = 0.01 and p = 0.04 respectively; **Supplemental Figure *5.4*** and **Supplemental Figure *5.5***). However, this did not invalidate the use of the Cox model.

5.2.1.3 Identification of LMS proteomic subtypes

Molecular subtypes of LMS have been identified based on the transcriptome[36,43,274,281–283]. Yet the presence of these subtypes at the proteome level is unexplored. Across the 80 LMS samples profiled, 3,263 proteins were identified and quantified with high confidence. To investigate proteomic heterogeneity within LMS, consensus clustering (CC) was performed. CC is a robust method to identify clusters within a dataset[601]. CC iteratively clusters sub-samples of the original data. Each sub-sampled data does not include all original data, and therefore introduces a variability that permits cluster stability to be inferred. CC is simulated for different numbers of clusters (k), and cluster stability used to determine the optimal value of k. In a dataset of unique samples, clustering will be 'perfect' at k = n of samples. However, this provides no biological or clinical insight. Instead, insight is derived from an optimal k value identified as the value beyond which only minimal improvements in cluster stability are seen. There are several ways to assess stability. Consensus matrices of CC illustrate sample assignment over all iterations, where a value of 1 indicates a sample was assigned to the cluster every time and a value of 0 indicates a sample was never assigned to the cluster. The consensus empirical cumulative distribution function (CDF) plot corresponds to the consensus matrix, where steps at 0 at 1 are sized relative to the number of 0's and 1's in the matrix. 'Perfect' clustering is indicated by a large step at consensus index 0, a flattening of the CDF between consensus index range 0 to 1, followed by a step at consensus index 1. The CDF can be partially summarised by calculating the change (Δ) in the area under the curves (AUC). When the Δ AUC is plotted, the inflection point can be used to identify the value of k beyond which only minimal improvements in clustering are seen. Additionally, a tracking plot can be used to visualise progression of sample assignment. Samples which repeatedly switch between clusters indicate poor stability. Finally, the assignment of a sample to a cluster over many iterations can also be numerically summarised by the silhouette width (S_i). An S_i close to 1 indicates highly robust clustering, an S_i close to 0 indicates clustering equal to random assignment, and a negative S_i (close to -1) indicates clustering is probably incorrect. Herein, CC was simulated up to k = 10. Visually, the consensus matrices from k = 2 to k = 5 showed similarly clean cluster separation with few intermediate values (i.e., close to 0.5; **Supplemental Figure 5.6A**).The CDF plot showed an obvious increase in the AUC between k = 2 and k = 3, with minor shifts in the curves beyond (**Supplemental Figure**

5.6B). The Δ area plot showed an inflection point at k = 4, with minimal changes in the AUC of the CDF beyond this (**Supplemental Figure 5.6C**). The CC tracking plot indicated good cluster stability at all values of k except k = 4. At k = 4, one case was separated from the cohort, then reassigned at k = 5 to the same group as in k = 3, before being separated again at k = 7 (**Supplemental Figure 5.6D**). Silhouette plots indicated k = 2 and k = 3 to show good clustering results with the average S_i for both > 0.8 (**Supplemental Figure 5.6E**). Given all visual observations, k = 3 was deemed as optimal.

SigClust was used to confirm the clustering at k = 3, by statistically assessing the significance of the results. The SigClust null hypothesis states that data is from a single Gaussian distribution[520]. Therefore, rejection of the null hypothesis indicates the presence of multiple significantly different Gaussian distributions (i.e., clusters) within the data. Running SigClust from the root of each node of the dendrogram found clusters at k = 3, k = 4, and k = 5 as significantly different dependent on the significance level ($p < 0.001, < 0.01, < 0.05$ respectively).

Using CC and SigClust, 3 objectively distinct proteomic subtypes of LMS were confidently identified: P1, P2, and P3 (**Figure 5.2A**). Complementary approaches to CC were used to visualise the subtypes. Hierarchical clustering, and dimension reduction by PCA, tSNE, and UMAP all illustrated a sample clustering pattern that is reflective of the CC results. Yet no dimension reduction method showed robust separation alone (**Figure 5.2B-E**). This illustrates the necessity of iterative and statistical cluster identification methods such as CC and SigClust.

5.2.1.4 Biological characterisation of LMS proteomic subtypes

To identify DEPs between the LMS proteomic subtypes, 2-class unpaired SAM tests were performed (P1 vs 'other'; P2 vs 'other'; P3 vs 'other'). In total, 101, 129, and 143 DEPs were significantly upregulated in P1, P2, and P3 LMS respectively, and 110, 203, and 181 DEPs were significantly downregulated in P1, P2, and P3 LMS respectively (FDR < 0.01 and fold change ≥ 2; **Supplemental Figure *5.7***). As a result of performing multiple paired tests, there was significant overlap in the proteins identified as up/downregulated in each proteomic subtype. To assess proteins specifically altered in each subtype, the protein lists were reduced to those uniquely upregulated in each. This revealed 75, 129, and 117 proteins as uniquely significantly upregulated in P1, P2, and P3 LMS respectively (**Figure 5.3A**). These proteins were used to construct protein-protein interaction (PPI) networks to allow inspection of the molecular subtype-specific

LMS proteomes. Networks were built based on interaction scores of the STRING Database (STRINGdb), which provides interaction measures based on biological database knowledge, experimental data, and literature records[602,603]. In each network, highly clustered and interconnected regions were manually inspected. The P1 PPI

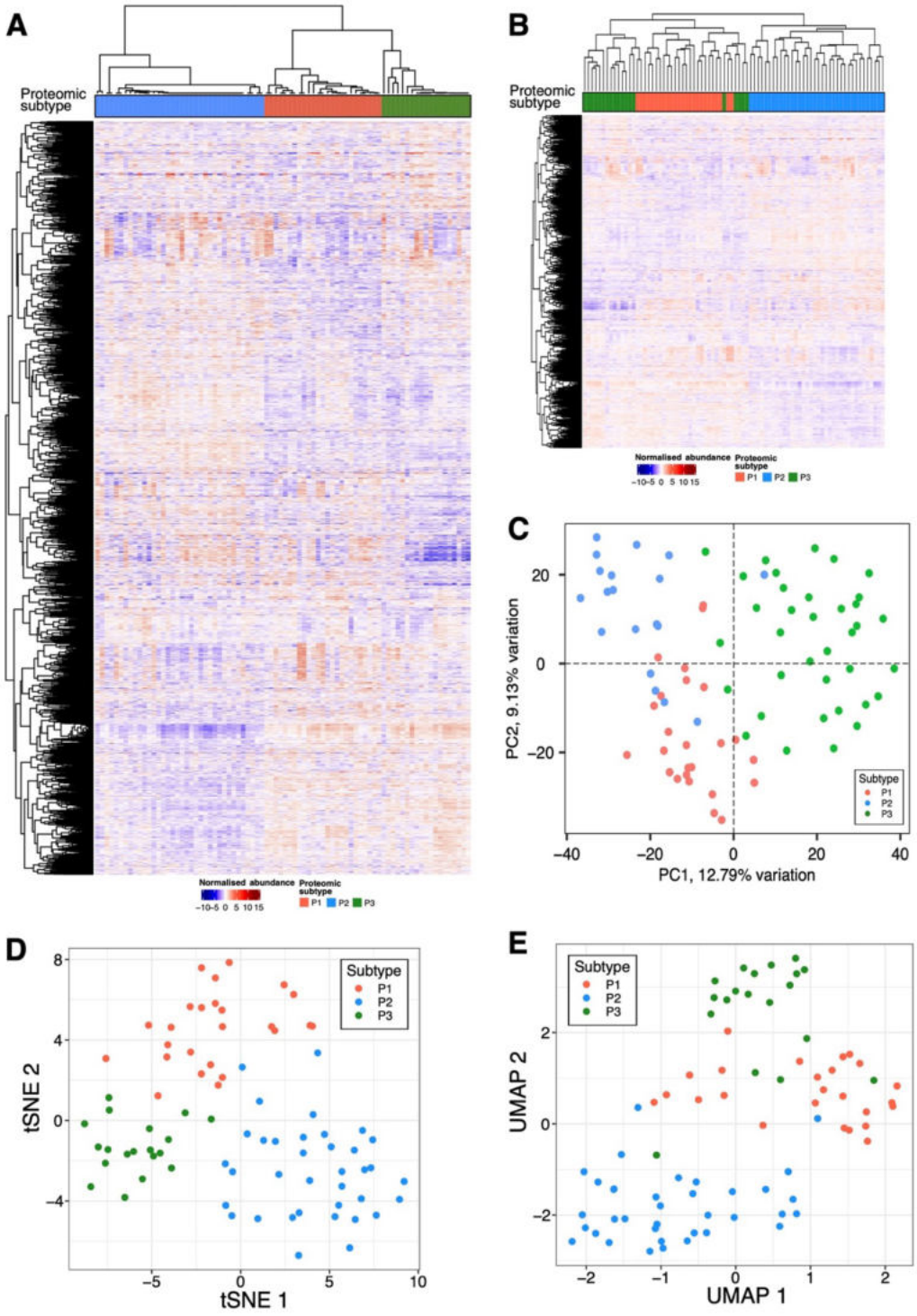


Figure 5.2 Proteomic subtypes of leiomyosarcoma (LMS)
(A) Heatmap showing the LMS consensus cluster dendrogram (at k = 3) and unsupervised clustering (Pearson's distance) of 3,263 proteins across 80 LMS cases. Top annotation panel indicates the proteomic subtypes of LMS **(B)** Heatmap showing the unsupervised clustering (Pearson's distance) of 80 LMS cases and 3,263 proteins. Top annotation panel indicates the proteomic subtypes of LMS **(C-E)** Dimension reduction of the proteomic data with individual cases coloured by proteomic subtype, using **(C)** principal component analysis (PCA), **(D)** t stochastic neighbour embedding (tSNE), and **(E)** uniform manifold approximation and projection (UMAP).

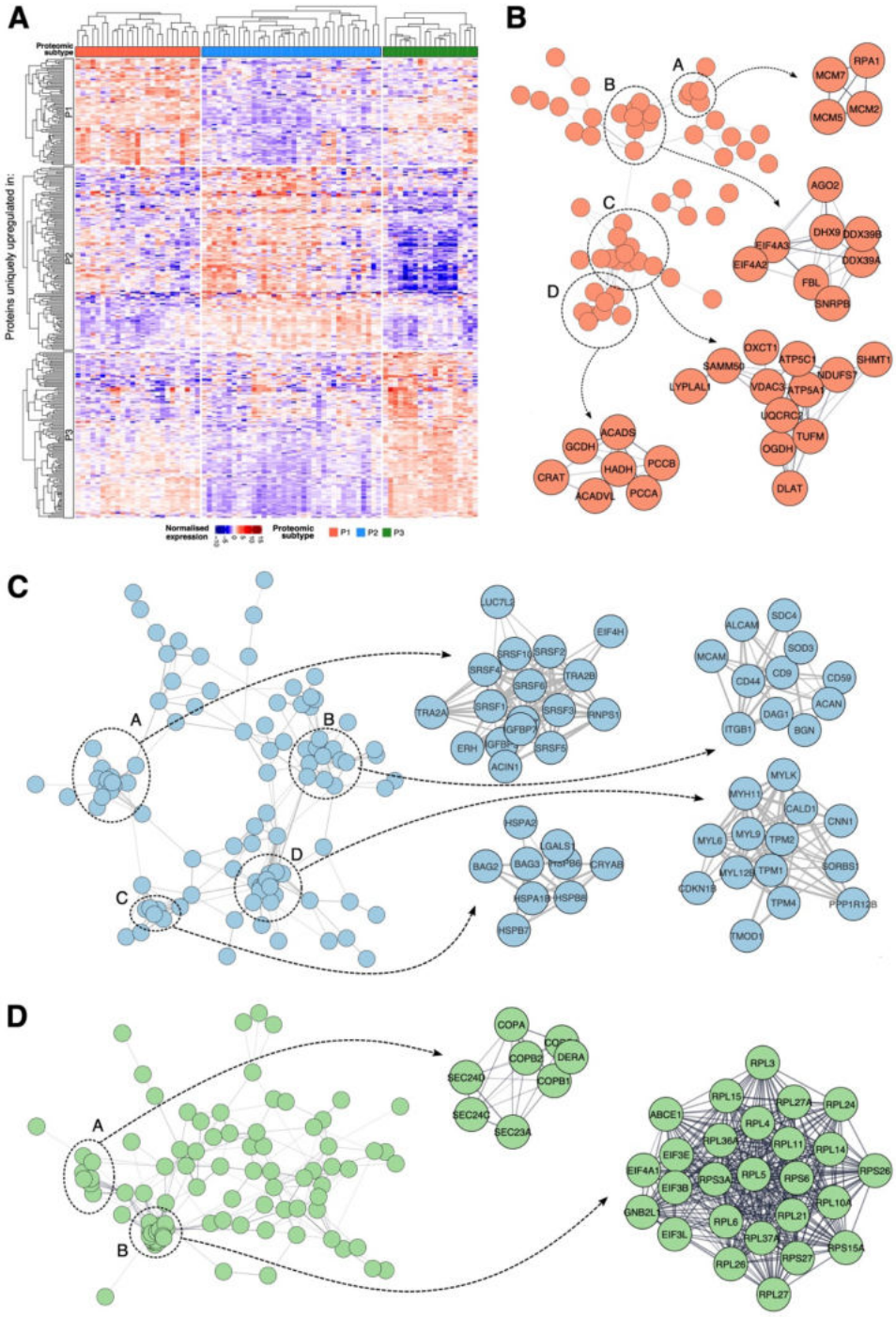


Figure 5.3 Leiomyosarcoma (LMS) proteomic subtype specific proteins
(A) Heatmap showing the supervised clustering of 321 differentially expressed proteins (DEPs) uniquely upregulated in each proteomic subtype of LMS. Top annotation panel indicates the proteomic subtypes of LMS, and the left annotation panel indicates which proteomic subtype the DEPs correspond to **(B-D)** Protein-protein interaction (PPI) networks, of **(B)** P1-, **(C)** P2-, and **(D)** P3-specific proteins. Regions of interest circled, and subnetworks constructed.

network comprised 4 clustered subnetworks (**Figure 5.3B**). Of these, 1 (subnetwork A) comprised components of the replication protein A (RPA) and minichromosome maintenance protein (MCM) complexes. RPA and MCM are both key to DNA replication activity and are considered pro-proliferative[604,605]. This suggests P1 may harbour a more proliferative phenotype. Subnetwork B comprised translation initiation factors and RNA processing proteins. Subnetworks C and D comprised predominantly mitochondrial enzymes. The P2 PPI network also comprised 4 subnetworks (**Figure 5.3C**). Here, subnetwork A contained several splicing factors and regulators. Subnetwork B contained cell adhesion and migration proteins, and subnetwork C contained mostly heat shock proteins. Interestingly, subnetwork D comprised muscle-specific proteins. This suggests P2 may show a more prominent smooth muscle phenotype relative to P1 and P3. The P3 PPI network comprised 2 subnetworks (**Figure 5.3D**). Subnetwork A exclusively contained proteins involved in vesicle budding: coatomers and coat protein complex II (COPII) proteins. Subnetwork B was notably large, comprising 26 proteins, most of which were ribosomal. Taken together, assessment of the proteins specific to each proteomic subtype of LMS suggest clear biological distinctions.

To investigate whether each set of DEPs contained shared biology, the significantly up and downregulated proteins were assessed by over-representation analysis against the gene ontology and hallmark gene sets (MSigDB)[506–508,512]. To ensure the robustness of any results, a background of the 3,263 LMS dataset proteins was used for analysis, as opposed to the whole genome. Over-representation identified no significant results in any proteomic subtype of LMS. This may be due to the small number of DEPs, or the reduced 3263-protein background that has limited coverage of the gene sets themselves. As an alternate and complementary method for exploring broad biological features, ssGSEA was performed[504,509]. As in DEP analysis, ssGSEA enrichment scores between samples were compared to identify differentially expressed biological features. To assess for such differences, ANOVA and post-hoc Tukey's multiple comparisons tests were used. This revealed 10 hallmarks as differentially expressed at the ssGSEA enrichment score level (Tukey's $p \leq 0.001$; **Figure *5.4***). P1 showed a notable and significant downregulation of all immune hallmarks ('allograft rejection', 'IL2 STAT5 signalling', 'complement', and 'inflammatory response') compared to P2 and P3. These immune hallmarks describe inflammatory signatures. P1 was therefore named the 'immune cold LMS' proteomic subtype. P3 showed significant downregulation of the hallmarks 'spermatogenesis' and 'myogenesis' relative to both P1 and P2. The biological basis for downregulated 'spermatogenesis' activity is unclear. The myogenesis hallmark captures genes involved in muscle development, suggesting tumours classified as P3

show lower expression of muscle specific proteins. Considering LMS are tumour of smooth muscle lineage this is indicative of poor differentiation or dedifferentiation. Accordingly, P3 was termed the 'dedifferentiated LMS' proteomic subtype. P2 showed upregulation of 'apoptosis'. Yet beyond this, P2 lacked any defining feature. Whilst the hallmark myogenesis was not significantly upregulated in P2, PPI analysis did identify specific muscle-related proteins as upregulated. P2 was therefore denoted the 'classical LMS' proteomic subtype.

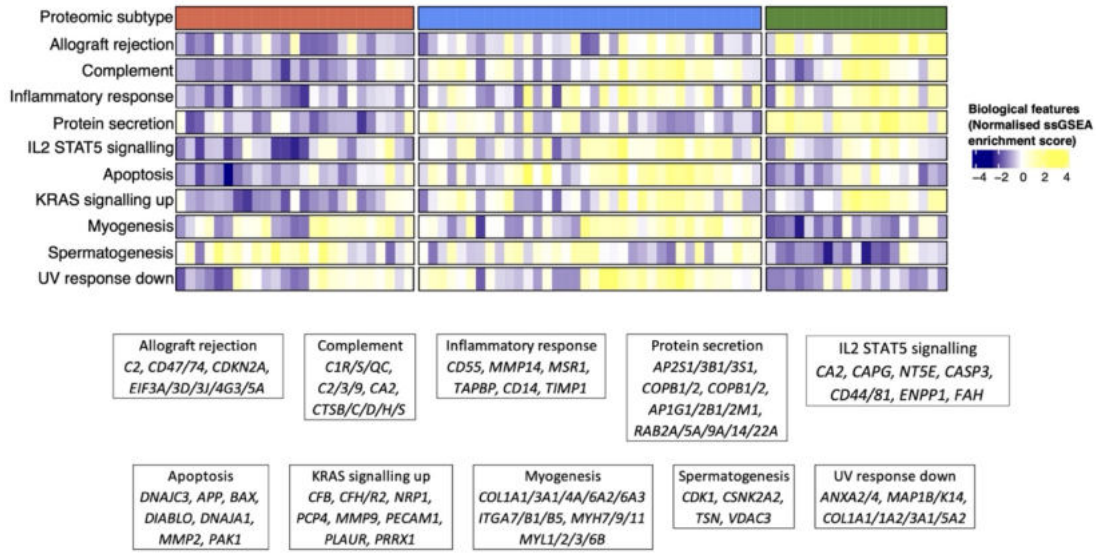


Figure 5.4 Hallmarks of the leiomyosarcoma (LMS) proteomic subtypes
Heatmap of significant (one-way ANOVA & Tukey's honestly significant difference (HSD) test; FDR < 0.001) biological features obtained from single sample Gene Set Enrichment Analysis (ssGSEA) of the MSigDB Hallmark gene sets, arranged by proteomic subtype (top annotation). Select enriched proteins from each hallmark are detailed in boxes.

5.2.1.5 Validation of immune cold and dedifferentiated LMS

To further investigate the characteristics of the identified immune cold LMS subtype, IHC data on TMAs generated and collected by previous lab members for a subset of MS-profiled cases was re-analysed. IHC data was available for 64/80 LMS samples in the cohort herein, covering assessment of CD3, CD4, and CD8; markers of total, helper, and cytotoxic T cell populations respectively. Data spanned 5 TMAs containing multiple 1 mm cores from each sample. To account for intra-tumoural heterogeneity, IHC measures were required to be available from at least 2 TMA cores per sample. Most samples had usable data from 3 cores, however for 1 sample, data from only 1 core was available (**Supplemental Figure 5.8A**). This case was excluded leaving a dataset of 63 for analysis. IHC scores were adjusted to TIL/mm^2 (**section 2.6)**, and the mean of all scores

used as the final sample measure. To assess to appropriateness of using mean in this data, the inter-core variability was explored. Across all markers, the largest differences between individual core measures and the mean value were observed in samples with higher immune infiltrate (**Supplemental Figure 5.8B**). This suggests that where immune infiltration is high within LMS tumours, spatial heterogeneity is observed. Such heterogeneity can undermine use of TMAs. In a study utilising data from the LMS samples profiled herein, intra-tumoural heterogeneity was investigated[606]. The authors found that whilst ≥ 11 cores were required for an estimate of absolute TILs, ≤ 3 cores were sufficient for the correct categorisation of most tumours into low and high based on the cohort median TIL counts. The mean count was therefore used, and data was assessed as both a continuous variable and dichotomised variable.

Overall, CD3+ TILs, CD4+ TILs, and CD8+ TILs all showed similar distributions across LMS samples. Density plots showed left tailing illustrating most samples have relatively low infiltrate of CD3/4/8+ cells (**Supplemental Figure *5.9***). CD3+ cells and CD4+ cells were present at higher levels than CD8+ cells, and there was a significant range of TIL burden across patients. The number of CD3+ TIL/mm^2 ranged from 0 – 843 TIL/mm^2 across the cohort (median = 72 TIL/mm^2), CD4+ ranged from 0 - 1040 TIL/mm^2 (median = 59 TIL/mm^2), and CD8+ ranged from 0 - 180 TIL/mm^2 (median = 16 TIL/mm^2). CD4+ TILs and CD8+ TILs are traditionally considered subpopulations of CD3+ TILs. In agreement with this, CD3+ TILs and CD8+ TILs, and CD3+ TILs and CD4+ TILs showed strong positive correlations (Pearson correlation coefficient = 0.94; $p < 0.001$, and Pearson correlation coefficient = 0.7; $p < 0.001$ respectively; **Supplemental Figure 5.8C**). CD4+ and CD8+ cells have complementary roles. As such they were positively correlated (Pearson correlation coefficient = 0.59; $p < 0.001$), although to a lesser extent than CD3/CD4 and CD3/CD8. As a total population, CD3+ TILs were expected to be higher than CD4+ TILs and CD8+ TILs. Yet, for 17 samples, the mean CD4+ cell infiltrate was higher than the mean CD3+ cell infiltrate. For 6 of these, CD4+ TILs were present at ≥ 1.5x the level of CD3+ TILs. Numerical interpretations between measures of total (CD3+) and helper (CD4+) T cell populations are therefore cautioned.

To assess TILs in the context of each proteomic subtype of LMS, density plots stratified by subtype were generated (**Figure 5.5A**). These showed most 'immune cold LMS/P1' tumours had very low levels of CD3+ TILs, CD4+ TILs, and CD8+ TILs. 'Classical LMS/P2' and 'dedifferentiated LMS/P3' showed a similar CD8+ TIL profile to 'immune cold LMS/P1' (Kruskal-Wallis test: $X^2 = 3.522$, $p = 0.1719$). However, in CD3+ TIL and CD4+ TIL measures, 'immune cold LMS/P1' showed a significantly lower TIL burden

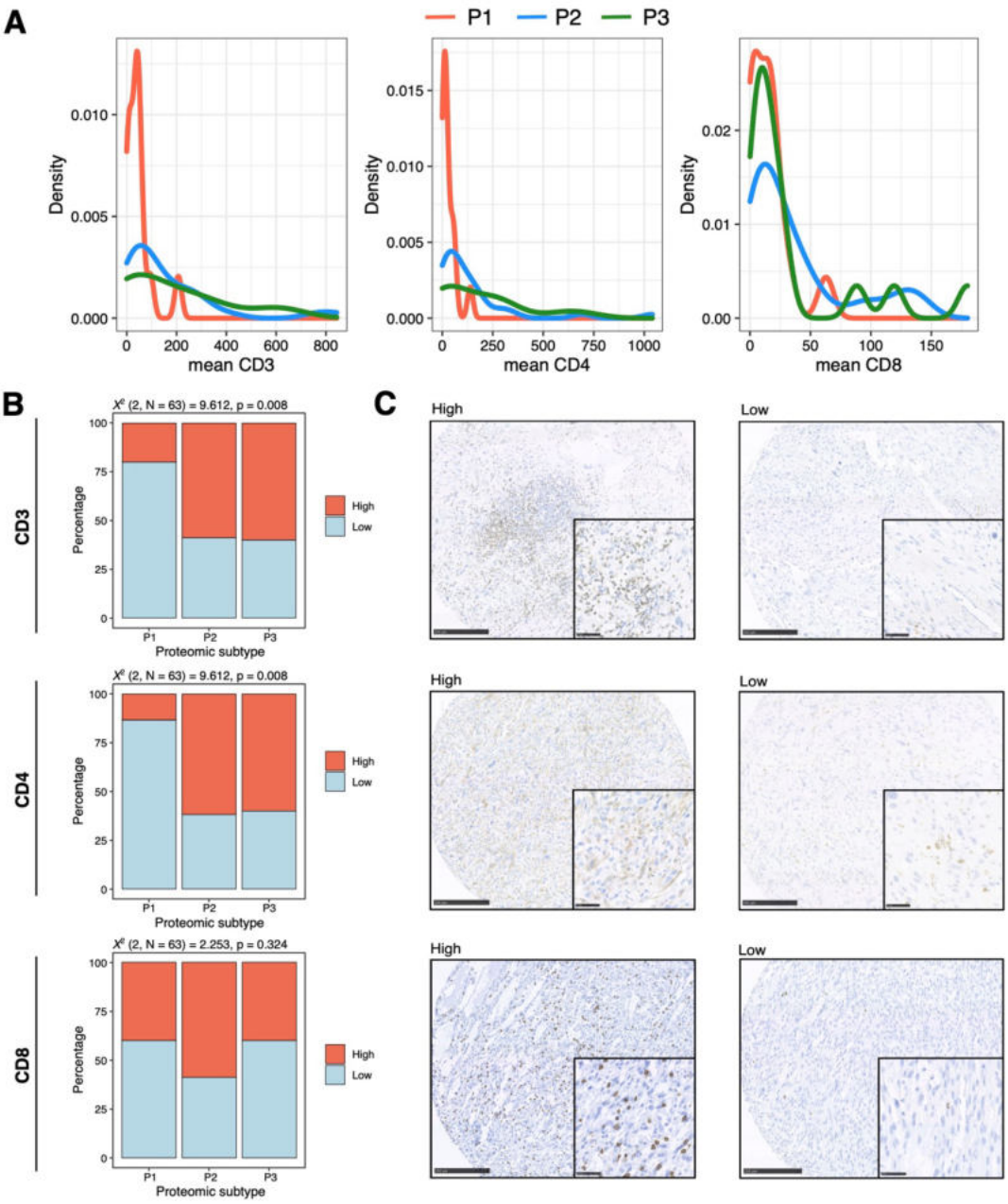


Figure 5.5 Characterisation of the tumour infiltrating lymphocyte (TIL) burden of leiomyosarcoma (LMS) proteomic subtypes.
(A) Density plots of CD3+/4+/8+ TILs across the 3 proteomic subtypes of LMS. **(B)** Stacked bar plots showing the proportion of high and low CD3+/4+/8+ TILs across each of the 3 proteomic subtypes of LMS. Samples were categorised as high and low based on median TIL density. Chi-squared test results reported at the top of each plot. **(C)** Representative images of high and low CD3+/4+/8+ TIL staining by immunohistochemistry in exemplar LMS tissue specimens.

(Kruskal-Wallis tests: CD3 X^2 = 6.442, p = 0.039; CD4 X^2 = 7.686, p = 0.0214). Interestingly, density plots revealed a subset of 'immune cold LMS/P1' tumours to have moderate CD3+, CD4+, and CD8+ TIL levels, as seen by a second peak at

approximately 200, 140, and 65 TIL/mm^2 respectively. This suggests some heterogeneity is present within 'immune cold LMS/P1', however relative to the other LMS

proteomic subtypes, levels of CD3+ and CD4+ TILs were consistently low. For further investigation and considering low numbers of cores have been shown as sufficient for categorised TIL counts, data was dichotomised at the median and reanalysed[606]. Consistent with use of the continuous variable form, this revealed the 'immune cold LMS/P1' subtype to comprise a significantly higher proportion of CD3+ low and CD4+ low tumours (Chi-squared tests: $X^2 = 9.612$, $p = 0.008$; **Figure 5.5B-C**). No significant difference in the proportions of CD8+ low and CD8+ high tumours was seen across the LMS proteomic subtypes (Chi-squared test: $X^2 = 2.253$, $p = 0.324$).

Subtype P3 was identified as dedifferentiated LMS. LMS is of smooth muscle origin and smooth muscle markers are routine in its diagnosis[237–239,542]. A dedifferentiated phenotype indicates loss or absence of such proteins. Indeed, the expression of known smooth muscle markers was markedly lower in 'dedifferentiated LMS/P3' (**Figure 5.6A**)[542]. Specifically, CLF2, SLMAP, ACTA2, MYLK, and MYH11 were significantly lower compared to both other molecular subtypes. Desmin was significantly lower compared to 'classical LMS/P2', and CALD1 was significantly lower in 'dedifferentiated LMS/P3' and 'immune cold LMS/P1' compared to 'classical LMS/P2'. To assess this dedifferentiation of LMS in the context of other STS, UMAP was utilised. This illustrated 9 LMS to cluster away from the bulk LMS tumour cluster. (**Figure 5.6B**). All except 1 of the LMS tumours clustering outside of this group were 'dedifferentiated LMS/P3'. Given, UPS is a tumour with no identifiable differentiation lineage, previous studies have suggested a disease spectrum between UPS and LMS tumours showing dedifferentiation[243,244]. However in my dataset, no co-clustering between UPS and 'dedifferentiated LMS/P3' was observed in the UMAP analysis (**Figure 5.6C**).

5.2.1.6 Clinical characterisation of LMS proteomic subtypes

The transcriptomic subtypes of LMS have been suggested to be reflective of anatomical site[36,274,283]. Specifically, a uterine-enriched subtype has been repeatedly noted, and the most recent works by Anderson *et al* have suggested transcriptomic subtypes of LMS correspond to the lineages of vascular, gynaecological and digestive tissue. To explore whether this is true for the proteomic subtype, and to assess whether other clinicopathological features are associated with subtype, Chi-squared and Kruskal-Wallis tests were conducted. No statistical association between the proteomic subtypes of LMS and any clinicopathological variable was identified (**Supplemental Table 5.4** and **Figure**

5.7A), however trends in anatomical distribution were observed (**Supplemental Table *5.4*** and **Figure 5.7B**). Due to the small number of trunk wall LMS, the representation of proteomic subtypes at this anatomical site could not be assessed. Overall, the distribution of 'classical LMS/P2' was similar across all anatomical sites, yet the proportion of 'immune cold LMS/P1' and 'dedifferentiated LMS/P3' differed.

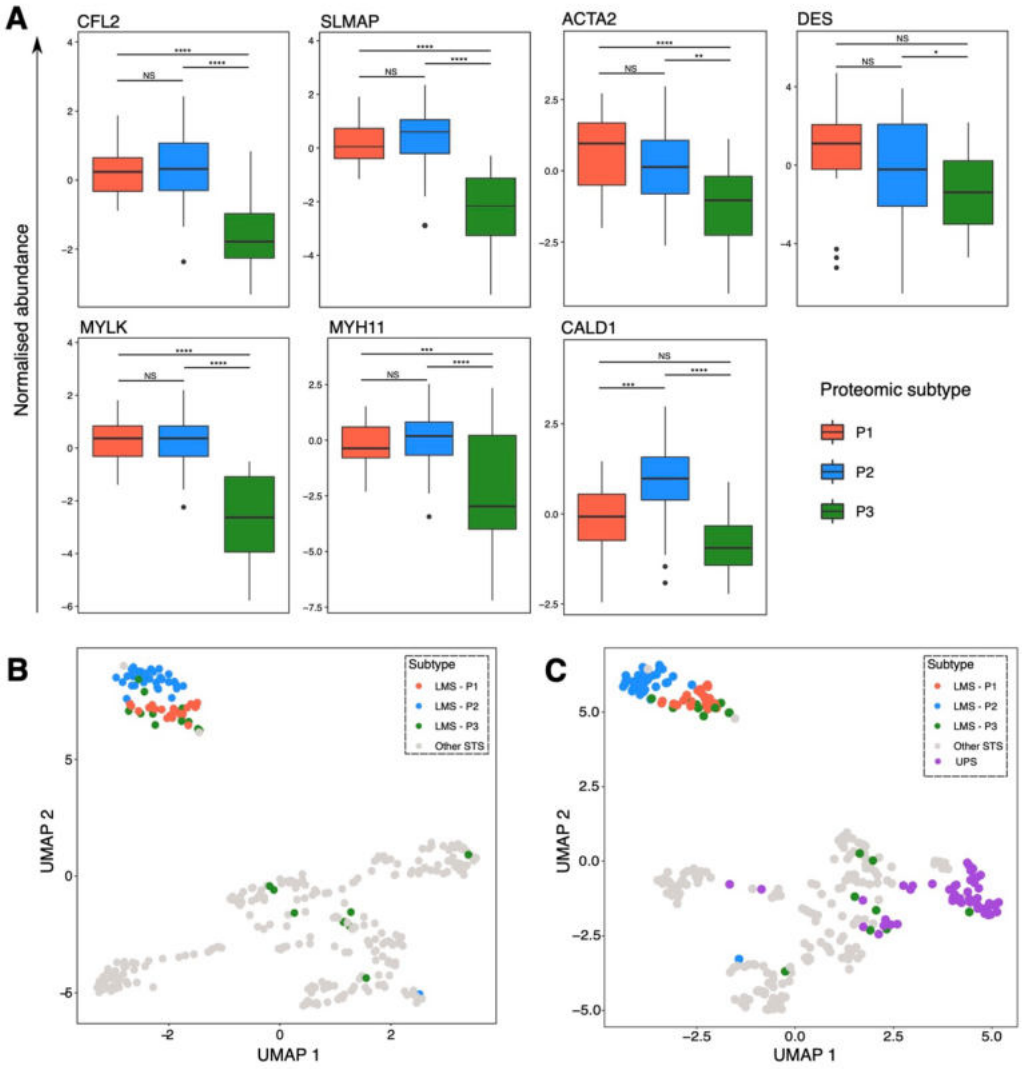


Figure 5.6 Characterisation of the dedifferentiated (P3) leiomyosarcoma (LMS) proteomic subtype
(A) Boxplots comparing expression of a subset of smooth muscle proteins between the three LMS proteomic subtypes. Boxes indicate 25th and 75th percentile, with median line in the middle, whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. Significance determined by Tukey's honestly significant difference (HSD) tests. NS = not significant, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$ **(B-C)** Uniform manifold approximation and projection (UMAP) plot showing clustering of the three LMS proteomic subtypes in relation to other soft tissue sarcomas (STS) samples. **(B)** Other STS in grey, **(C)** UPS in purple and other STS in grey.

Retroperitoneal and intra-abdominal LMS tumours showed an over-representation of the 'immune cold LMS/P1' subtype, which accounted for 48% and 50% of tumours respectively. Of the remaining tumours, all except 1 retroperitoneal and 1 intra-abdominal were classified as 'classical LMS'. By contrast, 'dedifferentiated LMS/P3' accounted for approximately 1/3rd of pelvis, uterine, and extremity tumours.

The clinical implications of LMS molecular subtyping are currently unclear. It is hypothesised that the distinctive biology observed between molecular subtypes may contribute to differences in disease progression and clinical course for the patient. Therefore, the LMS proteomic subtypes were assessed in the context of patient outcome (LFRS, MFS, and OS). Univariable analysis (Cox regression; summarised in **Table 5.2**) showed no significant association between any subtype and any outcome measure, and all PH assumptions were met. However, trends were observed by inspection of the Kaplan Meier curves (**Figure 5.7C**). 'Dedifferentiated LMS/P3' appeared to show an increased risk of local recurrence between 1- and 5-years following surgery. Similarly, 'dedifferentiated LMS/P3' showed a shorter median MFS (1.9 years) than 'immune cold LMS/P1' (2.8 years) and 'classical LMS/P2' (3.9 years). In LMS care, local recurrence is considered a relatively low risk event for patients, with metastases more common and the cause of fatality[607,608]. Yet when LMS proteomic subtypes were compared to the full STS cohort, it was evident that a population of LMS ('dedifferentiated LMS/P3') are at higher risk of local recurrence compared to other LMS. The Kaplan Meier curve suggested 'immune cold LMS/P1' and 'classical LMS/P2' possess the longest LRFS of all histological and proteomic subtypes assessed (**Supplemental Figure *5.10***). 'Dedifferentiated LMS/P3' showed a LRFS comparable to that of SS, UPS, and EPS. However, the sample size was limited, and these similarities were not statistically significant based on the univariable Cox regression.

Further to the univariable analyses, multivariable Cox regression was also performed to adjust for other clinicopathological variables and assess whether the significance of LMS proteomic subtype exists independent of these variables (summarised in **Table 5.3**). Whilst proteomic subtype remained non-significant in OS analyses, it was found to be a significant prognostic marker for LRFS and MFS. Specifically, 'dedifferentiated LMS/P3' was found to be associated with a significantly poorer LRFS compared to 'immune cold LMS/P1' (HR = 8.04, 95% CI = 1.7 – 38, p = 0.009). Additionally, patient age, sex (male), and a PS of 1 were all significant features associated with a poorer LRFS in multivariable analysis (HR = 0.956, CI = 0.918 – 0.996, p = 0.031, HR = 5.554, 95% CI = 1.537 - 20.002, p = 0.009, and HR = 7.819, 95% CI = 1.745 - 35.04, p = 0.007 respectively).

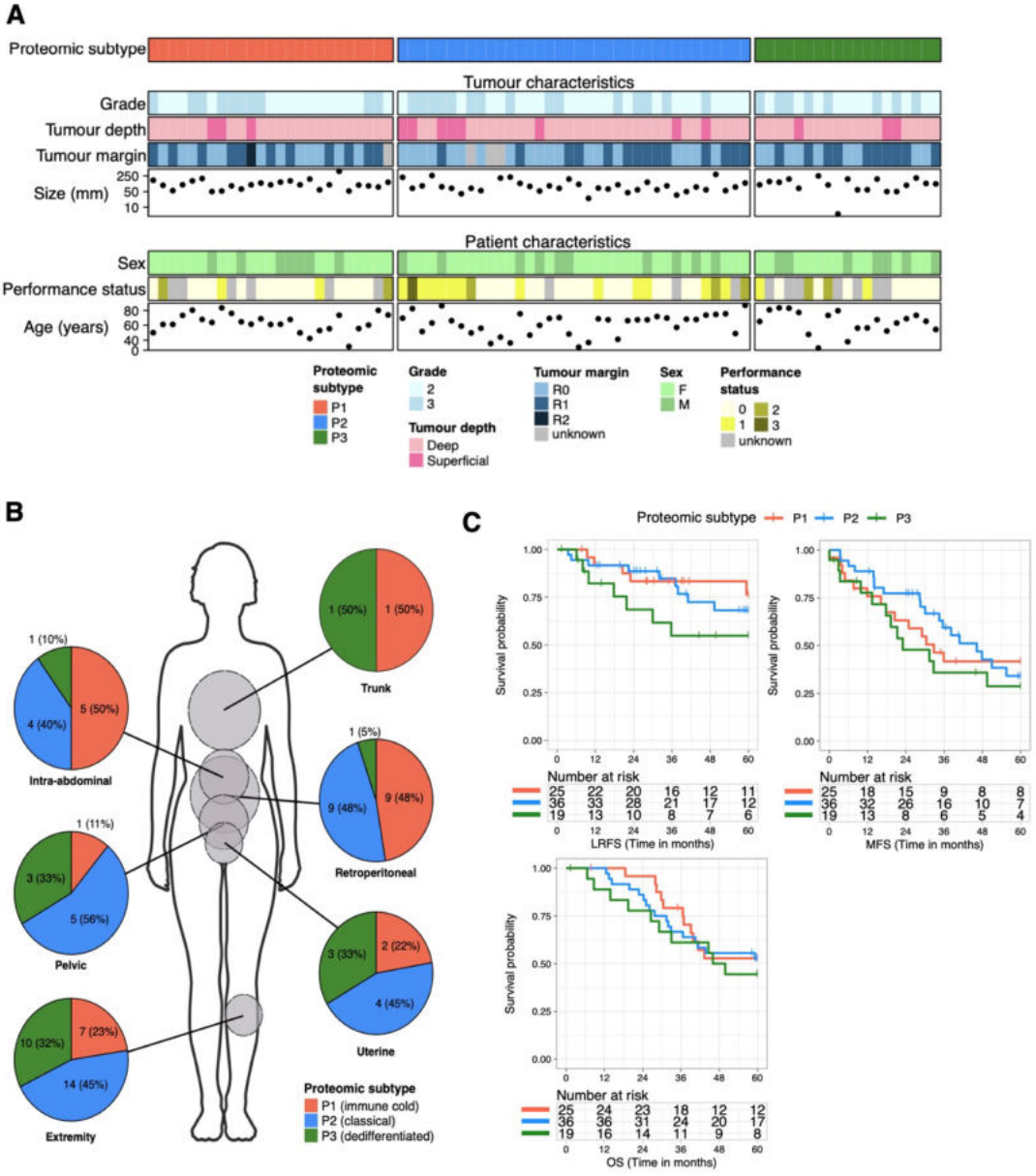


Figure 5.7 Clinical characterisation of leiomyosarcoma (LMS) proteomic subtypes
(A) Summary plot showing clinicopathological variables across LMS cases (n=80), arranged by proteomic subtype. **(B)** Pie charts depicting the breakdown of LMS proteomic subtypes at different anatomical sites. **(C)** Kaplan-Meier plot of local recurrence free survival (LRFS) metastasis free survival (MFS), and overall survival (OS) up to 5-years post-surgery across the LMS proteomic subtypes.

Tumours of the extremities and trunk wall were associated with a significantly improved LRFS (HR = 0.089, 95% CI = 0.018 - 0.452, p = 0.004). In the MFS model, 'dedifferentiated LMS/P3' was associated with a significantly poorer MFS compared to 'immune cold LMS/P1' (HR = 2.629, 95% CI = 1.065 - 6.489, p = 0.036). Grade was also associated with a significantly poorer MFS (HR = 3.282, 95% CI = 1.568 - 6.867, p =

0.002), and anatomical site ('trunk wall and extremity') and tumours of the largest size category (> 5 log(mm)) were associated with an improved MFS (HR = 0.316, 95% CI = 0.131 - 0.76, p = 0.01 and HR = 0.306, 95% CI = 0.115 - 0.815, p = 0.018 respectively). A minor PH violation was observed in the MFS model form LMS proteomic subtype (p =

0.033; **Supplemental Figure *5.11***), all other variables satisfied PH. The shift in significance and effect size of LMS proteomic subtype between univariable and multivariable analyses is of note. As in detailed in **section 5.2.1.2**, this can suggest statistical suppression. The cause of this suppression, if present, is unclear as no clinicopathological variable was statistically associated with proteomic subtype (**section 5.2.1.6**). The strongest trends were observed between subtype and anatomical site, although notably the anatomical site groupings were transformed for survival analyses. Analysis of the relationship between the transformed anatomical groups and proteomic subtype revealed no significant association (Chi-squared test: $X^2 = 5.092$, d.f = 4, p = 0.278). Irrespective of suppression, it is evident that the proteomic subtypes of LMS hold prognostic value for LRFS and MFS. To assess the importance of proteomic LMS subtype in the multivariable models, each variable was added sequentially, and ANOVA used to compare the fit of the sequential models. The multivariable models with and without the LMS proteomic subtype variable were compared. There was no significant improvement in model fit between the MFS clinicopathological only model (**section 5.2.1.2**) and the model inclusive of proteomic subtype (ANOVA $X^2 = 4.461$, d.f = 2, p = 0.108). However, the LRFS model including proteomic subtype was shown to fit the data significantly better than the model without proteomic subtype (**section 5.2.1.2**; ANOVA $X^2 = 8.752$, d.f = 2, p = 0.013).

Table 5.2 Univariable Cox regression assessing leiomyosarcoma (LMS) proteomic subtypes.
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval.

		LRFS		MFS		OS	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
Proteomic subtype	*P2 (ref)*	-	-	-	-	-	-
	P1	0.761 (0.255-2.27)	0.625	1.14 (0.571-2.27)	0.711	0.915 (0.428-1.95)	0.819
	P3	1.86 (0.691-5)	0.219	1.52 (0.737-3.13)	0.258	1.25 (0.57-2.72)	0.582

Given the distinct biology of the LMS proteomic subtypes, it is hypothesised that subtypes may show differential drug responses. Indeed, a previous LMS transcriptomic study has alluded to targetable proteins specific to LMS transcriptomic subtypes[282]. However, this was based on the expression of individual differentially expressed genes

and did not statistically consider the full targeting profiles of drugs. To assess drug profiles herein, ssGSEA was performed on the LMS proteome dataset using the DSigDB D1 database[511]. Clustering of D1 NES across LMS proteomic subtypes did not reveal any obvious association between proteomic subtype and drug target profiles (**Figure 5.8**).

Table 5.3 Multivariable Cox regression assessing leiomyosarcoma (LMS) proteomic subtypes. Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. Anatomical site of 'Other' indicates retroperitoneal, Intra-abdominal and pelvic cases. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval

		LRFS		MFS		OS	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
Age at excision (years)		**0.956 (0.918-0.996)**	**0.031**	1 (0.974-1.03)	0.894	0.983 (0.953-1.01)	0.291
Sex	*F (ref)*	-	-	-	-	-	-
	M	**5.54 (1.54-20)**	**0.009**	1.03 (0.442-2.4)	0.944	**3.1 (1.35-7.11)**	**0.008**
Anatomical site	*Other (ref)*	-	-	-	-	-	-
	Extremity & trunk wall	**0.089 (0.018-0.452)**	**0.004**	**0.316 (0.131-0.76)**	**0.01**	**0.258 (0.095-0.701)**	**0.008**
	Uterine	0.148 (0.01-2.29)	0.172	1 (0.278-3.62)	0.996	0.57 (0.166-1.95)	0.37
FNCLCC grade	*2 (ref)*	-	-	-	-	-	-
	3	2.89 (0.926-9)	0.068	**3.28 (1.57-6.87)**	**0.002**	**2.74 (1.25-6.03)**	**0.012**
Performance status	*0 (ref)*	-	-	-	-	-	-
	1	**7.82 (1.74-35)**	**0.007**	1.91 (0.749-4.86)	0.176	2.69 (0.986-7.35)	0.053
	2-3	5.52 (0.334-91.1)	0.233	2.14 (0.537-8.53)	0.281	**24.6 (6.29-96.4)**	**< 0.001**
	unknown	1.59 (0.361-6.96)	0.542	0.665 (0.254-1.74)	0.406	0.976 (0.329-2.9)	0.965
Tumour depth	*Deep (ref)*	-	-	-	-	-	-
	Superficial	0.514 (0.036-7.41)	0.625	0.418 (0.087-2.01)	0.276	1.3 (0.303-5.61)	0.722
Tumour margin	*R0 (ref)*	-	-	-	-	-	-
	R1 & R2	1.8 (0.628-5.14)	0.274	0.821 (0.414-1.63)	0.574	1.27 (0.605-2.65)	0.532
Log(Tumour size [mm])	4-5 (ref)	-	-	-	-	-	-
	< 4	1.18 (0.158-8.73)	0.875	0.55 (0.135-2.24)	0.404	0.85 (0.172-4.2)	0.842
	> 5	0.538 (0.14-2.07)	0.368	**0.306 (0.115-0.815)**	**0.018**	1.07 (0.434-2.63)	0.887
Proteomic subtype	*P2 (ref)*	-	-	-	-	-	-
	P1	1.19 (0.308-4.61)	0.8	1.61 (0.719-3.59)	0.248	1.08 (0.441-2.66)	0.862
	P3	**8.04 (1.7-38)**	**0.009**	**2.63 (1.07-6.49)**	**0.036**	2.18 (0.819-5.78)	0.119

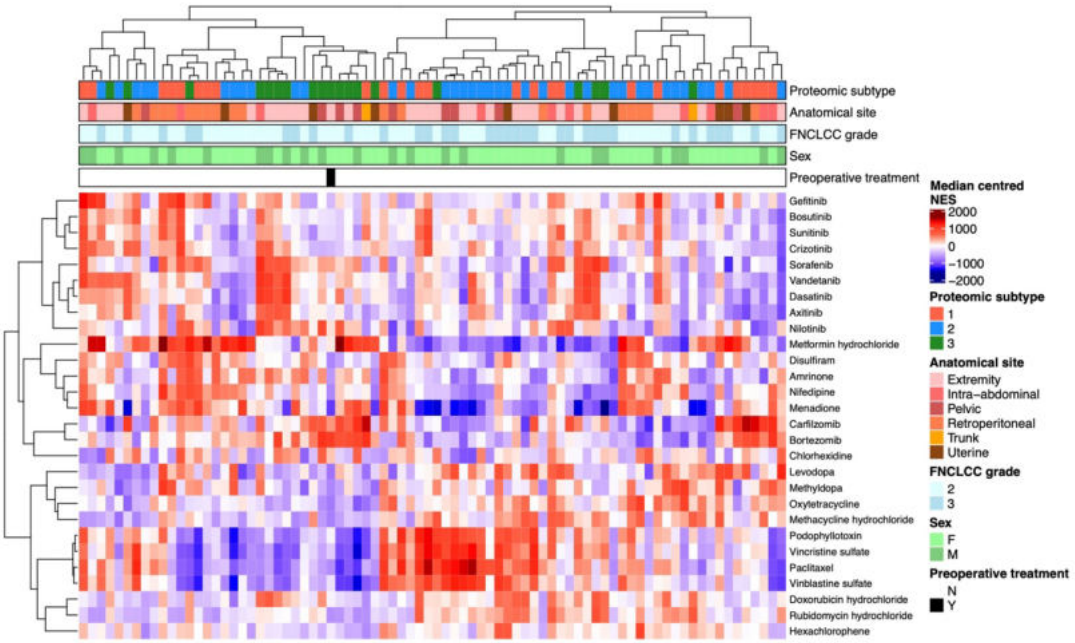


Figure 5.8 Drug target profile expression across leiomyosarcoma (LMS) proteomic subtypes
(A) Annotated heatmap showing the unsupervised clustering (Pearson's distance) of 27 Drug Signature database (DSigDB) D1 profiles across the LMS cohort. From top to bottom, panels indicate proteomic subtype, anatomical site, tumour grade, patient sex, and preoperative treatment status.

5.2.1.7 Comparison of the proteomic and transcriptomic subtypes of LMS

There are several transcriptomic studies describing molecular LMS subtypes (as discussed in **section 1.4.1.2**). It was therefore of interest to assess whether the proteomic LMS subtypes recapitulate, complement, or contrast transcriptomic subtypes. It was not possible to perform transcriptomic profiling on the cohort herein, nor is there a publicly available dataset of both transcriptomic and MS data for LMS patients. A direct assessment of whether the proteomic subtypes are recapitulating the transcriptomic subtypes was therefore not possible.

With these limitations in mind, the TCGA RNAseq data was queried based on the proteins identified herein; to offer an informal comparison of the transcriptomic and proteomic subtypes[36]. Whilst individual RNA-protein correlations are typically poor, the expression of groups of RNA/proteins representing overarching biological features is hypothesised to have higher similarity[448,609,610]. Therefore, by assessing numerous genes/proteins, it is anticipated that, if present, similarities between the proteomic and transcriptomic subtypes will be observed. The TCGA dataset was selected due to its

annotation with transcriptomic subtypes from 3 independent LMS studies, thus facilitating multi-study comparisons. The annotations describe 2 LMS subtypes from Abeshouse *et al*, 3 LMS subtypes from Hemming *et al*, and 4 LMS subtypes from Anderson *et al*[36,274,283]*.* To reveal whether cohort differences may impact cross-study translation of findings, the TCGA cohort and MS cohort were compared (summarised in **Supplemental Table *5.5***). Both cohorts comprised primary tumours from 80 LMS patients. The cohorts had a near identical number of females and males. However, all other overlapping clinicopathological variables assessed differed significantly. The significant difference of anatomical site was mostly attributable to the inclusion of more uLMS and fewer extremity tumours in TCGA. Tumour depth differences were mostly due to a high missingness for this variable in TCGA and low numbers of superficial tumours. Tumour margins were different due to more R1 margins in the MS cohort, and the distribution of grade was different due to the exclusion of grade 1 tumours in the MS cohort, and a coordinate increase in the inclusion of high grade tumours. The TCGA and MS cohort therefore represent two different patient populations within LMS. As a result, it is possible that intrinsic tumour biology may differ between the populations, restricting comparisons of the proteomic and transcriptomic subtypes.

To assess whether the proteins identified herein can recapitulate the transcriptomic heterogeneity of LMS, the TCGA RNAseq dataset was reduced to only those proteins/genes identified by MS (n = 3,290). Strikingly, the clustering achieved by this reduced gene list in the TCGA cohort was highly comparable to the LMS transcriptomic subtypes identified by Abeshouse *et al,* Hemming *et al,* and Anderson *et al* (**Figure *5.9*A**). There were some exceptions. Most notable, a subset of 'Abeshouse stLMS-like', 'Hemming cLMS' and 'Anderson C2A/B' clustered away from most other tumours of these classifications. However, within the 2 main clusters of 'Abeshouse uLMS-like' and 'Abeshouse stLMS-like'; the subtypes identified by Hemming *et al* and Anderson *et al* (iLMS and uLMS; C1 and C3) clustered separately within 'Abeshouse uLMS-like', as did the further subtypes identified by Anderson *et al* (C2A and C2B) within 'Abeshouse stLMS-like'. To assess whether these transcriptomic patterns are driven by the same DEPs as the proteomic subtypes and to investigate whether the proteomic and transcriptomic subtypes are the same; the TCGA data was reduced to DEPs uniquely upregulated between the proteomic subtypes (n = 321). Clustering of this dataset generated comparable results to the data from 3290 genes (**Figure *5.9*B** vs **Figure *5.9*A**). The heatmap showed 2 near exclusive clusters of 'Abeshouse uLMS-like' and 'Abeshouse stLMS-like' cases. Most 'Hemming uLMS' and 'Anderson C3' clustered as a subset of 'Abeshouse uLMS-like', and most 'Anderson C2B' and 'C2A' were clustered s

a subset of ‘Abeshouse stLMS-like’ and ‘Hemming cLMS’. However, the clustering of transcriptomic subtypes was less robust; in particular ‘Anderson C2B’ cases clustered across ‘Abeshouse stLMS-like’ and ‘Hemming cLMS’. Whilst use of the proteomic subtype DEPs did reproduce similar clusters to the published transcriptomic subtypes of LMS, visual inspection of the heatmap suggested each set of subtype-specific proteomic DEPs did not correspond to a particular transcriptomic subtype. Based on this, it was not possible to determine which proteomic subtypes map to which transcriptomic subtypes.

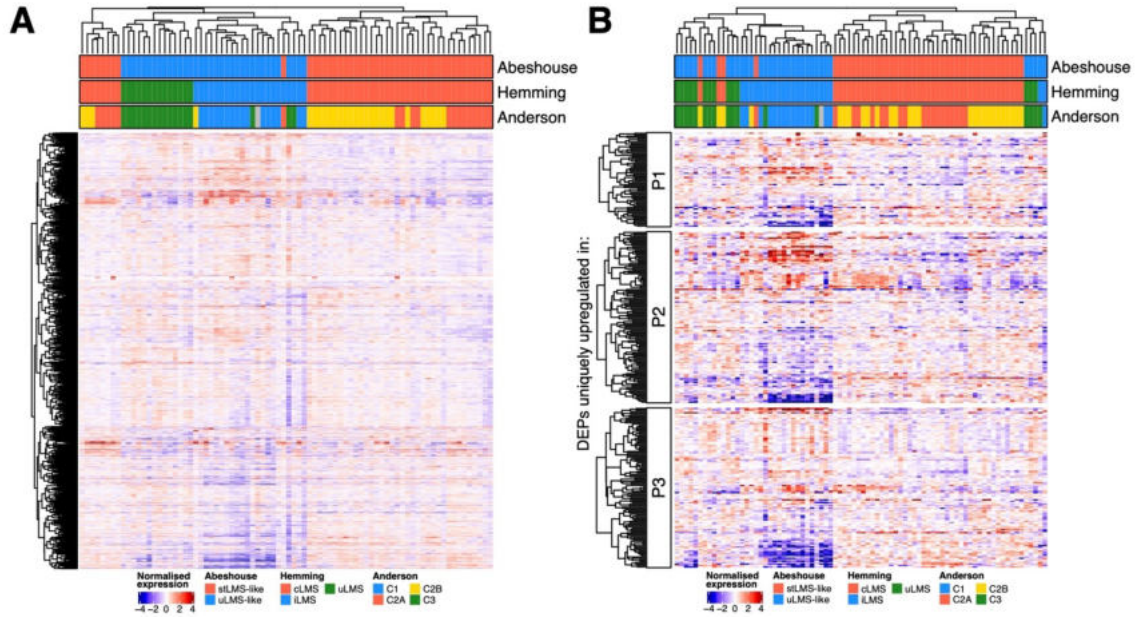


Figure 5.9 Leiomyosarcoma (LMS) proteomic subtypes in The Cancer Genome Atlas (TCGA) RNAseq cohort.
(A) Annotated heatmap showing the unsupervised clustering (Pearson’s distance) of all MS-identified proteins in the TCGA RNAseq dataset. Annotations show transcriptomic subtypes identified by Abeshouse et al, Hemming et al, and Anderson et al. **(B)** Annotated heatmap showing the clustering (Pearson’s distance) of all proteomic subtype specific differentially expressed proteins (DEPs; **Figure 5.3**) in the TCGA RNAseq dataset. Annotations show transcriptomic subtypes identified by Abeshouse et al, Hemming et al, and Anderson et al.[36,274,283]

5.2.2 The immune landscape of UPS and DDLPS

5.2.2.1 Clinicopathological features of the UPS and DDLPS cohort

Primary tumour specimens from 53 UPS and 39 DDLPS patients were profiled (total = 92). Clinicopathological features are summarised in **Table *5.4***. Briefly, patients had a median age of 68.6 years at the time of surgery and a median tumour size of 120 mm. There were approximately equal numbers of females and males (43 and 49 respectively), and a strong enrichment of high grade tumours (75% vs 23.9% intermediate grade). Most tumours were deep (88%) and located in either the extremities (40%) or retroperitoneum (32%). 5 UPS patients had radiation-associated disease, 2 DDLPS patients had metastatic disease at surgery, 1 DDLPS patient had multifocal disease at surgery, and 1 DDLPS patient received preoperative treatment (chemotherapy). Most surgical margins were either R0 (38%) or R1 (56.5%), and most patients had a PS of either 0 (42.4%) or 1 (29.3%). Data was not available for 16 patients with missing PS data, 5 patients with missing tumour margin data, and 1 patient with missing grade data. Interactions between clinicopathological features were assessed as before and are summarised in **Supplemental Table *5.6***. This revealed expected histological subtype differences. Most extremity tumours were UPS, and most retroperitoneal tumours were DDLPS (**Supplemental Figure 5.12A**). Histology was also significantly associated with grade, age, and size, where DDLPS were of lower grade (**Supplemental Figure 5.12B**), present in younger patients (**Supplemental Figure 5.12C)**, and larger than UPS (**Supplemental Figure 5.12D**). Beyond this, significant associations were noted as in the full cohort (**section 4.2.2**): significant between anatomical site and size, size and tumour depth, and age and grade. In addition, specific to the DDLPS and UPS cohort, an association between anatomical site and grade was noted; with grade 2 tumours almost exclusively intra-abdominal and retroperitoneal (**Supplemental Figure 5.12E**). Size was also associated with grade (lower grade tumours were larger; **Supplemental Figure 5.12F**), and age was associated with anatomical site (head and neck, and retroperitoneal tumours occurred in a younger population; **Supplemental Figure 5.12G**).

5.2.2.2 Cohort outcomes and the prognostic significance of clinicopathological variables

Survival data was censored at 5 years and therefore information on longer term outcomes was not available. Within the mixed DDLPS and UPS cohort, median LRFS was ~ 44 months, as was median OS (**Figure *5.10*A,C**). Median MFS for the cohort was not reached (**Figure *5.10*B**). At 5-years post-surgery, 46% of patients had experienced

Table 5.4 Clinicopathological features of the dedifferentiated liposarcoma (DDLPS) and undifferentiated pleomorphic sarcoma (UPS) cohort.
Features of total cohort and individual histological subtypes. Continuous variables detailed as median, minimum (min), and maximum (max). Categorical variables detailed as count (n) and percentage. Abbreviations: F = female; M = male; CTX = chemotherapy.

		Total	DDLPS	UPS
	n	**92**	**39**	**53**
Age at excision (years)	**median**	68.6	63	73.5
	min	28.2	35.1	28.2
	max	90	81.3	90
Tumour size (mm)	**median**	120	190	80
	min	15	35	15
	max	1090	1090	360
Sex [n (%)]	**F**	43 (46.7)	15 (38.5)	28 (52.8)
	M	49 (53.3)	24 (61.5)	25 (47.2)
Grade [n (%)]	**2**	22 (23.9)	19 (48.7)	3 (5.7)
	3	69 (75)	20 (51.3)	49 (92.5)
	unknown	1 (1.1)	-	1 (1.9)
Anatomical site [n (%)]	**Extremity**	40 (43.5)	2 (5.1)	38 (71.7)
	Head/neck	4 (4.3)	-	4 (7.5)
	Intra-abdominal	4 (4.3)	3 (7.7)	1 (1.9)
	Retroperitoneal	32 (34.8)	32 (82.1)	-
	Trunk	10 (10.9)	2 (5.1)	8 (15.1)
	Pelvic	2 (2.2)	-	2 (3.8)
Tumour depth [n (%)]	**Deep**	81 (88)	38 (97.4)	43 (81.1)
	Superficial	11 (12)	1 (2.6)	10 (18.9)
Status at excision [n (%)]	**Local**	89 (96.7)	36 (92.3)	53 (100.0)
	Metastatic	2 (2.2)	2 (5.1)	-
	Multifocal	1 (1.1)	1 (2.6)	-
Radiation associated [n (%)]	**No**	87 (94.6)	39 (100.0)	48 (90.6)
	Yes	5 (5.4)	-	5 (9.4)
Tumour margins [n (%)]	**R0**	35 (38)	9 (23.1)	26 (49.1)
	R1	52 (56.5)	25 (64.1)	27 (50.9)
	unknown	5 (5.5)	5 (12.8)	-
Performance status [n (%)]	**0**	39 (42.4)	17 (43.6)	22 (41.5)
	1	27 (29.3)	12 (30.8)	15 (28.3)
	2	6 (6.5)	2 (5.1)	4 (7.5)
	3	4 (4.3)	1 (2.6)	3 (5.7)
	unknown	16 (17.4)	7 (17.9)	9 (17.0)
Pre-op treatment [n (%)]	**CTX**	1 (1)	1 (2.6)	-
	None	91 (99)	38 (97.4)	53 (100.0)

a local recurrence event, 39% had experienced a metastatic event, and 57% were deceased.

There were extensive differences between the clinicopathological features of UPS and DDLPS. By extension, it was hypothesised that histology, as well as other clinicopathological variables may influence patient outcomes. To assess this Kaplan Meier curves were plotted and univariable Cox regressions performed (summarised in **Supplemental Table *5.7***). As in the LMS-specific cohort, some anatomical sites contained low numbers of patients. Therefore, intra-abdominal and retroperitoneal (i.e., intracavity) tumours were grouped, and tumours of all other anatomical sites grouped.

All other variables were handled as before (**section 4.2.2**). Univariable Cox regression revealed histological subtype as associated with a significantly different LRFS and MFS. DDLPS showed a significantly shorter LRFS (HR = 2.63, 95% CI = 1.4 – 4.94, p = 0.003), but significantly longer MFS (HR = 0.426, 95% CI = 0.205 – 0.889, p = 0.023; **Supplemental Figure *5.13*A**). Lower grade tumours (grade 2 v 3) were associated with a superior MFS (HR = 0.284, 95% CI = =0.1 – 0.809, p = 0.018) and OS (HR = 0.435, 95% CI = 0.204 – 0.926, p = 0.031; **Supplemental Figure 5.13B**), tumours of the largest size category were associated with a significantly poorer LRFS (HR = 2.14, 95% CI = 1.08 – 4.22, p = 0.029; **Supplemental Figure 5.13C**), and age was associated with a superior OS (HR = 1.04, 95% CI = 1.02 – 1.07, p = 0.001). All PS categories were associated with a significantly inferior OS compared to the reference group of PS 0 (PS 1: HR = 2.66, 95% CI = 1.35 – 5.24, p = 0.005; PS 2-3: HR = 4.21, 95% CI = 1.76 – 10.1, p = 0.001; PS unknown: HR = 2.7, 95% CI = 1.21 – 6.02, p = 0.015; **Supplemental Figure *5.13*D**).

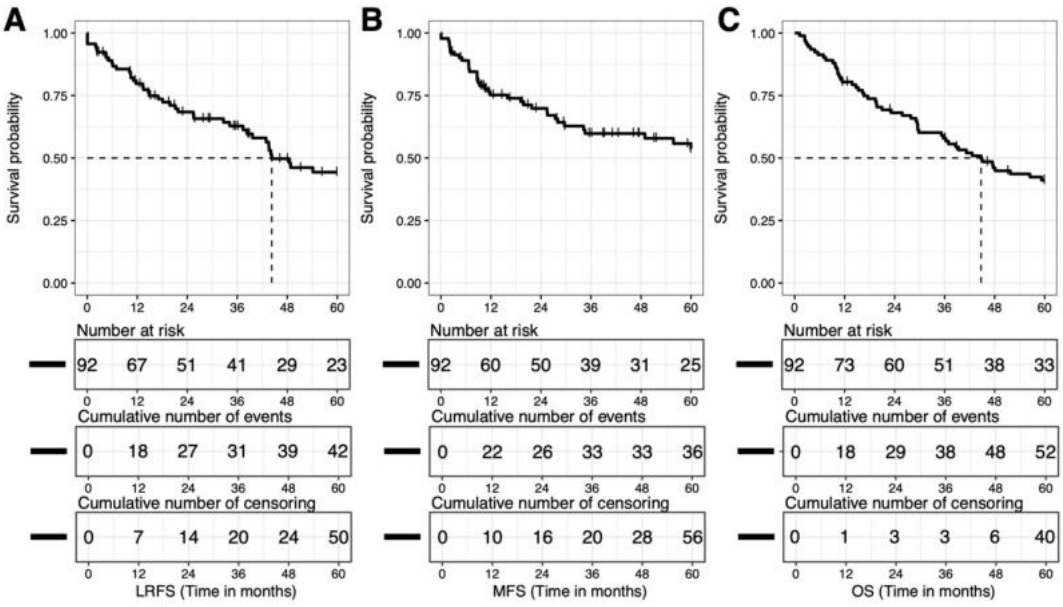


Figure 5.10 Clinical outcome of the dedifferentiated liposarcoma (DDLPS) and undifferentiated pleomorphic sarcoma (UPS) cohort. Kaplan Meier plots showing local recurrence free survival (LRFS; **A**), metastasis free survival (MFS; **B**), and overall survival (OS; **C**) up to 5-years post-surgery. Dashed line indicates median survival.

Following multivariable adjustment (summarised in **Supplemental Table *5.8***), age and a PS of 1 were the only variables to retain significance as prognosticators (both for OS). Whilst anatomical site and tumour size gained significance in the LRFS and MFS models respectively, suggesting the presence of suppressor variables. Both anatomical site and tumour size both showed extensive interactions with other clinicopathological variables

in this cohort (**section 5.2.2.1**). Namely, depth with size, age with anatomical site, histology and grade with both size and anatomical site, and anatomical site and size with each other. It is probable that these interactions underlie the differences seen between the univariable and multivariable model. Given suppressor variables show weak influence over the IV (explained in **section 5.2.1.2**), it is unlikely that histology (in LRFS and MFS) or grade (in MFS) are suppressive as both were found significant in univariable assessment. Therefore, it is hypothesised that depth or age drive the significance of tumour size and anatomical site.

For all univariable and multivariable models, the PH assumption was met. Minor violations were observed for the PS variable in the univariable MFS model ($p = 0.021$), univariable OS model ($p = 0.01$), multivariable MFS model ($p = 0.048$), and multivariable OS model ($p = 0.013$; **Supplemental Figure *5.14*** and **Supplemental Figure *5.15***). However, these did not invalidate the use of the Cox model.

5.2.2.3 Heterogeneity in TIL burden in UPS and DDLPS

Given the previously reported immune heterogeneity in UPS and DDLPS that was also shown herein (**Chapter 4, Figure 4.10**), clinical trial results illustrating favourable responses to ICB in a subset of UPS and DDLPS patients, and the reported association between TIL levels and immune checkpoint expression levels across cancer types, the immune infiltrate of these two subtypes in our cohort was characterised[139,140,611,612]. IHC data on TMAs generated and collected by previous lab members and corresponding to CD3, CD4 and CD8 expression was available for a subset of the MS-profiled UPS and DDLPS cohort. Data was collected from 5 TMAs containing multiple 1mm cores from each sample. As in LMS analyses (**section 5.2.1.5**), 2 TMA cores were required for analysis to ensure the robustness of any findings. This resulted in a CD3+ TIL (total T cell) dataset of 50 UPS and 32 DDLPS samples (total = 82); a CD4+ TIL (helper T cell) dataset of 50 UPS and 35 DDLPS samples (total = 85); and a CD8+ TIL (cytotoxic T cell) dataset of 47 UPS and 32 DDLPS samples (total = 79; **Supplemental Figure *5.16*A**). Scores were adjusted to TIL/mm^2 (**section 2.6)**, and the mean of all scores used as the final sample measure. The suitability of using mean as a summary statistic was assessed as in LMS (**section 5.2.1.5**). Briefly, inter-core variability was largest in samples with higher immune infiltrate (**Supplemental Figure *5.16*B**). Therefore, one caveat of this approach is that the mean TIL measures may not accurately portray TIL burden. In LMS, it has been shown that where data is only available for a small number of cores, dichotomisation at the median can accurately classify most tumours as either 'high' or 'low'[606]. Data for each case was therefore dichotomised at each median TIL level.

Across the UPS and DDLPS cohort, CD3+ TILs, CD4+ TILs, and CD8+ TILs were similarly distributed. Density plots of each TIL population showed left tailing suggesting a relatively low infiltrate of CD3/4/8+ cells in most samples (**Figure *5.11***). However, CD3+ TILs did show less extreme tailing than CD4+ and CD8+ TILs, alluding to a larger population with intermediate CD3+ TIL levels. CD3+ and CD4+ TILs were generally higher than CD8+ TILs, with CD3+ ranging from 1 - 1238 TIL/mm^2 (median = 107 TIL/mm^2), CD4+ from 1 - 1735 TIL/mm^2 (median = 89 TIL/mm^2), and CD8+ from 0 - 869 TIL/mm^2 (median = 31 TIL/mm^2). All TIL measures were positively correlated but decreased in strength from CD3+ and CD4+ (Pearson correlation coefficient = 0.97; $p < 0.001$), to CD3+ and CD8+ (Pearson correlation coefficient = 0.75; $p < 0.001$), to CD4+ and CD8+ (Pearson correlation coefficient = 0.68; $p < 0.001$; **Supplemental Figure *5.16*C**). As in LMS (**section 5.2.1.5**), despite CD4+ TILs being a theoretical subpopulation of CD3+ TILs, in some cases the mean CD4+ cell infiltrate was higher. It was not possible to identify the cause of this and therefore CD4+ TIL results require cautious interpretation.

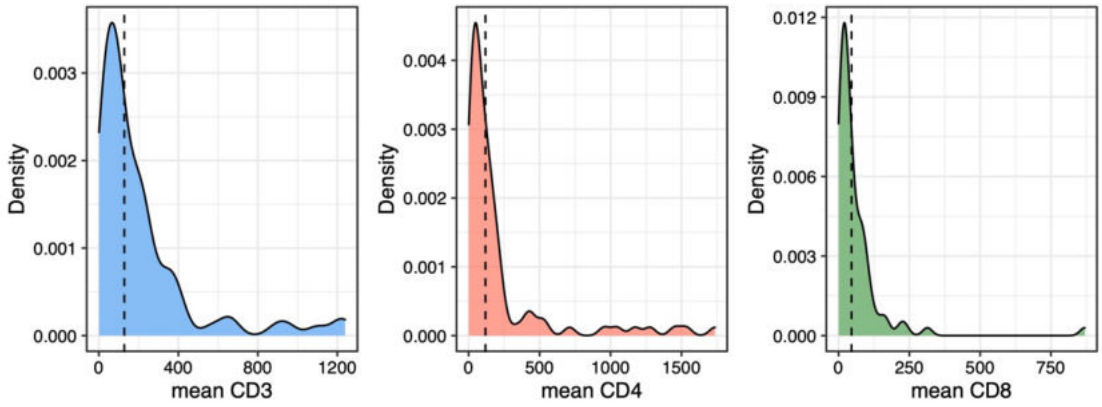


Figure 5.11 CD3+/4+/8+ tumour infiltrating lymphocyte (TIL) burden in dedifferentiated liposarcoma (DDLPS) and undifferentiated pleomorphic sarcoma (UPS)
Density plots showing the distribution of CD3+/4+/8+ TILs in DDLPS and UPS cases. Dashed line indicates median.

5.2.2.4 Clinical characterisation of UPS and DDLPS immune subtypes

As a result of this cohort comprising 2 STS subtypes, significant variation in clinicopathological features was introduced (**section 5.2.2.1**). Similarly, these subtypes or any other clinicopathological variable group may show intrinsically different TIL burdens. It was therefore important to assess whether TIL burden correlated with clinicopathological features. Moreover, previous studies have highlighted an association

between immune activity and clinical outcome, thus the relationship between TIL burden and LRFS, MFS, and OS was also explored[36,222,231].

There was no statistical association between CD3+ TIL burden and any clinicopathological variable (**Figure *5.12*A** and **Supplemental Table *5.9***). However, CD3+ TIL burden was associated with outcome. Use of the Kaplan Meier curve (**Figure *5.12*B**) and univariable Cox regression (**Table *5.5***), revealed high CD3+ TILs to be significantly associated with a superior LRFS and OS (HR = 0.489, 95% CI = 0.247 – 0.969, p = 0.04, and HR = 0.43, 95% CI = 0.241 – 0.767, p = 0.004 respectively). Following multivariable adjustment, a high CD3+ TIL level retained independent significance in the OS model; associated with a superior OS compared to low CD3+ TILs

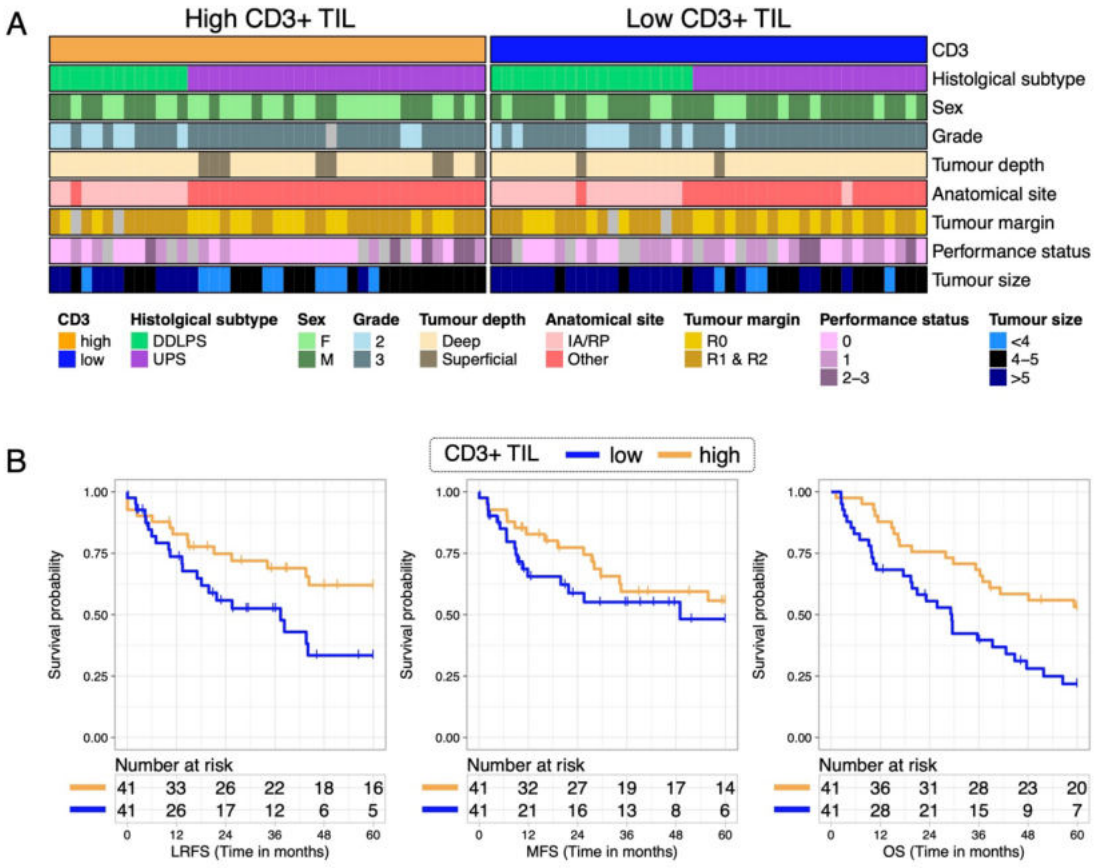


Figure 5.12 Clinical features of high and low CD3+ tumour infiltrating lymphocyte (TIL) cases
(A) Overview clinicopathological features of high and low CD3+ TIL cases. High and low determined by median value. Corresponding statistical tests for associations between variables are detailed in **Supplemental Table 5.9**. **(B)** Kaplan-Meier plots of local recurrence free survival (LRFS) metastasis free survival (MFS), and overall survival (OS) up to 5-years post-surgery for high CD3+ TIL patients compared to low CD3+ TIL patients. Corresponding univariable Cox regression results are detailed in **Table 5.5**.

(HR = 0.484, 95% CI = 0.236 – 0.992, p = 0.048; **Table *5.6***). The only other variable significantly associated with OS following multivariable adjustment was age. As with CD3+ TILs, CD4+ TILs also showed no association with any clinicopathological variable (**Figure *5.13*A** and **Supplemental Table *5.9***). High CD4+ TILs were significantly associated with a superior LRFS and OS in the univariable setting (HR = 0.499, 95% CI = 0.258 – 0.967, p = 0.04, and HR = 0.532, 95% CI = 0.303 – 0.936, p = 0.029 respectively; **Figure *5.13*B** and **Table *5.5***). Following multivariable adjustment, the significance of high/low CD4+ TILs was lost (**Supplemental Table *5.10***). As with CD3+ and CD4+ TILs, CD8+ TILs also showed no associated with any clinicopathological variable (**Figure *5.14*A** and **Supplemental Table *5.9***). High CD8+ TILs were significantly

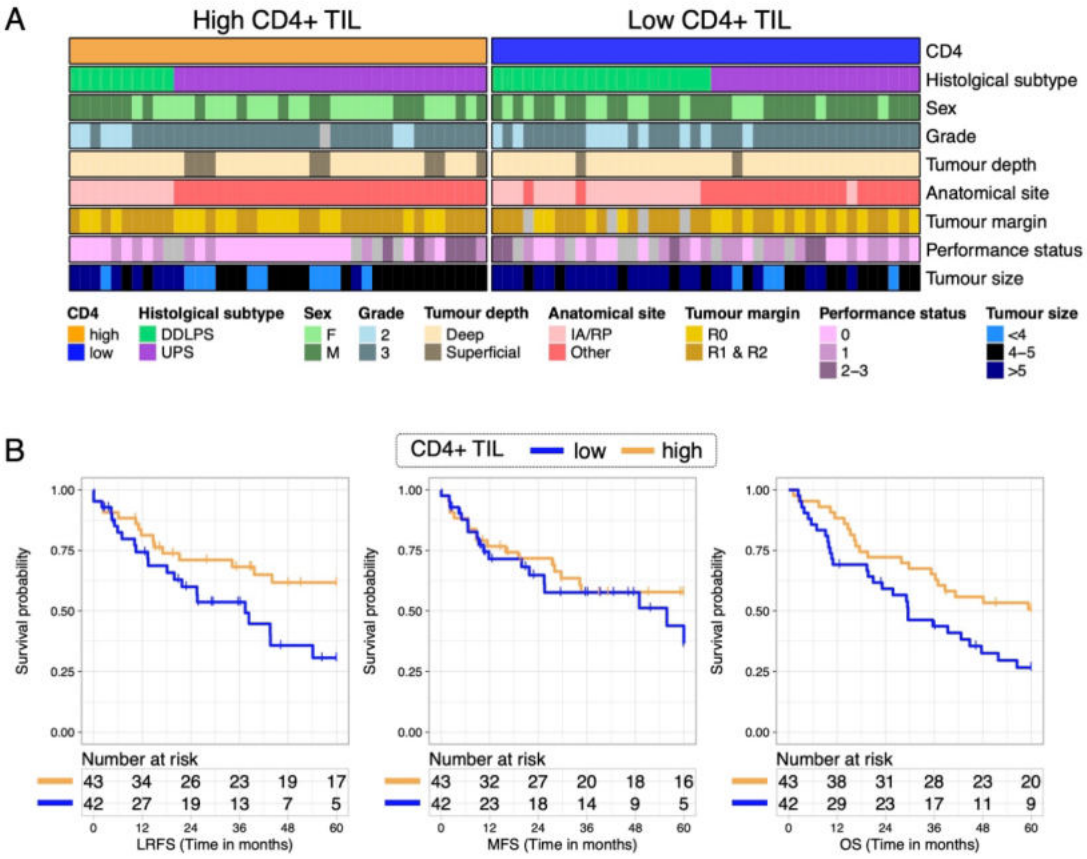


Figure 5.13 Clinical features of high and low CD4+ tumour infiltrating lymphocyte (TIL) cases
(A) Overview clinicopathological features of high and low CD4+ TIL cases. High and low determined by median value. Corresponding statistical tests for associations between variables are detailed in **Supplemental Table 5.9**. **(B)** Kaplan-Meier plots of local recurrence free survival (LRFS) metastasis free survival (MFS), and overall survival (OS) up to 5-years post-surgery for high CD4+ TIL patients compared to low CD4+ TIL patients. Corresponding univariable Cox regression results are detailed in **Table 5.5**.

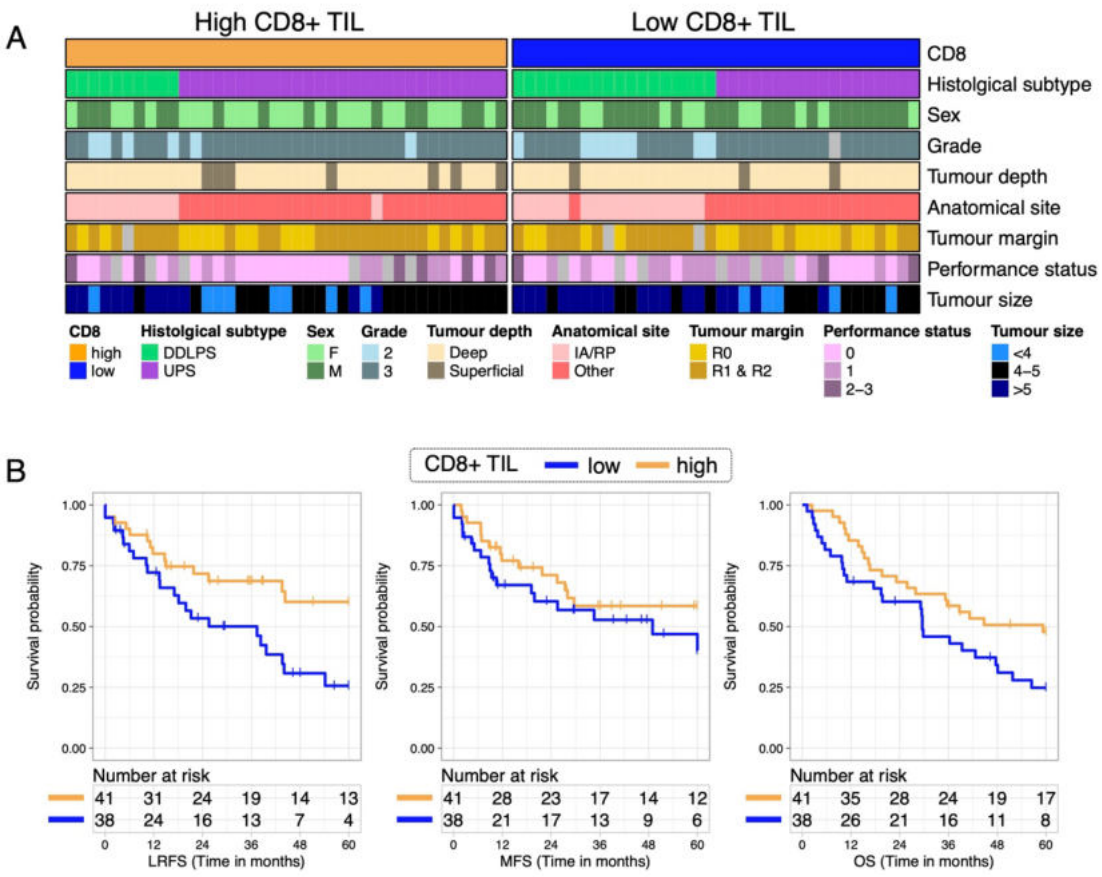


Figure 5.14 Clinical features of high and low CD8+ tumour infiltrating lymphocyte (TIL) cases
(A) Overview clinicopathological features of high and low CD8+ TIL cases. High and low determined by median value. Corresponding statistical tests for associations between variables are detailed in **Supplemental Table 5.9**. **(B)** Kaplan-Meier plots of local recurrence free survival (LRFS) metastasis free survival (MFS), and overall survival (OS) up to 5-years post-surgery for high CD8+ TIL patients compared to low CD8+ TIL patients. Corresponding univariable Cox regression results are detailed in **Table 5.5**.

associated with a superior LRFS (HR = 0.452, 95% CI = 0.232 – 0.881, p = 0.02; **Figure *5.14*B** and **Table *5.5***). As was the case for CD4+ TILs, multivariable Cox models showed no significant relationship between high/low CD8+ TIL burden and outcome (**Supplemental Table *5.11***).

As before, assumptions of the Cox regression model were assessed. This revealed minor PH violations in the CD4+ OS model for sex, and in the CD8+ MFS model for subtype, anatomical site, and PS. Inspection of the scaled Schoenfeld's residual plots (**Supplemental Figure *5.17***), illustrated no obvious violations that would invalidate the model. However, PH violations (Schoenfeld's $p < 0.01$) were observed in all OS models for PS (**Supplemental Figure *5.18***). Given the strong relationship between PS and outcome, not only identified in the UPS and DDLPS cohort (**section 5.2.2.2**), but also in

Table 5.5 Univariable Cox regression assessing CD3+/CD4+/CD8+ tumour infiltrating lymphocyte (TIL) burden in dedifferentiated liposarcoma and undifferentiated pleomorphic sarcoma cases
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval.

		LRFS		MFS		OS	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
CD3	*low (ref)*	-	-	-	-	-	-
	high	**0.489 (0.247-0.969)**	**0.04**	0.706 (0.355-1.4)	0.32	**0.43 (0.241-0.767)**	**0.004**
CD4	*low (ref)*	-	-	-	-	-	-
	high	**0.499 (0.258-0.967)**	**0.04**	0.737 (0.378-1.44)	0.37	**0.532 (0.303-0.936)**	**0.029**
CD8	*low (ref)*	-	-	-	-	-	-
	high	**0.452 (0.232-0.881)**	**0.02**	0.66 (0.332-1.31)	0.235	0.568 (0.321-1.01)	0.053

Table 5.6 Multivariable Cox regression assessing CD3+ tumour infiltrating lymphocyte (TIL) burden in dedifferentiated liposarcoma (DDLPS) and undifferentiated pleomorphic sarcoma (UPS) patients.
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. Anatomical site of 'Other' indicates extremity, trunk wall, and head/neck cases. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval; IA = Intra-abdominal; RP = retroperitoneal

		LRFS		MFS		OS	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
Age at excision (years)		1.04 (0.995-1.08)	0.085	1.02 (0.979-1.07)	0.317	**1.04 (1.01-1.09)**	**0.027**
Sex	*M (ref)*	-	-	-	-	-	-
	F	1.23 (0.536-2.83)	0.625	1.4 (0.554-3.56)	0.475	1.32 (0.668-2.62)	0.423
Histological subtype	*UPS (ref)*	-	-	-	-	-	-
	DDLPS	0.937 (0.192-4.58)	0.936	0.53 (0.094-2.99)	0.472	0.559 (0.119-2.62)	0.461
Anatomical site	*Other (ref)*	-	-	-	-	-	-
	IA/RP	5.47 (0.935-32)	0.059	0.216 (0.036-1.32)	0.097	1.59 (0.316-8.01)	0.574
FNCLCC grade	*3 (ref)*	-	-	-	-	-	-
	2	0.91 (0.348-2.38)	0.848	0.388 (0.105-1.43)	0.155	0.618 (0.238-1.6)	0.323
Performance status	*0 (ref)*	-	-	-	-	-	-
	1	1.8 (0.665-4.87)	0.247	1.6 (0.611-4.21)	0.338	2.16 (0.914-5.09)	0.079
	2-3	1.11 (0.247-5.02)	0.888	0.383 (0.07-2.09)	0.267	1.98 (0.58-6.76)	0.275
	unknown	1.18 (0.353-3.93)	0.79	0.878 (0.208-3.71)	0.86	1.92 (0.62-5.95)	0.258
Tumour depth	*Deep (ref)*	-	-	-	-	-	-
	Superficial	0.57 (0.104-3.12)	0.517	0.401 (0.097-1.66)	0.207	0.807 (0.231-2.82)	0.738
Tumour margin	*R1 & R2 (ref)*	-	-	-	-	-	-
	R0	0.767 (0.305-1.93)	0.572	0.994 (0.43-2.3)	0.989	0.911 (0.421-1.97)	0.813
	unknown	1.58 (0.284-8.78)	0.601	1.62 (0.182-14.4)	0.667	0.774 (0.079-7.55)	0.825
Log(Tumour size [mm])	4-5 (ref)	-	-	-	-	-	-
	< 4	0.796 (0.204-3.1)	0.743	0.391 (0.118-1.29)	0.124	0.325 (0.103-1.02)	0.054
	> 5	0.855 (0.282-2.59)	0.781	**5.31 (1.52-18.6)**	**0.009**	1.75 (0.718-4.26)	0.219
CD3	*low (ref)*	-	-	-	-	-	-
	high	0.517 (0.228-1.17)	0.114	0.872 (0.356-2.14)	0.764	**0.484 (0.236-0.992)**	**0.048**

the full cohort (**section 4.2.2**), it was deemed inappropriate to exclude the PS variable from the model. Instead, PS was included but interpretation cautioned. Importantly, in the CD4+ and CD8+ OS models inclusive of PS, the global Schoenfeld test identified no PH violation ($p = 0.153$ and $p = 0.065$ respectively). In the CD3+ OS model only a minor global PH violation (global Schoenfeld $p = 0.03$) was observed. Therefore, as complete models, interpretation is valid.

5.2.2.5 Biological features associated with CD3+ TILs in UPS and DDLPS

A significant association between CD3+ TILs and OS was revealed herein. To better understand the biological basis that may underpin differing outcomes in these patients, the wider immune biology in high and low CD3+ TIL groups was investigated. Targeted transcriptomic data (NanoString) corresponding to 21 immune components (detailed in **section 2.4)** was collected by previous lab members and available for analysis. NanoString data was present for 41 UPS and 26 DDLPS cases (total = 67) with both MS and IHC CD3 data. Gene expression profiles were compared between the low and high CD3+ TIL cases using Kruskal-Wallis tests (**Figure 5.15A**). This highlighted several genes, including *CD3G* and *CD8G* as expressed at a significantly higher levels in the high CD3+ TIL group. Several immune checkpoint regulation genes were also highlighted as enriched in the high CD3+ TIL group. These included *PDCD1*, the PD-1 receptor, *PDCD1LG2*, the PD-L2 ligand, as well as checkpoint genes *IDO* and *LAG3*. Following multiple testing adjustment, only *PDCD1* remained significant. Together, these results suggest high CD3+ TIL burden to be associated with increased activity in immune checkpoint processes. Conversely, low CD3+ TIL patients harboured low expression of immune checkpoint genes. As immune checkpoint genes are suggested to be associated with ICB response (as discussed in **section 1.5.3.1**), ICB intervention may not be of benefit to the low CD3+ TIL population.

Given the likely ineffectiveness of ICB in low CD3+ TIL patients and the significantly poorer OS seen in low CD3+ TIL compared to high CD3+ TIL cases (**section 5.2.2.4**); there is an evident and pressing need to identify clinically actionable biological pathways in these patients. Targeted gene expression analyses of selected immune components failed to highlight any upregulated genes within the low CD3+ TIL population (**Figure 5.15A**). Therefore, to reveal such pathways, the more comprehensive proteomic data was interrogated. GSEA was performed on the complete dataset against the Hallmark and GO BP databases of MSigDB. Strikingly, the top 20 significant (based on NES; adjusted $p < 0.05$) gene sets enriched in both high and low CD3+ TIL cases were exclusively immune-related. Furthermore, within the top 40 gene sets enriched in high

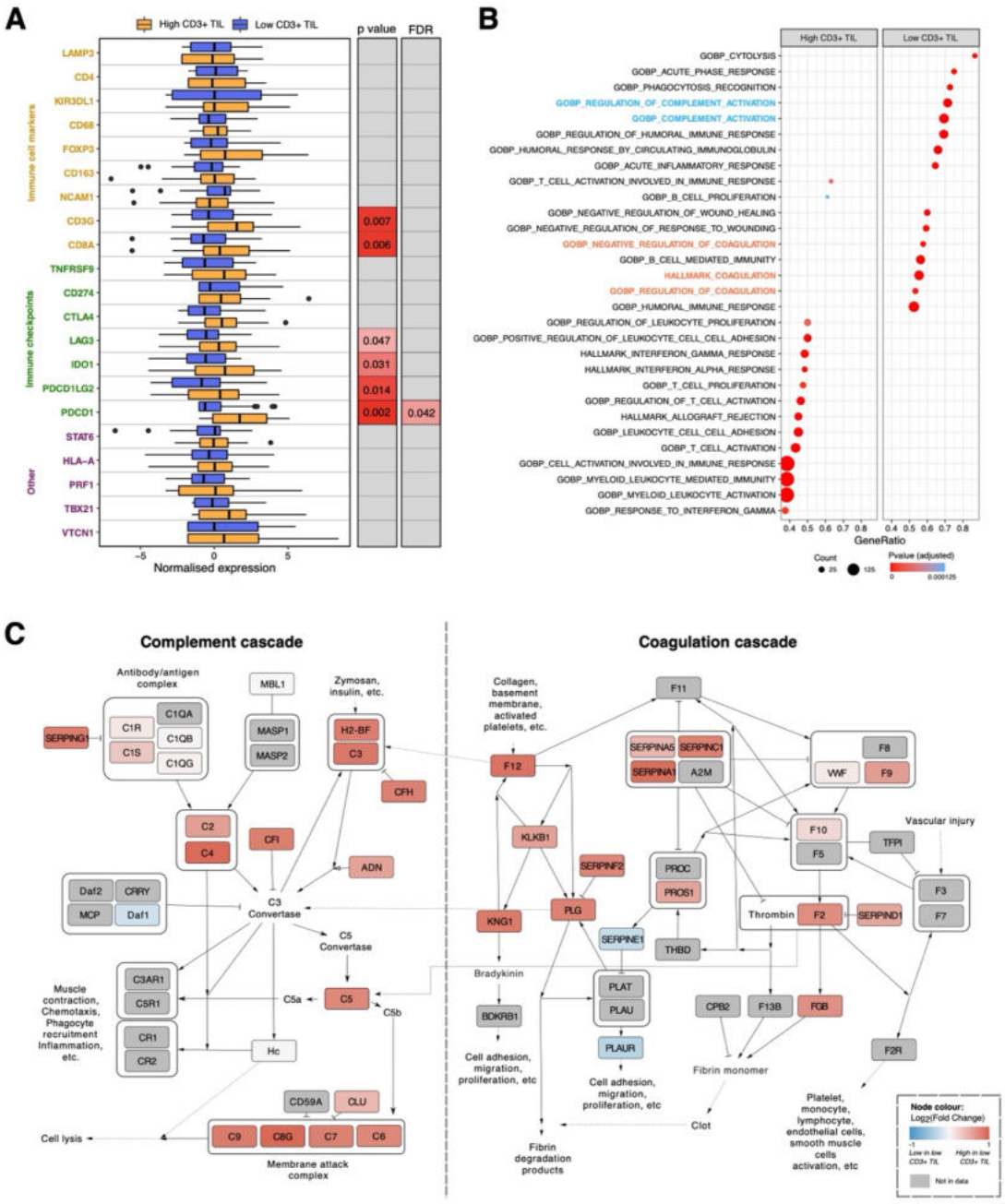


Figure 5.15 Characterisation of the immune profiles of dedifferentiated liposarcoma (DDLPS) and undifferentiated pleomorphic sarcoma (UPS)

(A) Boxplots comparing expression of 21 immune-related genes in low and high CD3+ TIL cases. Boxes indicate 25th and 75th percentile, with median line in the middle, whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. p values determined by Kruskal-Wallis tests and adjusted to false discovery rate (FDR). **(B)** Gene Set Enrichment Analysis (GSEA) results applied to the proteomic dataset showing the top 15 gene sets enriched in CD3+ TIL-high and and-low cases based on normalised enrichment score (NES) with gene sets related to complement activity (blue) and coagulation processes (orange) highlighted. **(C)** Protein-protein interaction network of the coagulation and complement cascades). Node colour indicates Log_2(Fold Change CD3+ TIL low: CD3+ TIL high) protein expression. Grey indicates nodes that are not in the proteomic data. Abbreviations: TIL = tumour infiltrating lymphocyte

CD3+ TILs and the top 40 gene sets enriched in low CD3+ TILs, only 6 and 15 described non-immune processes respectively. In agreement with IHC and transcriptomic data, GSEA revealed a robust enrichment of T cell immune responses in the high CD3+ TIL group (**Figure 5.15B**). The enriched gene sets included leukocyte and T cell specific activation and proliferation processes. Additionally, interferon α and γ responses were also enriched. In cancer, interferons α and γ are cytokines which show complex and reciprocal interplay with T cells: T cells can secrete interferons and interferons support T cell differentiation, priming, and activation[613–616]. By contrast, low CD3+ TILs showed an enrichment of the humoral immune response, of which key players include complement, antibodies, plasma cells, and B cells[617]. Additionally, the coagulation pathway, which is known to interact with the complement cascade was enriched[618]. To inspect the proteins contributing to the enrichment of complement and coagulation in these tumours, PPI networks were constructed based on the KEGG and WikiPathways databases (**Figure 5.15C**). This analysis highlighted the serpin family of serine proteases to be strongly upregulated in low CD3+ TIL patients (SERPINA1/A5/C1/D1/F2/G1). Several complement proteins were also upregulated in low CD3+ TIL patients, including those of the membrane attack complex (MAC). Therefore, despite low CD3+ TIL patients showing a low TIL infiltration and low immune checkpoint activity, they are not 'immune cold'. Instead, these patients harbour a distinctive and active immune component, which may have implications for disease progression, patient outcome, and response to treatment.

5.3 Discussion and summary

This chapter investigated the intra-subtype heterogeneity of LMS and the immune-specific features of DDLPS and UPS. This heterogeneity was molecularly defined using multiple complementary datasets (MS, IHC, NanoString). In doing so, this chapter revealed clinical applications corresponding to molecular features of STS and identified multiple areas of interest for future research.

The cohorts analysed describe LMS patients, and DDLPS and UPS patients. Limitations of these cohort designs due to the use of primary tissue only and an absence of normal tissue were discussion in **section 4.3.** Cohort features were largely in line with clinical presentation of each subtype. For example, most LMS were deep seated and occurred predominately in females, DDLPS tended to be large and retroperitoneal, and UPS were mostly high grade[587,619–621]. The DDLPS and UPS cohort showed extensive interactions between clinicopathological variables. This is hypothesised to be the result of histological subtype differences (e.g., the somewhat counterintuitive observation that larger tumours

were typically lower grade, can be explained by DDLPS tumours being large and lower grade). Therefore, despite reports suggesting similar immune profiles, there are limited, if any, clinicopathological similarities between DDLPS and UPS. Furthermore, expected clinicopathological variables such as anatomical site and grade were associated with clinical outcome[45,46]. One LMS-specific limitation was an underrepresentation of uLMS tumours relative to incidence (9% of the LMS cohort vs 25% of all LMS diagnoses)[303]. This may limit the application of our findings in the uLMS group. Beyond this and given the inclusion criteria (detailed in **section 3.2.1**), these cohorts were deemed as representative of the disease population.

5.3.1 Molecular heterogeneity in LMS

This chapter identified 3 robust proteomic subtypes of LMS with different clinical outcomes, and distinct biological features. Some of these features are in agreement with the previously reported transcriptomic subtypes of LMS[36,43,274,281–283].

Specifically, network analysis of the 'immune cold LMS/P1' proteome revealed an upregulation of pro-proliferative complexes involved with DNA repair (RPA and MCM). This has not been previously reported in the transcriptomic subtyping of LMS. 'immune cold LMS/P1' were also biologically characterised by a low expression of pro inflammatory immune hallmarks, including IL2-STAT5 signalling, complement and allograft rejection. IL2 is an inflammatory cytokine that triggers STAT5-mediated transcriptional activity, the targets of which include immune genes[622–624]. Complement is a key component of the innate immune system and a regulator of inflammation, and inflammation has been reported as a trigger for allograft rejection[625–627]. These indicate that 'immune cold LMS/P1' tumours exhibit a markedly low immune response in-situ. The identification of an immune cold proteomic subtype is consistent with transcriptomic reports. Namely, Abeshouse *et al*, Chudasama *et al*, Hemming *et al*, and Anderson *et al* highlight variations in immune infiltrate across subtypes[36,43,274,283]. These studies focus on a singular immune hot subtype, the implication being that other subtypes are immune cold. The transcriptomic studies highlight immune activity through transcriptomic data assessment, both by overrepresentation analysis and in some cases immune cell deconvolution. The efforts herein are the first to complement immune-based molecular profiling findings of LMS with IHC. Using IHC, the low immune activity in 'immune cold LMS/P1' was validated, and specifically revealed a significantly lower CD3+ and CD4+ TIL burden in these tumours compared to 'classical LMS/P2' and 'dedifferentiated LMS/P3'. It is notable that both IL2-STAT5 signalling and CD4+ TILs were significantly downregulated in 'immune cold LMS/P1', as the CD4+ TIL transcriptional program sits

downstream of STAT5[624]. CD4+ TILs primarily function to activate CD8+ TILs, which elicit cytotoxic effects[628,629]. Previous studies have highlighted the presence of a subset of CD4+ TILs (follicular T helper cells) as correlated with better patient outcomes[630]. Additionally, high CD4+ TIL burden has been hypothesised as predictive of ICB response, and STS subtypes with high immune activity (i.e., 'immune hot' UPS and DDLPS) show the most favourable ICB responses[140,631,632]. In ICB basket trials, LMS have shown limited responses, and in a uLMS specific phase II trial evaluating nivolumab, no patients showed treatment response (n = 12)[139,633]. The work herein leads to the hypothesis that further efforts to assess ICB across LMS may be beneficial, and that non 'immune cold LMS/P1' patients would be the most promising candidates. The major restriction of this hypothesis is that the immune differences herein are relative to LMS tumours only. It is widely accepted that as a group of diseases STS show lower immune activity compared to other malignancies (as discussed in **section 1.3.1.2**). Furthermore, within STS, LMS are not highlighted as a typical immune hot histology. Indeed, IHC analysis of DDLPS and UPS within the latter analyses of this chapter showed higher CD3+/4+/8+ levels in these subtypes than LMS. Therefore, the immune hot LMS population may not harbour sufficient immune activity to warrant immunotherapy intervention. Future research directions of interest include the assessment of LMS immune profiles in the context of other cancers. Data from this project could facilitate comparisons within STS. Whilst comprehensive pan-cancer assessments linked to this cohort would require MS and/or IHC analysis of non-STS samples.

The 'dedifferentiated LMS/P3' proteome showed a specific enrichment of numerous ribosomal proteins. In line with this observation, one transcriptomic study has reported an LMS subtype enriched in ribosomal gene expression[281]. Increased ribosomal expression implies increased ribosomal activity (i.e., protein synthesis) within these tumours. Aberrant protein synthesis can impact the fidelity of translation and drastically alter cell behaviour, which may contribute to tumourigenesis[634]. 'Dedifferentiated LMS/P3' also showed markedly lower expression of smooth muscle markers than 'immune cold LMS/P1' and 'classical LMS/P2', and a low expression of the broader 'myogenesis' hallmark. Mechanistically, this suggests a reduction, loss, or absence of smooth muscle lineage signatures, and as such indicates 'dedifferentiated LMS/P3' to harbour a dedifferentiated phenotype. Inspection of the proteomic dataset revealed 79 'myogenesis' proteins to be present, the vast majority of which correspond to myosin chains, integrin subunits and ECM components. Notably, the hallmark 'spermatogenesis' was also significantly downregulated in 'dedifferentiated LMS/P3'. The underlying

biology of this is unclear. Solid tumours have been reported to express germ cell (GC) specific genes (or 'cancer testis (CT)' antigens), markers usually only observed in reproductive development; the expression of which may be driving this observation in LMS[635,636]. Inspection of the overlap in genes/proteins between revealed 16 spermatogenesis components in the proteomic data (ACE, AGFG1, CDK1, CSNK2A2, GSTM3, HSPA2, HSPA4L, IDE, LDHC, PEBP1, PGK2, PRKAR2A, RFC4, TALDO1, TSN, VDAC3). In support of the loss of a smooth muscle signature in a subset of LMS, dimension reduction revealed several LMS profiles, most of which were 'dedifferentiated LMS/P3', to cluster away from the LMS-specific cluster. Although we hypothesised that these 'dedifferentiated LMS/P3' may show similar profiles to the dedifferentiated subtype UPS, clustering did not reflect this. Despite this, the revelation of a dedifferentiated subtype is consistent with the transcriptomic LMS subtype studies, all of which highlight an LMS subtype with a more dedifferentiated phenotype[36,43,274,281–283].

Despite observed biological similarities between the proteomic and transcriptomic subtypes of LMS, clinical associations were not consistent. Unlike claims made for the transcriptomic LMS subtypes, no proteomic subtype was enriched in uLMS tumours[36,43,274,281–283]. The reason for this discordance is unknown. It may be the case that the uLMS features driving transcriptomic observations are not detectable at the proteomic level. Alternatively, compositional differences in the analysed cohorts may limit comparisons. Indeed, it is notable that the cohort herein had poor representation of uLMS cases. In addition, whilst no associations with clinical outcome have been robustly identified with the transcriptomic LMS subtypes, proteomic subtypes were associated with outcome. Specifically, following multivariable adjustment, 'dedifferentiated LMS/P3' showed a significant poorer LRFS and MFS compared to the reference group ('classical LMS/P2'). Model comparisons with and without proteomic subtype suggested proteomic data can provide significant added value to LRFS prognostication. These associations with outcome are consistent with carcinoma literature, which report dedifferentiation to confer a more aggressive malignancy[297–299]. Furthermore, in an LMS specific study which assessed smooth muscle marker expression by IHC, loss of myogenic differentiation markers was shown to be prognostic for a poorer OS[542]. Whilst this may also be the case here, many biological processes differ between the LMS proteomic subtypes, and the exact molecular driver(s) underpinning the variation in clinical outcome are unclear. All observations describe associations and are not necessarily causative. Yet these findings could have important clinical implications for LMS patients. In LMS, disease recurrence and metastasis are common events and the latter the cause of patient death[607,608]. The identification of a high risk subpopulation can stratify patients for further and more

aggressive treatment such as adjuvant regimens, as well as prolonged and more frequent monitoring.

There is no publicly available, paired transcriptomic and MS data, and therefore to assess the relationship between proteomic and transcriptomic subtypes the TCGA RNAseq data was analysed. Use of the MS-identified proteins to cluster the RNAseq data illustrated an impressive ability of the MS-derived genes to capture transcriptomic subtype heterogeneity. However, this does not indicate the proteomic and transcriptomic subtypes are the same. Therefore, the RNAseq data was clustered using the proteomic subtype-specific DEPs. This showed comparable clustering with approximate separation of the transcriptomic LMS subtypes. Yet the subtype-specific DEPs did not appear to drive the clustering. It is therefore unclear as to whether the proteomic and transcriptomic subtypes are the same, and it was not possible to assign which proteomic subtype corresponded to which transcriptomic subtype. There are several caveats to this analysis. Specifically, comparing the datasets relied on the translation of proteomic findings to transcriptomic data. Protein-RNA correlations are known to be poor[448,609,610]. From the translation of RNA to proteins, extensive processing (e.g, to generate different proteoforms) occurs. Proteins are under regulation by PTMs, which alter activity, and in the case of ubiquitination can target proteins for degradation[451]. Therefore, the final protein levels and activity within a cell can vastly differ from measures of gene expression. This hinders proteomic-to-transcriptomic translation of findings. Additionally, the DEPs utilised were are not optimised for proteomic subtype classification. The DEPs are unlikely to perfectly recapitulate proteomic separation itself, therefore it is unreasonable to expect the same of transcriptomic data. As a result, interpretation of the clustering patterns must consider this as a limit to analysis. Future efforts should include a more in-depth investigation into the relationship between proteomic and transcriptomic subtypes, ideally through matched MS and RNAseq profiling of the same samples.

There are several limitations to my study. It is important to note that these findings, whilst revealed in a relatively large cohort by rare disease standards, are derived from a small dataset from a single institution (RMH). These results are therefore highly overfitted to the cohort herein. To validate proteome-outcome associations, an independent validation cohort would be required. Additionally, this work would benefit from the development of a classifier for LMS proteomic subtypes. If a reduced number of proteins were identified as suitable for classification, these could be translated to IHC measures for low cost and rapid classification of patients. Notably, if the intention was to stratify patients for neoadjuvant therapy, these findings would need to be assessed in biopsy

samples as opposed to the resection samples analysed herein. Furthermore, if the intention was to stratify advanced disease patients, these findings would need to be assessed in metastatic and recurrent disease to determine whether LMS subtypes persist throughout disease progression. It is therefore evident that significant steps are required before clinical translation can be considered.

Other efforts for future research could be focused on the assessment of drug targets for these patients. 'Dedifferentiated LMS/P3' showed poor outcome and therefore a high clinical need for treatment options, and 'immune cold LMS/P1' showed a lack of immune activity reducing the likelihood of immunotherapy utility. Disappointingly, the DSigDB analyses herein yielded little insight into targetable axes. This is likely attributable to a poor representation of protein targets within the MS data. As a result, there remains a pressing need for druggable axes to be identified in LMS. With this in mind, future efforts utilising the MS data could include investigating the GO BP, hallmark, and KEGG features, as assessed in the full cohort (**section 4.2.3**), within the LMS cohort. Starting from a broad signature standpoint may facilitate the identification of groups of drugs (e.g., those targeting metabolic activity), whose target profiles can individually be queried. These drugs could be assessed *in vitro*, where large scale drug screening can be performed. Such co-ordination between *in vitro* experimentation and bioinformatic profiling of tumour specimens may reveal promising candidate therapies.

5.3.2 The immune landscape of DDLPS and UPS

This chapter also characterised immune heterogeneity across UPS and DDLPS. Stratification of the cohort based on CD3+ TIL burden identified 2 subtypes with distinctive immune components and differing OS. These subtypes were independent of histology and all other clinicopathological features that are typically associated with outcome. Moreover, the association of CD3+ TIL burden remained significant following multivariable adjustment of the Cox model. This therefore illustrates the added prognostic value immune cell characterisation can provide. However, the validity of this model was questioned. The PS variable showed a strong association with outcome but did not satisfy the PH assumption. Herein, PS is a measure of functional status following at diagnosis. Although some analyses show PS as associated with long-term outcome (~ years), PS has also been noted to have prognostic value for survival in the short-term (~ months)[45,46,637]. It follows that patients with low functional ability at diagnosis are at higher risk of death, however as time passes and treatment commences and progresses, this risk may decrease. It is therefore unsurprising that PS may not show a constant relationship with OS over time, as is required to meet the PH assumption. In addition,

PS can impact treatment choices, further complicating its association with risk over time: patients with low functional ability often cannot tolerate treatment toxicities well and thus may not receive the aggressive regimens needed to induce remission. PH violation invalidated the univariable Cox assessment of PS, but in the multivariable models, the global PH was still met. Next steps could include optimisation of this model, by use of a PS-time interaction variable, or a time stratified model. This would provide a more statistically robust assessment of outcome.

To better understand the boarder immune context of low/high CD3+ TIL-stratified UPS and DDLPS tumours, targeted transcriptomics was utilised. This revealed the high CD3+ TIL subtype to show concordant high expression of *CD3G* and *CD8G*, as well as several immune checkpoint genes. The observed high *CD3G*, encoding the γ subunit of CD3, is consistent with stratification of these groups based on CD3+ TIL burden. The high *CD8G*, encoding the γ subunit of CD8, in high CD3+ TIL cases may be reflective of the positive correlation observed between CD3+ TILs and CD8+ TILs in these samples. However, no differential transcriptomic expression of *CD4* was observed, despite CD3+ TIL and CD4+ TIL correlation. This may be due to methodological differences: IHC measures protein abundance, whilst NanoString measures gene expression (RNA). Indeed, this would be agreement with many reports that note low concordance between individual RNA and protein levels[448,609,610]. The enrichment of immune checkpoint expression in high CD3+ TIL samples was consistent with current literature noting TIL and checkpoint expression correlation in many cancer types[231,611,612].

In addition to IHC and transcriptomic characterisation, the richness of the MS data was leveraged to identify broad biological activities differentially associated with low and high CD3+ TILs. Notably the most significant biological features were immune associated. This suggests these tumours present with highly comparable biological profiles in all aspects except immune response. Given these groups show significantly different OS, this highlights the importance of the immune environment in disease progression. MS analysis revealed that whilst the high CD3+ TIL subtype showed evidence of cell-mediated immunity, the low CD3+ TIL subtype showed evidence of a humoral immune response (enrichment of B cell activity, complement and phagocytosis). Cell-mediated immunity relies on the activity of T cells[638]. By contrast, the humoral immune response is mediated by antibodies produced by plasma cells which differentiate from B cells[617]. Much focus in tumour immunology is directed towards T cell responses, and less so towards humoral responses. In sarcoma, B cells, players in humoral immunity, have been show as prognostic for improved OS and predictive of favourable responses to ICB

(discussed in **section 1.5.3.1**)[231]. However, this study also showed co-ordinately high T cell infiltrate (CD8+) in high B cell tumours, resultant of the presence of T and B cell-containing TLS. The observations herein in low CD3+ TIL patients contrast this. There is no detectable enrichment in CD8+ TIL levels in these patients, and the opposite impact on OS is observed; low CD3+ TIL patients (i.e., enriched in humoral activity) were associated with a poorer outcome. In addition to B cells and antibodies, the humoral immune response also comprises the complement cascade. The complement cascade amplifies the activities of antibodies and is highly interconnected with coagulation[562,595,596,618,639]. Both complement and coagulation were enriched in CD3+ TIL samples. Within the complement and coagulation cascades, the MAC complex was highlighted. MAC binds to and disrupts the membrane of a target cell inducing cell lysis and death. The role of MAC in cancer is complex and recent reports provide evidence that MAC binding to cancer cells can activate pathways which inhibit cell death signals and promote long term cell survival[640–642]. Additionally, in melanoma and lung cancer activation of complement has been reported to promote tumour growth and, consistent with IHC data herein, suppress CD4+ and CD8+ TILs[643–645].

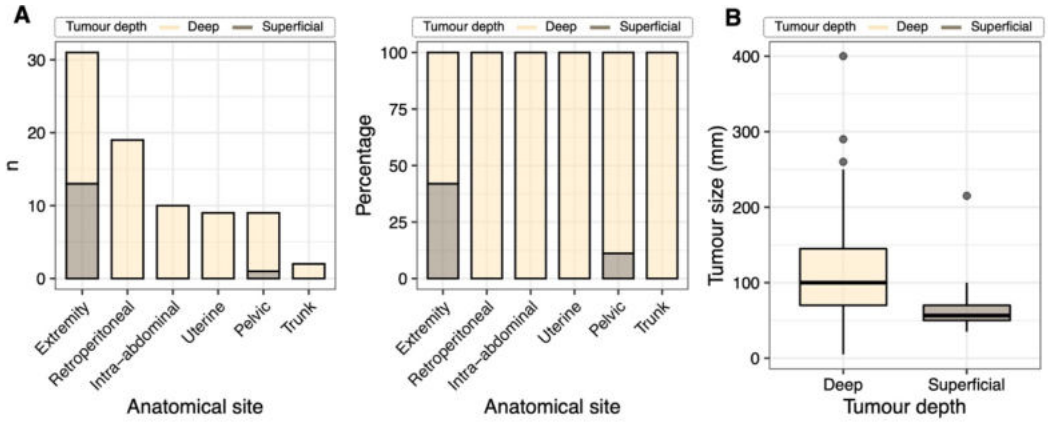
Given the poor outcome of low CD3+ TIL patients, and low checkpoint expression suggesting ICB response will be poor in these patients, there is high a clinical need to identify candidate treatment approaches. Targeting complement components could represent a viable option. Several inhibitors of complement are currently approved, or under investigation for a range of non-oncology uses including treatment of paroxysmal nocturnal haemoglobinuria and rheumatoid arthritis, and coronary artery bypass grafting[646–648]. Clinical trials for these drugs have not been conducted in cancer, although pre-clinical evidence suggests promise. In lung and colon cancer mouse models, co-inhibition of complement and PD1/PD-L1 treatment led to a synergistic antitumour immune response[649,650].

Future avenues to investigate in DDLPS and UPS immunity include assessment of the other immune cell subsets and other TILs analysed in this project (CD4+ and CD8+). Herein these were not followed up due to a lack of association with clinical outcome. However, Kaplan Meier curves did show trends between CD4+ and CD8+ and outcome, and statistical assessment was based on an arbitrary cut point (the median). The median is highly unlikely to be the optimal point for identifying high and low TIL burden patients with clinical relevance. Furthermore, dichotomising results in groups where the highest low TIL sample and the lowest high TIL samples are highly similar. Therefore, future options for analysis include optimising a cut point based on outcome; various methods

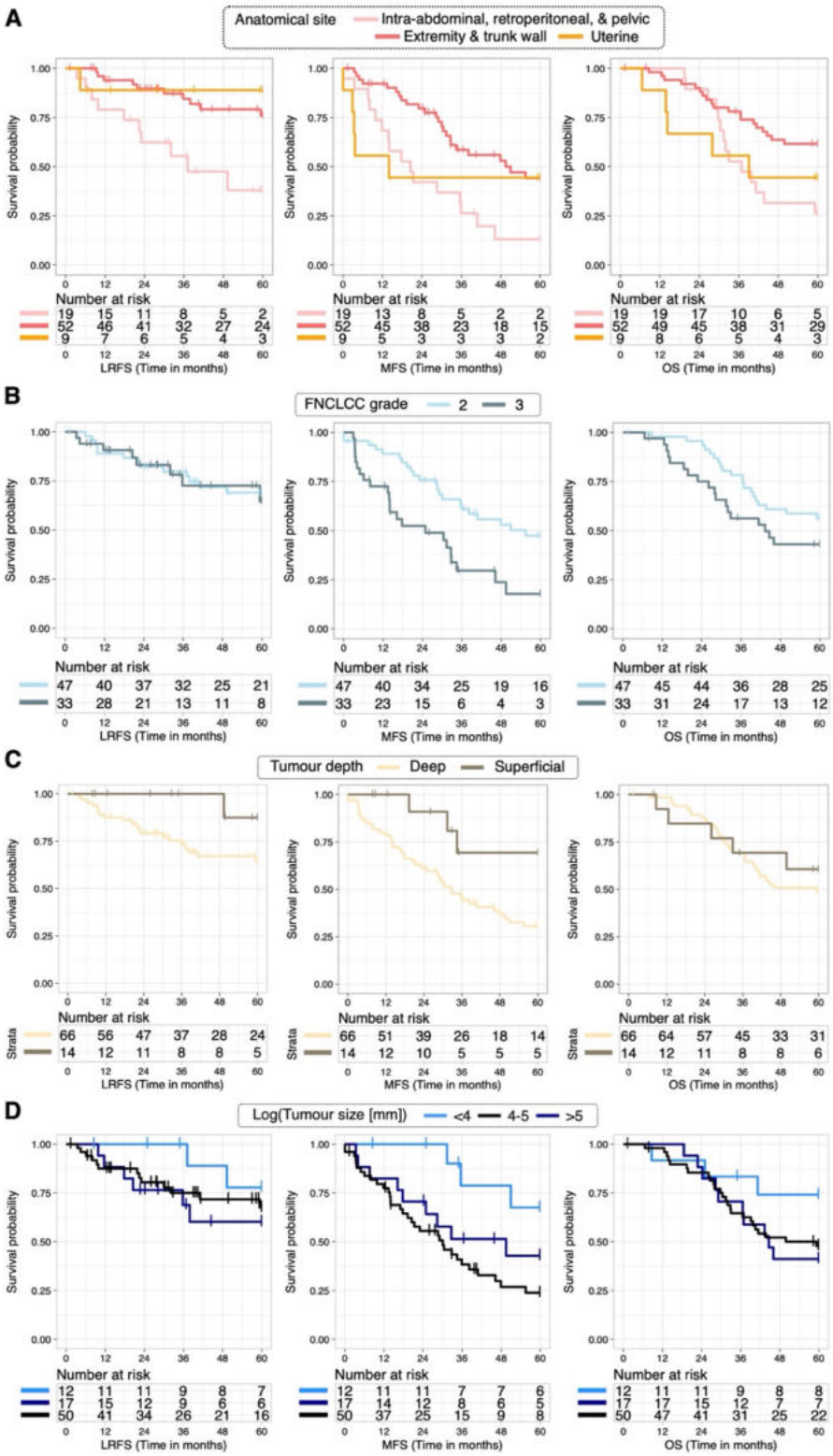
are available for this purpose[651–654]. Additionally, as with the LMS analyses, it would be desirable to expand the cohort to include samples from independent research/clinical institutions to ensure results are not overfitted to this RMH cohort. Expansion efforts should also include metastasis/recurrent samples to assess whether findings are applicable in the advanced disease setting.

5.4 Supplemental material

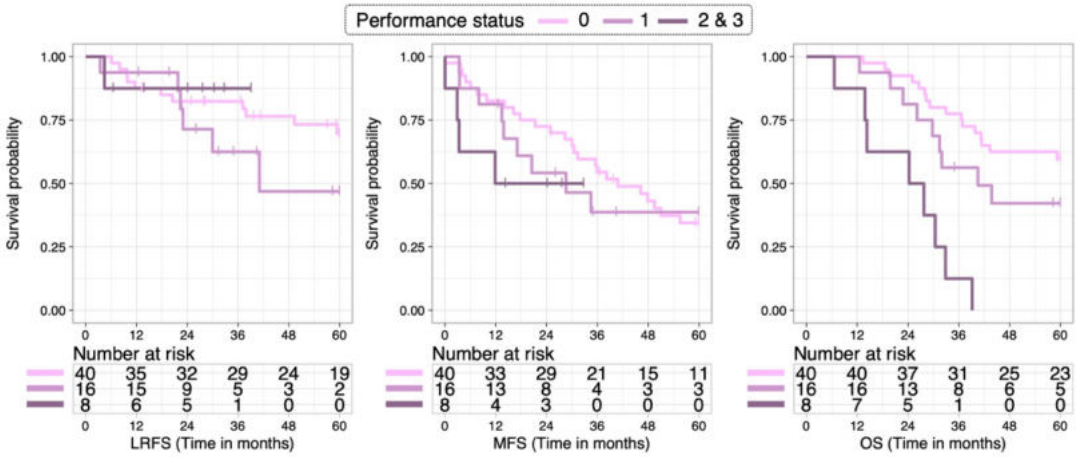
5.4.1 Supplemental figures



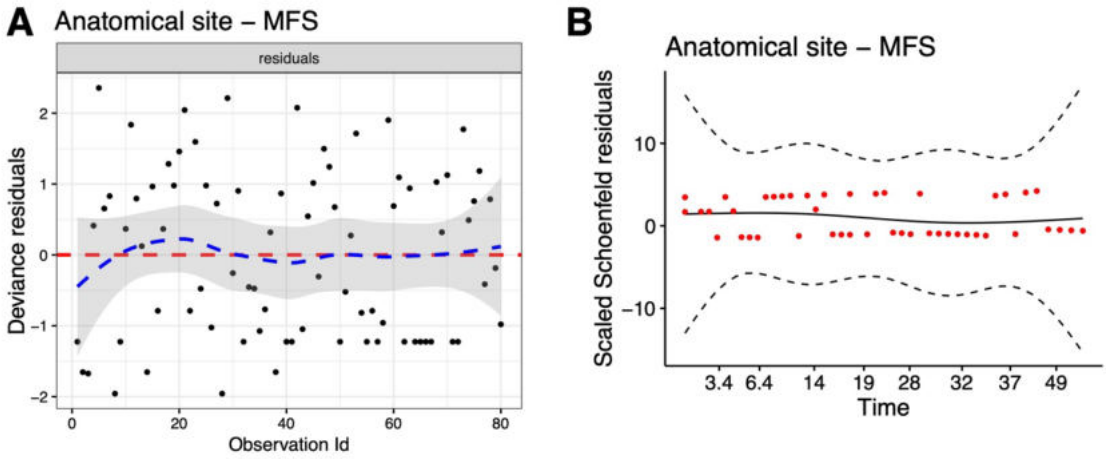
Supplemental Figure 5.1 Associations between clinicopathological variables within the leiomyosarcoma (LMS) cohort.
(A) Stacked bar plots (number and percentage) illustrating the association between anatomical site and tumour depth. **(B)** Box plots shown for associations between tumour size and depth. Boxes indicate 25th and 75th percentile, with median line in the middle, whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. Corresponding statistical tests are detailed in **Supplemental Table 5.1**.



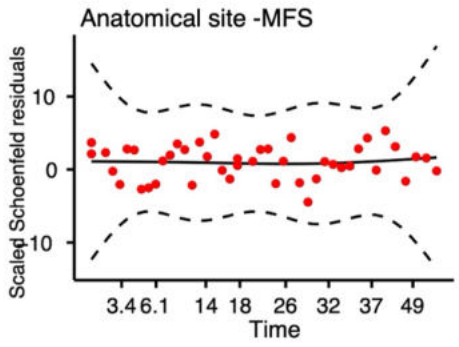
Supplemental Figure 5.2 Clinical outcome of the leiomyosarcoma (LMS) cohort stratified by significant tumour characteristics.
(A-D) Kaplan Meier plots showing from left to right, local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) up to 5-years post-surgery. **(A)** Stratification by anatomical site. **(B)** Stratification by grade. **(C)** Stratification by tumour depth. **(D)** Stratification by tumour size. Corresponding univariable Cox regression results are detailed in **Supplemental Table 5.2**.



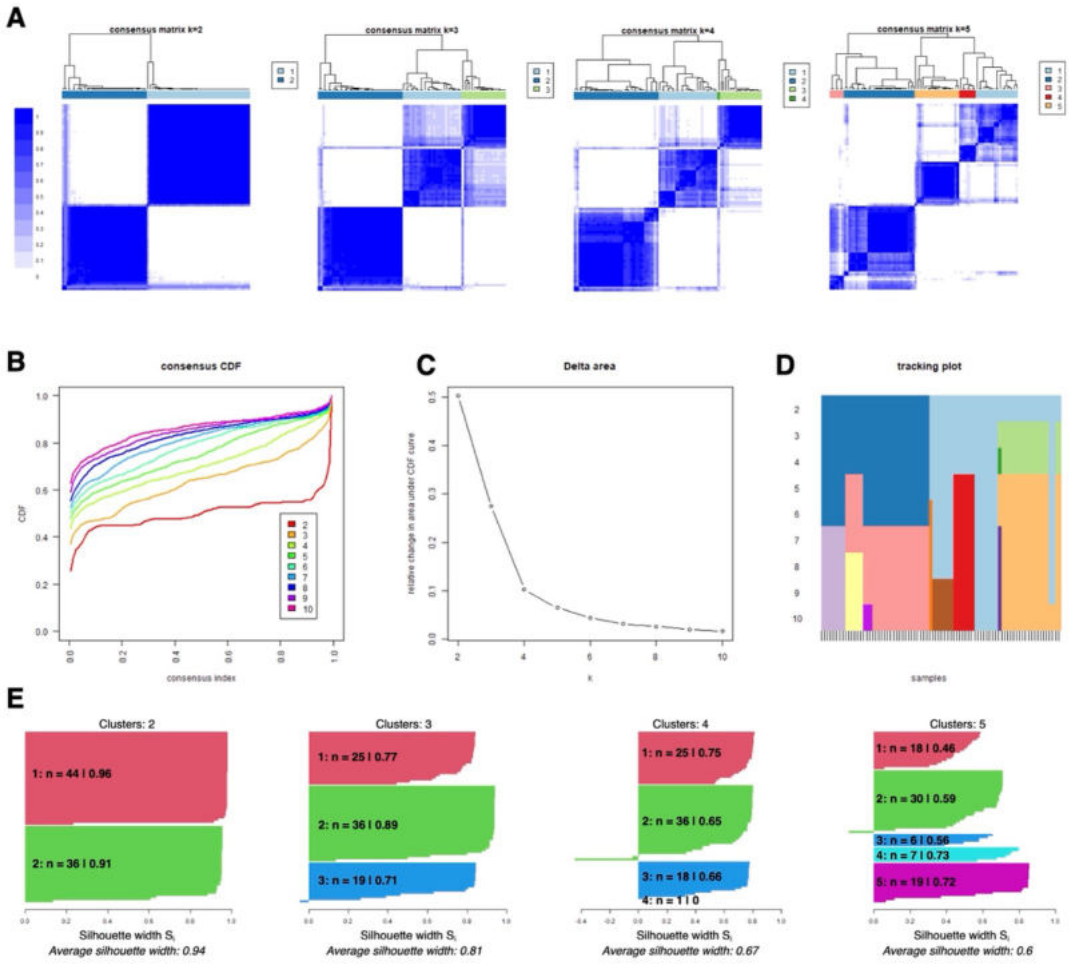
Supplemental Figure 5.3 Clinical outcome of the leiomyosarcoma (LMS) cohort stratified by significant patient characteristics.
Kaplan Meier plots showing from left to right, local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) up to 5-years post-surgery, stratified by performance status. Corresponding univariable Cox regression results are detailed in **Supplemental Table 5.2**.



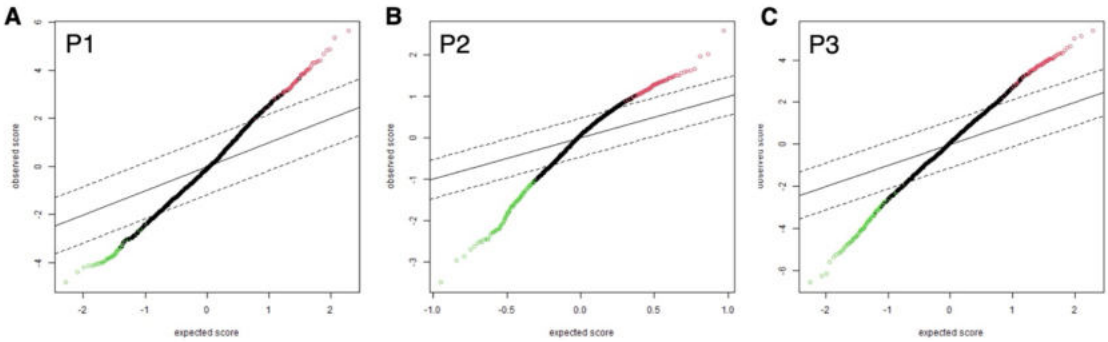
Supplemental Figure 5.4 Assessment of the proportional hazards (PH) assumption in the null univariable Cox model for leiomyosarcoma patients
Plot shown for variable-model combinatios where a minor violation of the PH assumption was identified**. (A)** Deviance residuals and **(B)** scaled Schoenfeld residuals plotted for anatomical site in the metastasis free survival (MFS) model. **(A)** Red dashed line at 0, blue line indicates a locally weighted smoothed fit and grey shading the coordinate 95% confidence intervals. **(B)** Solid black line indicates a smoothed spline fit of residuals and dashed black lines indicate +/- 2-standard error.



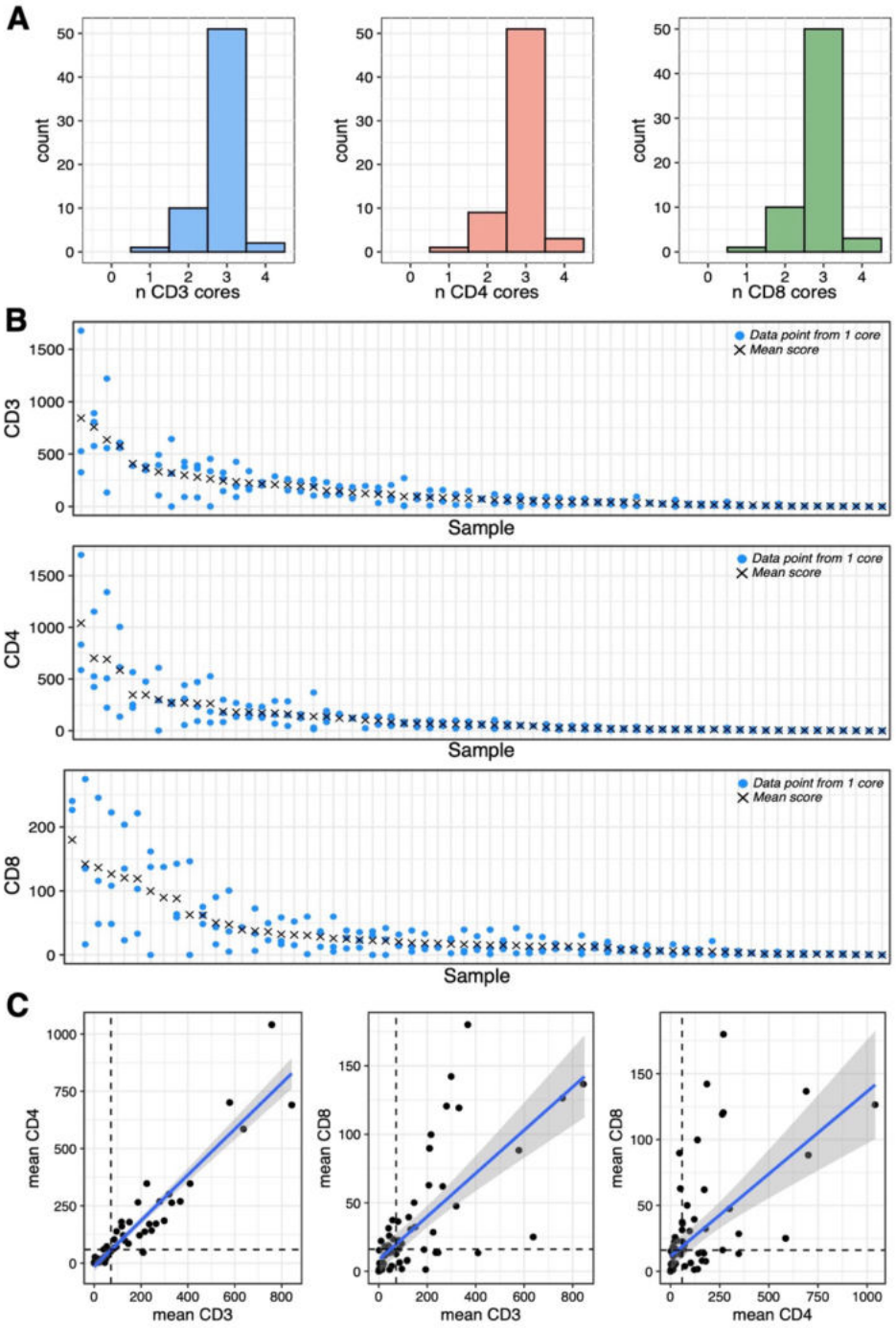
Supplemental Figure 5.5 Assessment of the proportional hazards (PH) assumption in the multivariable Cox model for leiomyosarcoma patients
Plot shown for variable-model combination where a minor violation of the PH assumption was identified. Scaled Schoenfeld residuals plotted for anatomical site in the metastasis free survival (MFS) model. Solid black line indicates a smoothed spline fit of residuals and dashed black lines indicate +/- 2-standard error.



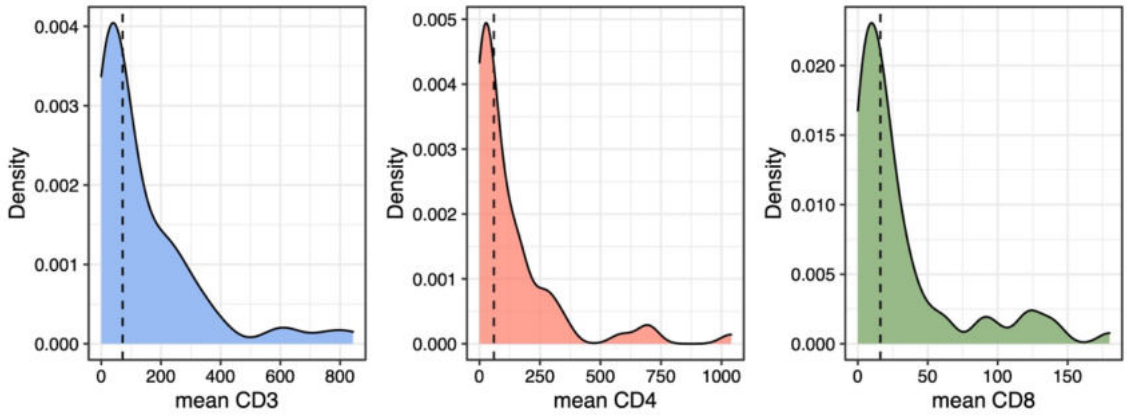
Supplemental Figure 5.6 Identification of leiomyosarcoma (LMS) proteomic subtypes. (A-E) Consensus clustering results. **(A)** Consensus matrices up to k = 5. **(B)** Consensus CDF plot up to k = 10. **(C)** Delta area plot showing relative change in area under the cumulative distribution function (CDF) curve up to k (n clusters) = 10. **(D)** Tracking plot up to k = 10. **(E)** Silhouette plots up to k = 5.



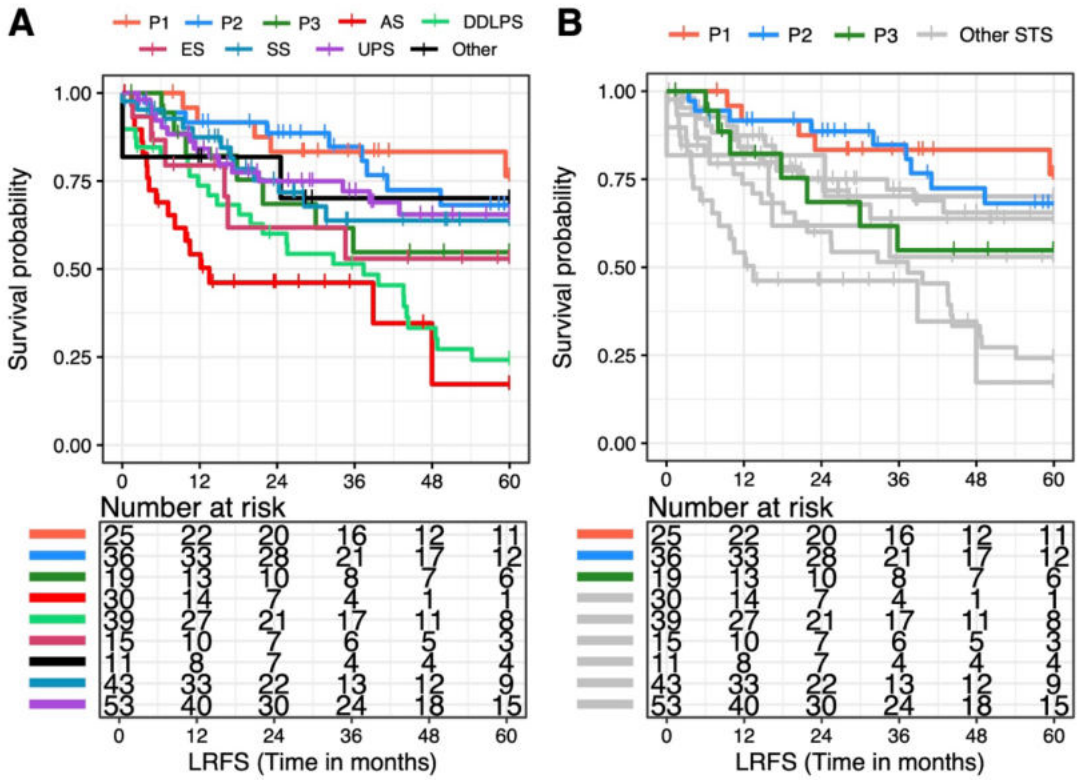
Supplemental Figure 5.7 Significant analysis of microarray (SAM) 2-class unpaired results for leiomyosarcoma (LMS) proteomic subtypes
SAM plots for LMS P1 (**A**), LMS P2 (**B**), and LMS P3 (**C**) compared to the rest of the LMS cohort. Each point is a protein. Proteins within the dashed lines have an FDR ≥ 0.01 and therefore are not significantly differentially expressed proteins (DEPs). Proteins in red are significantly upregulated DEPs (fold change ≥ 1.5) in the subtype, and proteins in green are significantly downregulated DEPs (fold change < 0.667) in the subtype.



Supplemental Figure 5.8 Assessment of the CD3+/CD4+/CD8+ tumour infiltrating lymphocyte (TIL) immunohistochemistry (IHC) tissue microarray (TMA) data in the leiomyosarcoma cohort.
(A) Histogram showing the number of TMA cores with usable CD3+/4+/8+ TIL data in the LMS cohort. **(B)** Dotplot showing inter-core variability as individual core scores and the summary mean score for each case. **(C)** Scatterplots showing the correlation between CD3+/4+/8+ TIL scores for each case. Blue line indicates the regression line of the correlation, grey shading indicates 95% confidence intervals, and dashed black lines indicate median scores.

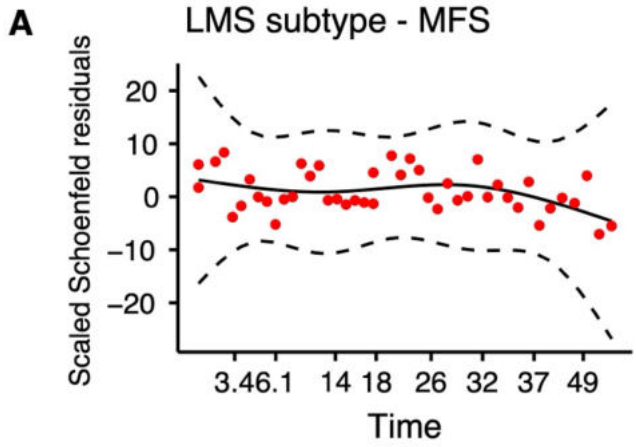


Supplemental Figure 5.9 CD3+/CD4+/CD8+ tumour infiltrating lymphocyte (TIL) burden in leiomyosarcoma (LMS)
Density plots showing the distribution of CD3+/4+/8+ TILs in LMS cases. Dashed line indicates median.

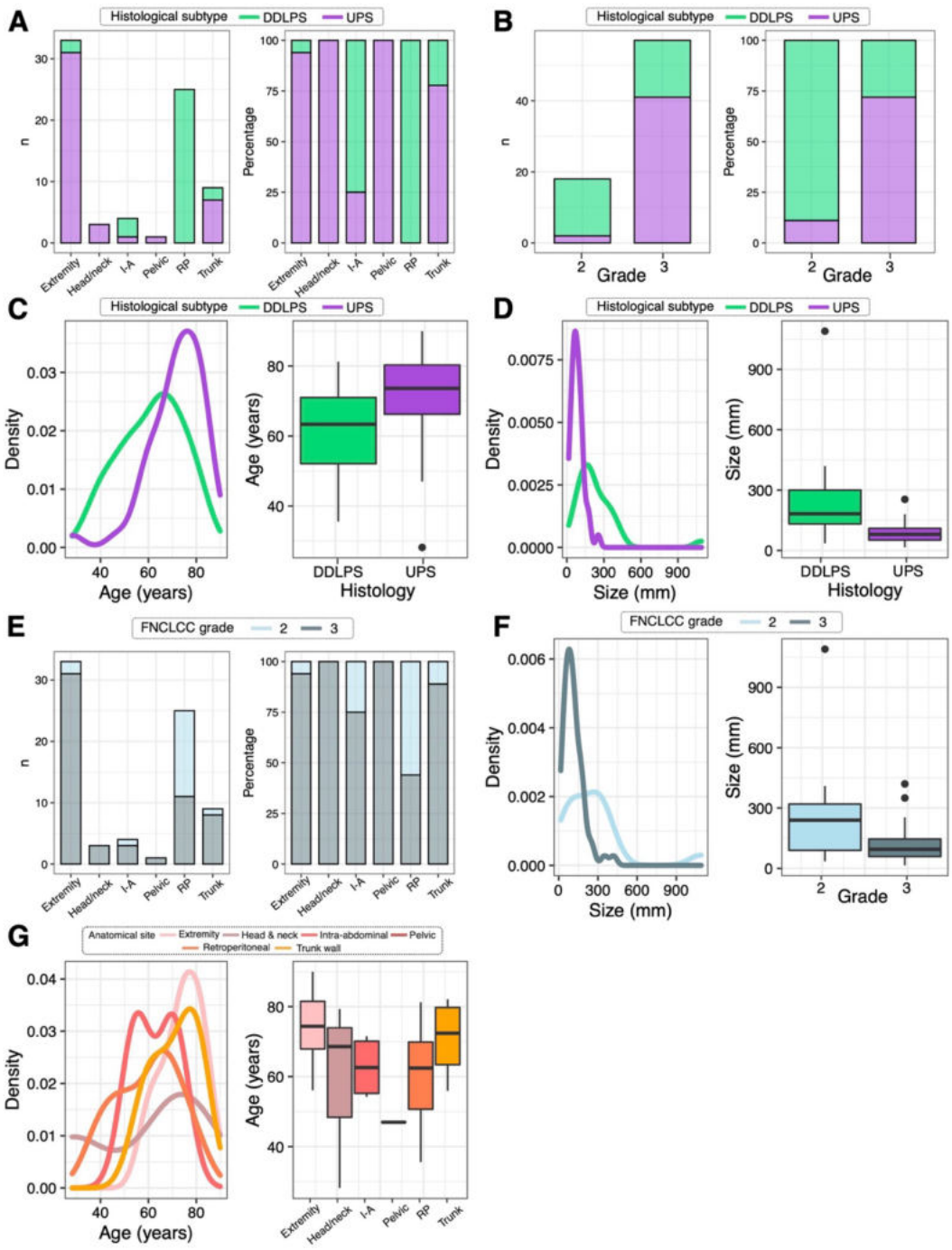


Supplemental Figure 5.10 Clinical outcome of the proteome-profiled cohort stratified by histological subtype and leiomyosarcoma (LMS) proteomic subtype.

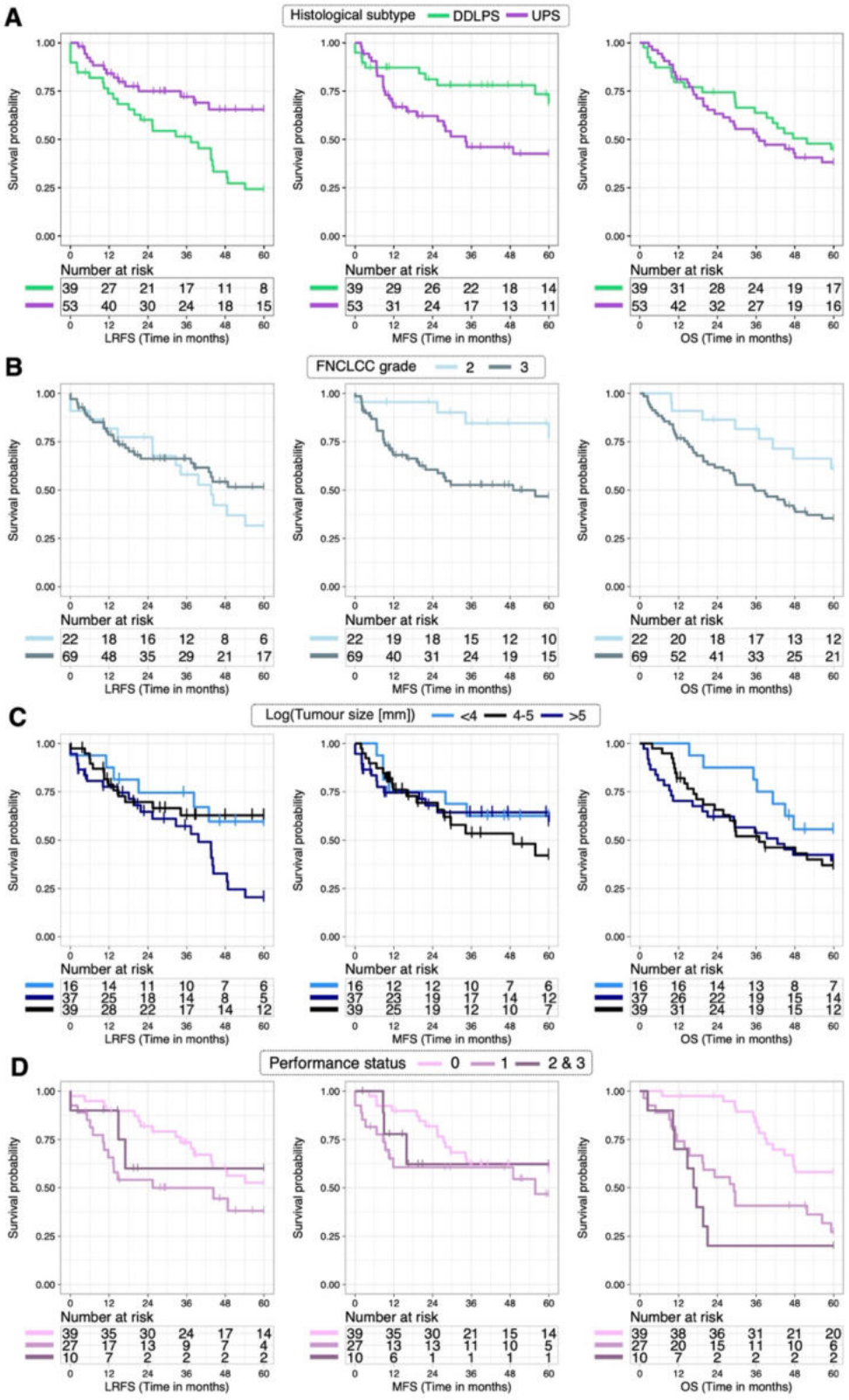
Kaplan Meier plots showing local recurrence free survival (LRFS) up to 5-years post-surgery. **(A)** Plot coloured by histological and proteomic subtype. **(B)** Plot coloured by proteomic subtype. All non-LMS cases in grey. Abbreviations: AS = angiosarcoma; DDLPS = dedifferentiated liposarcoma; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma; STS = soft tissue sarcoma



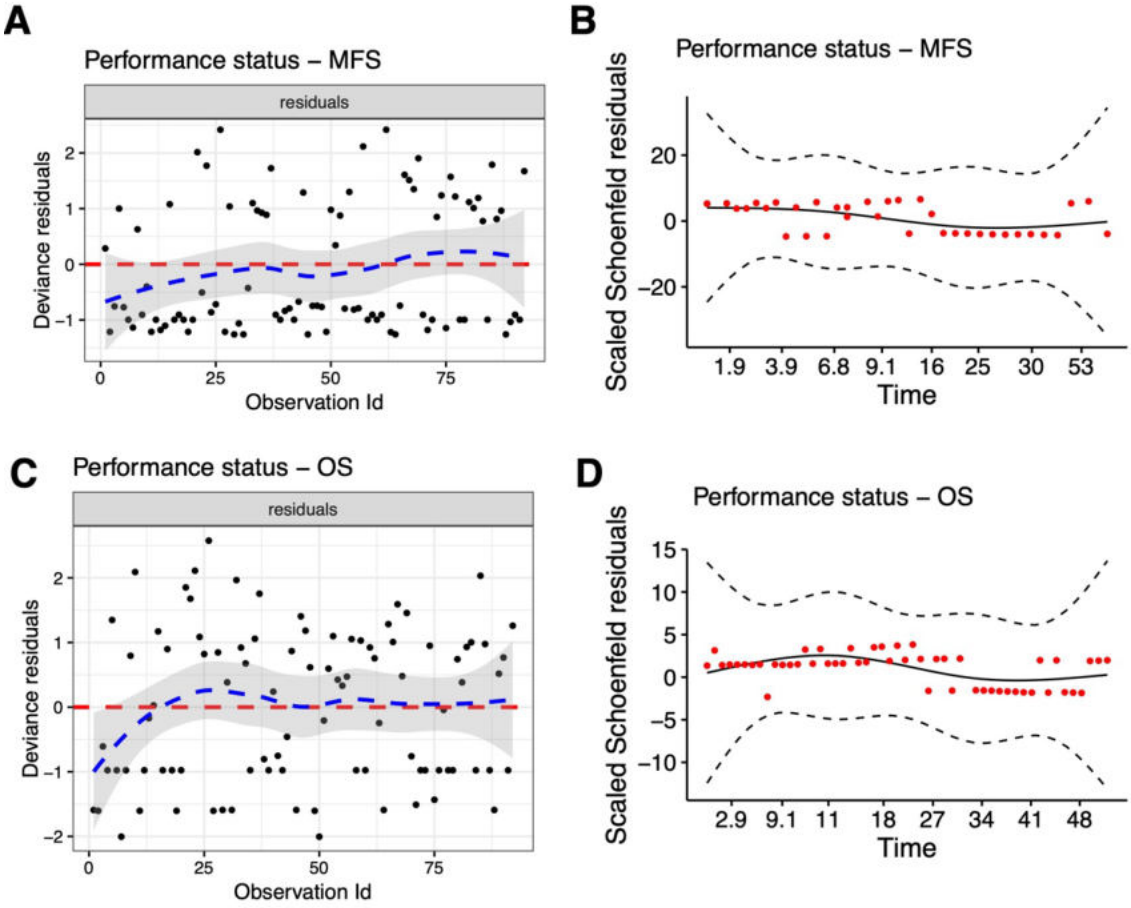
Supplemental Figure 5.11 Assessment of the proportional hazards (PH) assumption in the multivariable Cox model for leiomyosarcoma (LMS) patients including proteomic subtype
Plot shown for variable-model combination where a minor violation of the PH assumption was identified. Scaled Schoenfeld residuals plotted for LMS proteomic subtype in the metastasis free survival (MFS) model. Solid black line indicates a smoothed spline fit of residuals and dashed black lines indicate +/- 2-standard error.



Supplemental Figure 5.12 Associations between clinicopathological variables within the dedifferentiated liposarcoma (DDLPS) and undifferentiated pleomorphic sarcoma (UPS) cohort.
(A-G) Density plots and box plots are shown for associations between continuous and categorical variables. Boxes indicate 25th and 75th percentile, with median line in the middle, whiskers extending from 25th percentile-(1.5*IQR) to 75th percentile+(1.5*IQR), and outliers plotted as points. Stacked bar plots for number and percentage are shown for associations between 2 categorical variables. Plots illustrate the relationship between **(A)** histological subtype and anatomical site, **(B)** histological subtype and grade, **(C)** histological subtype and age, **(D)** histological subtype and tumour size, **(E)** anatomical site and grade, **(F)** grade and tumour size, **(G)** anatomical site and age. Abbreviations: FNCLCC = French Federation of Cancer Center Sarcoma Group; DDLPS = dedifferentiated liposarcoma; UPS = undifferentiated pleomorphic sarcoma; I-A = intra-abdominal; RP = retroperitoneal. Corresponding statistical tests are detailed in **Supplemental Table 5.6**

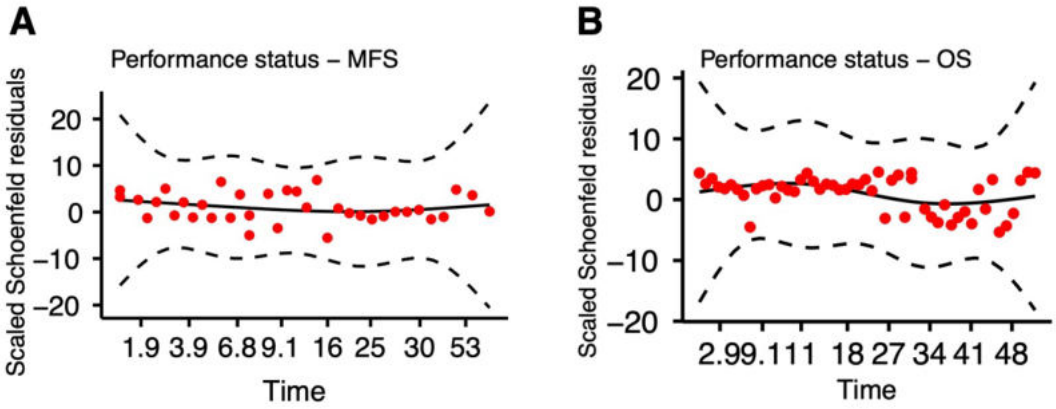


Supplemental Figure 5.13 Clinical outcome of the dedifferentiated liposarcoma (DDLPS) and undifferentiated pleomorphic sarcoma (UPS) cohort stratified by significant characteristics.
(A-D) Kaplan Meier plots showing from left to right, local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) up to 5-years post-surgery. **(A)** Stratification by histological subtype. **(B)** Stratification by grade. **(C)** Stratification by tumour size. **(D)** Stratification by performance status. Corresponding univariable Cox regression results are detailed in **Supplemental Table 5.7**

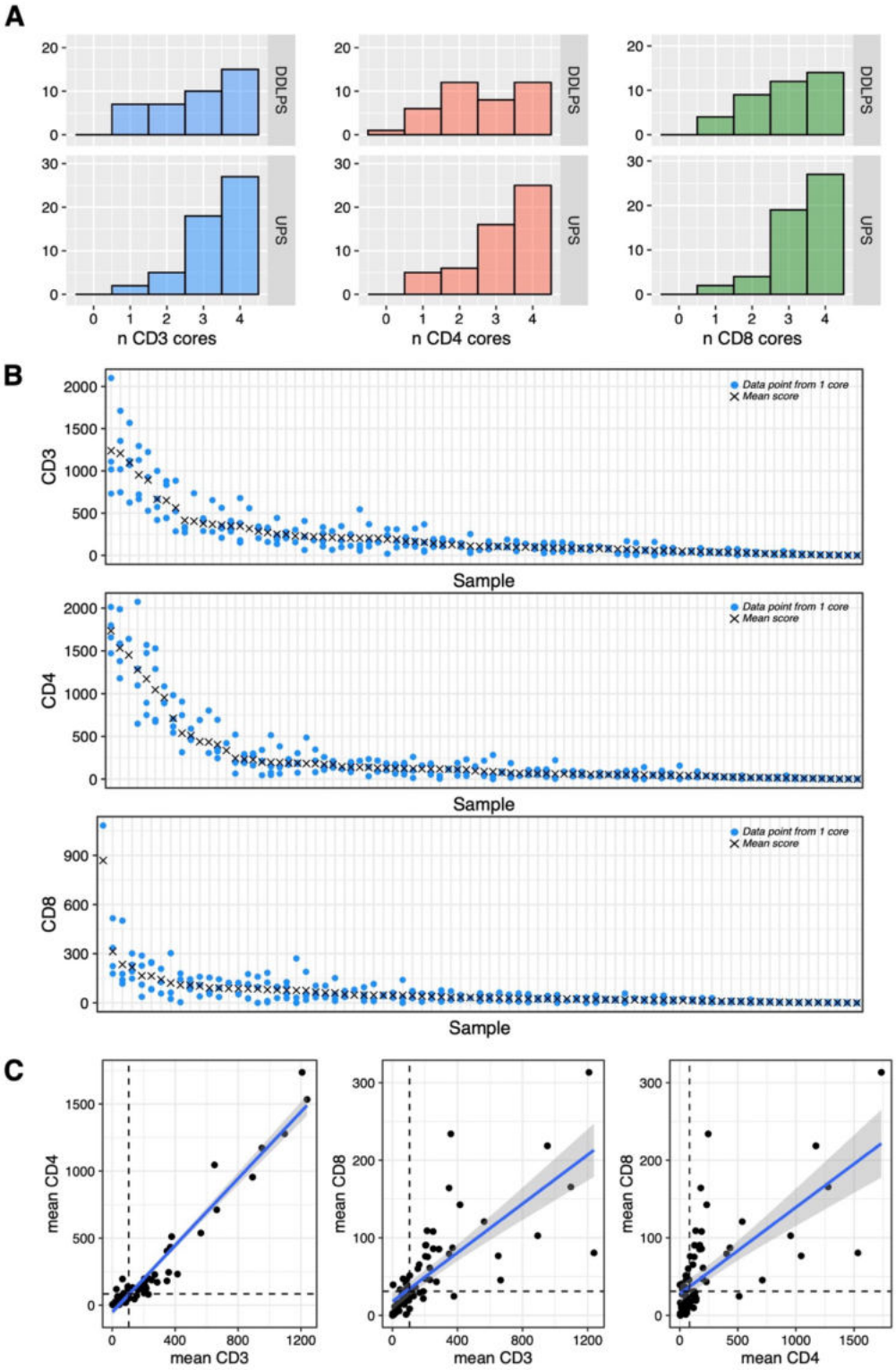


Supplemental Figure 5.14 Assessment of the proportional hazards (PH) assumption in the null univariable Cox models of dedifferentiated liposarcoma and undifferentiated pleomorphic sarcoma patients.

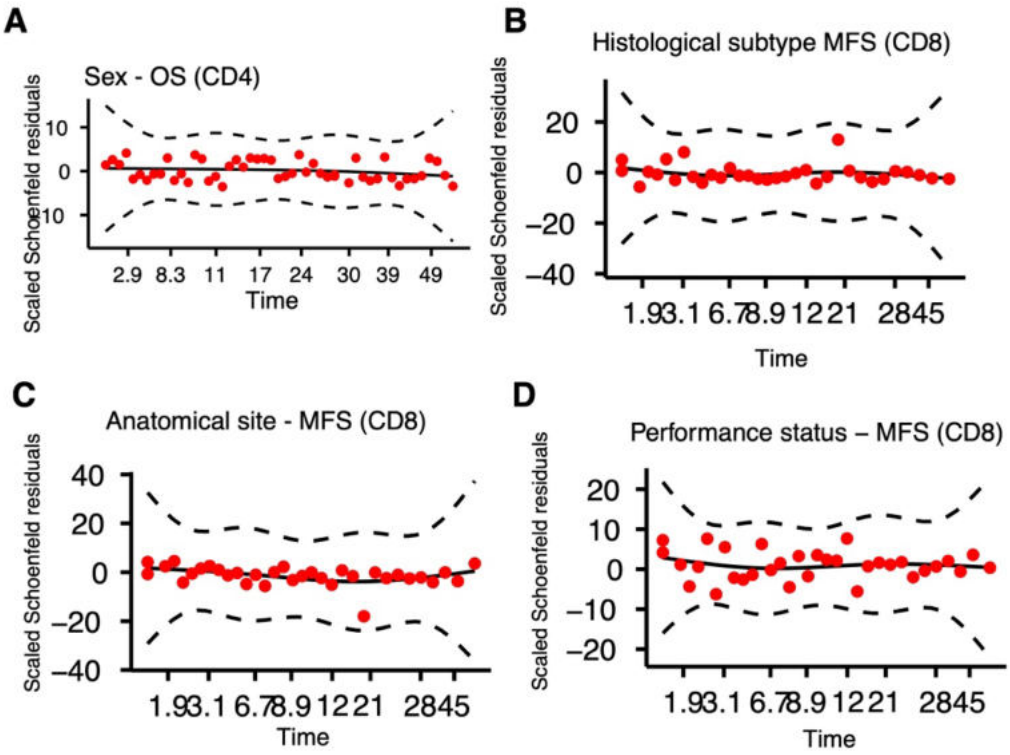
Plots shown for variable-model combinations where a minor violation of the PH assumption was identified: **(A-B)** performance status and metastasis free survival (MFS); **(C-D)** performance status and overall survival (OS). Deviance residuals **(A)** plotted for each observation. Red dashed line at 0, blue line indicates a locally weighted smoothed fit and grey shading the coordinate 95% confidence intervals. Scaled Schoenfeld residuals **(B,D)** plotted over time for each observation. Solid black line indicates a smoothed spline fit of residuals and dashed black lines indicate +/- 2-standard error.



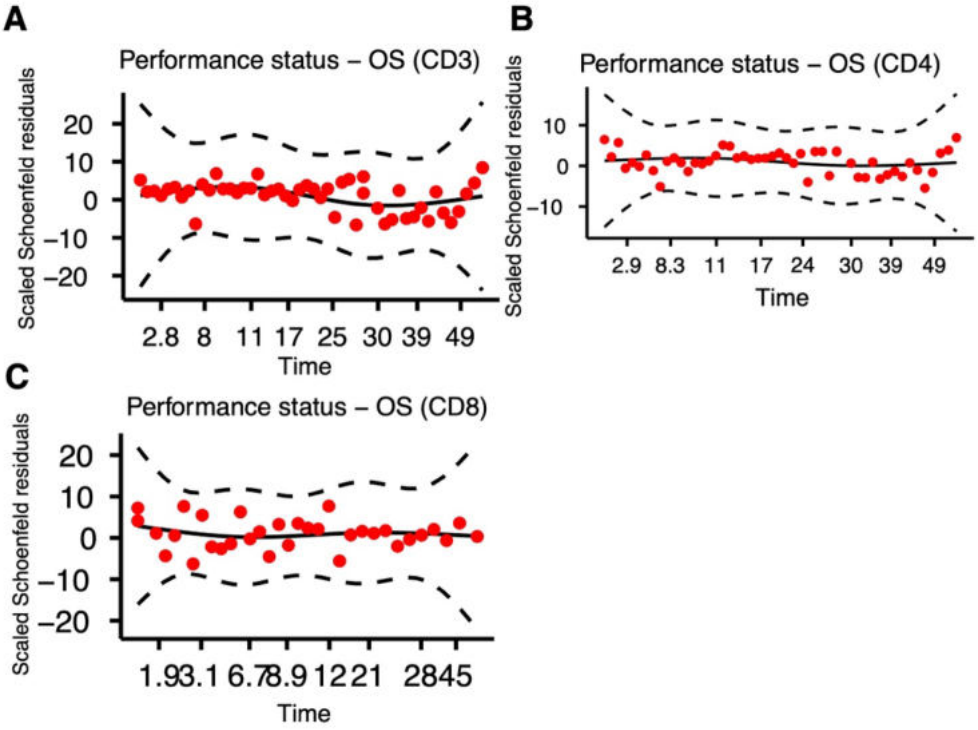
Supplemental Figure 5.15 Assessment of the proportional hazards (PH) assumption in the multivariable Cox models of dedifferentiated liposarcoma and undifferentiated pleomorphic sarcoma patients.
Scaled Schoenfeld residuals plotted over time for each observation. Solid black line indicates a smoothed spline fit of residuals and dashed black lines indicate +/- 2-standard error. Plots shown for variable-model combinations where a minor violation of the PH assumption was identified: **(A)** performance status and metastasis free survival (MFS); **(B)** performance status and overall survival (OS).



Supplemental Figure 5.16 Assessment of the CD3+/CD4+/CD8+ tumour infiltrating lymphocyte (TIL) immunohistochemistry (IHC) tissue microarray (TMA) data in the dedifferentiated liposarcoma and undifferentiated pleomorphic sarcoma cohort.
(A) Histogram showing the number of TMA cores with usable CD3+/4+/8+ TIL data in the LMS cohort. **(B)** Dotplot showing inter-core variability as individual core scores and the summary mean score for each case. **(C)** Scatterplots showing the correlation between CD3+/4+/8+ TIL scores for each case. Blue line indicates the regression line of the correlation, grey shading indicates 95% confidence intervals, and dashed black lines indicate median scores



Supplemental Figure 5.17 Assessment of the proportional hazards (PH) assumption in the multivariable Cox models of dedifferentiated liposarcoma and undifferentiated pleomorphic sarcoma patients including tumour infiltrating lymphocyte (TIL) burden.
Scaled Schoenfeld residuals plotted over time for each observation. Solid black line indicates a smoothed spline fit of residuals and dashed black lines indicate +/- 2-standard error. Plots shown for variable-model combinations where a minor violation of the PH assumption was identified: **(A)** sex and overall survival (OS) in the CD4+ TIL model; **(B)** histological subtype and metastasis free survival (MFS) in the CD8+ TIL model; **(C)** anatomical site and MFS in the CD8+ TIL model **(D)** performance status and MFS in the CD8+ TIL model.



Supplemental Figure 5.18 Assessment of proportional hazards (PH) assumption violations in the multivariable Cox models of dedifferentiated liposarcoma and undifferentiated pleomorphic sarcoma patients including tumour infiltrating lymphocyte (TIL) burden.
Scaled Schoenfeld residuals plotted over time for each observation. Solid black line indicates a smoothed spline fit of residuals and dashed black lines indicate +/- 2-standard error. Plots shown for variable-model combinations where a violation of the PH assumption was identified: performance status and overall survival (OS) in the **(A)** CD3+ TIL model, **(B)** CD4+ TIL model, and **(C)** CD8+ TIL model.

5.4.2 Supplemental tables

Supplemental Table 5.1 Statistical associations between clinicopathological features of the leiomyosarcoma cohort.
Significant results in bold.

Variable 1	Variable 2	Test performed	Test statistic	Degrees of freedom	p	FDR
Anatomical site	**Tumour depth**	**Chi-squared**	**X2 = 21.56**	**5**	**< 0.001**	**0.013**
Tumour depth	Performance status	Chi-squared	X2 = 7.965	3	0.047	0.178
Anatomical site	Performance status	Chi-squared	X2 = 24.924	15	0.051	0.178
Sex	Anatomical site	Chi-squared	X2 = 10.015	5	0.075	0.233
Sex	Grade	Chi-squared	X2 = 1.415	1	0.234	0.459
Anatomical site	Grade	Chi-squared	X2 = 5.943	5	0.312	0.514
Tumour margin	Grade	Chi-squared	X2 = 2.107	2	0.349	0.543
Sex	Tumour depth	Chi-squared	X2 = 0.697	1	0.404	0.595
Sex	Performance status	Chi-squared	X2 = 2.619	3	0.454	0.611
Grade	Performance status	Chi-squared	X2 = 2.476	3	0.48	0.611
Anatomical site	Tumour margin	Chi-squared	X2 = 9.543	10	0.481	0.611
Tumour depth	Tumour margin	Chi-squared	X2 = 1.198	2	0.549	0.641
Sex	Tumour margin	Chi-squared	X2 = 0.881	2	0.644	0.716
Tumour depth	Grade	Chi-squared	X2 = 0.188	1	0.665	0.716
Tumour margin	Performance status	Chi-squared	X2 = 1.383	4	0.847	0.878
Tumour size	**Tumour depth**	**Kruskal-Wallis**	**X2 = 10.996**	**1**	**<0.001**	**0.013**
Age	Anatomical site	Kruskal-Wallis	X2 = 13.992	5	0.016	0.149
Age	Performance status	Kruskal-Wallis	X2 = 7.559	2	0.023	0.161
Age	Grade	Kruskal-Wallis	X2 = 4.436	1	0.035	0.178
Tumour size	Sex	Kruskal-Wallis	X2 = 3.919	1	0.048	0.178
Age	Tumour margin	Kruskal-Wallis	X2 = 4.655	2	0.098	0.273
Tumour size	Anatomical site	Kruskal-Wallis	X2 = 14.212	5	0.143	0.364
Age	Tumour depth	Kruskal-Wallis	X2 = 1.905	1	0.168	0.392
Tumour size	Tumour margin	Kruskal-Wallis	X2 = 2.822	2	0.244	0.459
Tumour size	Performance status	Kruskal-Wallis	X2 = 2.807	2	0.246	0.459
Tumour size	Grade	Kruskal-Wallis	X2 = 1.041	1	0.308	0.514
Age	Sex	Kruskal-Wallis	X2 = 0.4515	1	0.502	0.611
Age	Tumour size	Spearman's rank correlation	Rho = 0.007	-	0.948	0.948

Supplemental Table 5.2 Univariable Cox regression for leiomyosarcoma patients.
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval

		LRFS		MFS		OS	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
Age at excision (years)		0.983 (0.956-1.01)	0.217	1 (0.984-1.02)	0.707	1.01 (0.99-1.04)	0.249
Sex	*F (ref)*	-	-	-	-	-	-
	M	1.6 (0.662-3.86)	0.298	0.596 (0.295-1.21)	0.15	1.6 (0.832-3.06)	0.16
Anatomical site	*Intra-abdominal & Pelvic & retroperitoneal (ref)*	-	-	-	-	-	-
	Trunk wall & extremity	**0.27 (0.111-0.657)**	**0.004**	**0.392 (0.207-0.742)**	**0.004**	**0.421 (0.21-0.844)**	**0.0147**
	Uterine	0.185 (0.024-1.45)	0.109	0.714 (0.259-1.97)	0.514	0.833 (0.299-2.32)	0.726
FNCLCC grade	*2 (ref)*	-	-	-	-	-	-
	3	1.07 (0.443-2.6)	0.875	**2.46 (1.35-4.46)**	**0.003**	1.62 (0.857-3.07)	0.137
Performance status	*0 (ref)*	-	-	-	-	-	-
	1	1.87 (0.681-5.14)	0.225	1.25 (0.58-2.68)	0.571	1.82 (0.802-4.13)	0.152
	2-3	0.859 (0.108-6.81)	0.886	2.12 (0.717-6.25)	0.175	**7.82 (3.15-19.4)**	**< 0.001**
	unknown	0.735 (0.205-2.64)	0.637	0.793 (0.343-1.84)	0.589	0.79 (0.289-2.16)	0.645
Tumour depth	*Deep (ref)*	-	-	-	-	-	-
	Superficial	0.225 (0.03-1.68)	0.145	**0.291 (0.09-0.94)**	**0.039**	0.767 (0.299-1.97)	0.58
Tumour margin	*R0 (ref)*	-	-	-	-	-	-
	R1 & R2	1.57 (0.651-3.81)	0.314	1.06 (0.583-1.91)	0.859	1.04 (0.549-1.96)	0.908
	unknown	1.97 (0.248-15.6)	0.522	0.694 (0.094-5.14)	0.72	3.97e-08 (0-Inf)	0.997
Log[tumour size(mm)]	4-5 (ref)	-	-	-	-	-	-
	< 4	0.505 (0.114-2.24)	0.369	**0.226 (0.069-0.74)**	**0.014**	0.426 (0.128-1.41)	0.162
	> 5	1.3 (0.494-3.42)	0.596	0.635 (0.303-1.33)	0.228	1.12 (0.537-2.33)	0.767

Supplemental Table 5.3 Multivariable Cox regression for leiomyosarcoma patients.
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval

		LRFS		MFS		OS	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
Age at excision (years)		0.966 (0.923-1.01)	0.14	1.02 (0.988-1.05)	0.233	0.995 (0.962-1.03)	0.798
Sex	*F (ref)*	-	-	-	-	-	-
	M	2.92 (0.827-10.3)	0.096	0.739 (0.329-1.66)	0.465	1.9 (0.803-4.49)	0.144
Anatomical site	*Intra-abdominal & Pelvic & retroperitoneal (ref)*	-	-	-	-	-	-
	Trunk wall & extremity	**0.303 (0.093-0.989)**	**0.048**	**0.288 (0.126-0.66)**	**0.003**	**0.345 (0.136-0.878)**	**0.026**
	Uterine	0.247 (0.023-2.66)	0.248	0.699 (0.191-2.56)	0.588	0.519 (0.151-1.79)	0.299
FNCLCC grade	*2 (ref)*	-	-	-	-	-	-
	3	1.78 (0.637-4.99)	0.271	**2.41 (1.22-4.77)**	**0.011**	**2.19 (1.04-4.6)**	**0.039**
Performance status	*0 (ref)*	-	-	-	-	-	-
	1	**4.66 (1.22-17.8)**	**0.024**	1.68 (0.691-4.07)	0.253	2.08 (0.803-5.41)	0.131
	2-3	3.55 (0.321-39.2)	0.302	2.1 (0.577-7.62)	0.26	**15.9 (4.39-57.2)**	**< 0.001**
	unknown	1.66 (0.375-7.3)	0.505	0.985 (0.395-2.45)	0.974	0.905 (0.3-2.73)	0.86
Tumour depth	*Deep (ref)*	-	-	-	-	-	-
	Superficial	0.253 (0.024-2.7)	0.255	0.282 (0.067-1.19)	0.085	0.839 (0.238-2.96)	0.784
Tumour margin	*R0 (ref)*	-	-	-	-	-	-
	R1 & R2	0.448 (0.158-1.27)	0.131	1.13 (0.558-2.27)	0.741	0.638 (0.292-1.39)	0.258
Log[tumour size(mm)]	4-5 (ref)	-	-	-	-	-	-
	< 4	0.784 (0.122-5.02)	0.797	0.499 (0.133-1.88)	0.303	0.862 (0.195-3.82)	0.846
	> 5	1.37 (0.352-5.3)	0.651	0.719 (0.28-1.85)	0.493	1.87 (0.694-5.07)	0.215

Supplemental Table 5.4 Statistical associations between leiomyosarcoma (LMS) clinicopathological features and proteomic subtype.
Abbreviations: RTX = radiotherapy.

Variable		LMS subtype			Test results			
		P1 (immune cold) n = 25	P2 (classical) n = 36	P3 (dedifferentiated) n = 19	Test performed	X^2	Degrees of freedom	p
Age at excision (years)	median	61.5	66.8	65.4	Kruskal Wallis	0.373	2	0.83
	min	31.4	30.5	29.3				
	max	83.6	86.9	83.5				
Tumour size (mm)	median	100	80	110	Kruskal Wallis	3.131	2	0.209
	min	50	25	5				
	max	400	290	250				
Sex [n (%)]	F	18 (72.0)	26 (72.2)	12 (63.2)	Chi-squared	0.556	2	0.757
	M	7 (28.0)	10 (27.8)	7 (36.8)				
Grade [n (%)]	2	15 (60.0)	21 (58.3)	11 (57.9)	Chi-squared	0.024	2	0.988
	3	10 (40.0)	15 (41.7)	8 (42.1)				
Performance status [n (%)]	0	15 (60.0)	17 (47.2)	8 (42.1)	Chi-squared	13.304	8	0.102
	1	2 (8.0)	12 (33.3)	2 (10.5)				
	2	2 (8.0)	3 (8.3)	2 (10.5)				
	3	-	1 (2.7)	-				
	unknown	6 (24.0)	3 (8.3)	7 (36.8)				
Pre-op treatment [n (%)]	RTX	-	-	1 (5.3)	Chi-squared	3.251	2	0.197
	None	25 (100.0)	36 (100.0)	18 (94.7)				
Anatomical site [n (%)]	Extremity	7 (28.0)	14 (38.9)	10 (52.6)	Chi-squared	12.032	10	0.283
	Intra-abdominal	5 (20.0)	4 (11.1)	1 (5.3)				
	Pelvic	1 (4.0)	5 (13.9)	3 (15.8)				
	Retroperitoneal	9 (36.0)	9 (25.0)	1 (5.3)				
	Trunk	1 (4.0)	-	1 (5.3)				
	Uterine	2 (8.0)	4 (11.1)	3 (15.8)				
Status at excision [n (%)]	Local	24 (96.0)	36 (100.0)	18 (94.7)	Chi-squared	1.749	2	0.417
	Metastatic	1 (4.0)	-	1 (5.3)				
Tumour depth [n (%)]	Deep	22 (88.0)	28 (77.8)	16 (84.2)	Chi-squared	1.118	2	0.572
	Superficial	3 (12.0)	8 (22.2)	3 (15.8)				
Tumour margins [n (%)]	R0	13 (52.0)	20 (58.8)	9 (47.4)	Chi-squared	5.342	6	0.501
	R1	11 (44.0)	14 (41.2)	10 (52.6)				
	R2	1 (4.0)	-	-				

Supplemental Table 5.5 Comparison of the baseline clinicopathological factors in the proteomic and The Cancer Genome Atlas (TCGA) leiomyosarcoma (LMS) cohorts

Chi-squared tests performed. Significant results in bold.

		TCGA cohort (n = 80)				Proteomic cohort (n = 80)				p	FDR
		Observed	Expected	Residuals	Contribution (%)	Observed	Expected	Residuals	Contribution (%)		
Sex	F	55	55.5	-0.067	NA	56	55.5	0.067	NA	1.000	1.000
	M	25	24.5	0.101	NA	24	24.5	-0.101	NA		
Anatomical site	Extremity	14	22.5	-1.792	16.867	31	22.5	1.792	16.867	0.004	**0.005**
	Head/neck	1	0.5	0.707	2.625	0	0.5	-0.707	2.625		
	Intra-abdominal	14	12	0.577	1.749	10	12	-0.577	1.749		
	Pelvic	4	6.5	-0.981	5.055	9	6.5	0.981	5.055		
	Retroperitoneal	18	18.5	-0.116	0.071	19	18.5	0.116	0.071		
	Trunk	2	2	0	0	2	2	0	0		
	Uterine	27	18	2.121	23.629	9	18	-2.121	23.629		
Tumour depth	Deep	67	66.5	0.061	0.016	66	66.5	-0.061	0.016	< 0.001	**< 0.001**
	Superficial	1	7.5	-2.373	24.195	14	7.5	2.373	24.195		
	UNK	12	6	2.449	25.769	0	6	-2.449	25.769		
Tumour margin	R0	56	48	1.155	7.919	40	48	-1.155	7.919	0.001	**0.001**
	R1	12	23.5	-2.372	33.401	35	23.5	2.372	33.401		
	R2	3	2	0.707	2.967	1	2	-0.707	2.967		
	Rx	9	6.5	0.981	5.713	4	6.5	-0.981	5.713		
Grade	1	12	6	2.449	34.702	0	6	-2.449	34.702	0.000	**< 0.001**
	2	51	49	0.286	0.473	47	49	-0.286	0.473		
	3	17	25	-1.6	14.812	33	25	1.6	14.812		

Supplemental Table 5.6 Statistical associations between clinicopathological features of the dedifferentiated liposarcoma and undifferentiated pleomorphic sarcoma cohort.
Significant results in bold.

Variable 1	Variable 2	Test performed	Test statistic	Degrees of freedom	p	FDR
Histological subtype	**Grade**	**Chi-squared**	**18.274**	**1**	**0.000**	**< 0.001**
Histological subtype	**Anatomical site**	**Chi-squared**	**57.895**	**5**	**0.000**	**< 0.001**
Anatomical site	**Grade**	**Chi-squared**	**21.943**	**5**	**0.001**	**0.004**
Tumour margin	Grade	Chi-squared	6.592	2	0.037	0.194
Histological subtype	Tumour depth	Chi-squared	3.610	1	0.057	0.241
Anatomical site	Tumour depth	Chi-squared	8.042	5	0.154	0.377
Tumour depth	Grade	Chi-squared	2.284	1	0.131	0.377
Tumour depth	Performance status	Chi-squared	3.644	2	0.162	0.377
Tumour margin	Performance status	Chi-squared	6.984	4	0.137	0.377
Performance status	Grade	Chi-squared	2.716	2	0.257	0.540
Histological subtype	Sex	Chi-squared	0.886	1	0.347	0.662
Sex	Anatomical site	Chi-squared	4.477	5	0.483	0.670
Sex	Tumour depth	Chi-squared	0.435	1	0.510	0.670
Sex	Tumour margin	Chi-squared	1.690	2	0.430	0.670
Anatomical site	Tumour margin	Chi-squared	9.227	10	0.511	0.670
Histological subtype	Tumour margin	Chi-squared	3.532	4	0.473	0.670
Anatomical site	Performance status	Chi-squared	8.838	10	0.548	0.676
Sex	Performance status	Chi-squared	1.077	2	0.584	0.681
Tumour depth	Tumour margin	Chi-squared	0.950	2	0.622	0.687
Histological subtype	Performance status	Chi-squared	0.757	2	0.685	0.719
Sex	Grade	Chi-squared	0.128	1	0.720	0.720
Tumour size	**Anatomical site**	**Kruskal-Wallis**	**33.341**	**5**	**0.000**	**< 0.001**
Tumour size	**Histological subtype**	**Kruskal-Wallis**	**28.181**	**1**	**0.000**	**< 0.001**
Age	**Histological subtype**	**Kruskal-Wallis**	**11.165**	**1**	**0.001**	**0.004**
Tumour size	**Tumour depth**	**Kruskal-Wallis**	**10.559**	**1**	**0.001**	**0.004**
Age	**Anatomical site**	**Kruskal-Wallis**	**19.373**	**5**	**0.002**	**0.005**
Tumour size	**Grade**	**Kruskal-Wallis**	**8.550**	**1**	**0.003**	**0.008**
Age	**Grade**	**Kruskal-Wallis**	**6.791**	**1**	**0.009**	**0.018**
Age	Performance status	Kruskal-Wallis	5.791	2	0.055	0.097
Age	Tumour depth	Kruskal-Wallis	2.843	1	0.092	0.129
Tumour size	Tumour margin	Kruskal-Wallis	2.851	1	0.091	0.129
Tumour size	Sex	Kruskal-Wallis	2.256	1	0.133	0.169
Age	Sex	Kruskal-Wallis	0.422	1	0.516	0.602
Tumour size	Performance status	Kruskal-Wallis	0.779	2	0.677	0.729
Age	Tumour margin	Kruskal-Wallis	0.085	1	0.771	0.771

Supplemental Table 5.7 Univariable Cox regression for dedifferentiated liposarcoma (DDLPS) and undifferentiated pleomorphic sarcoma (UPS) patients.
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. Anatomical site of 'Other' indicates extremity, trunk wall, and head/neck cases. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval; IA = Intra-abdominal; RP = retroperitoneal

		LRFS		MFS		OS	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
Age at excision (years)		1.01 (0.983-1.03)	0.624	1.03 (0.998-1.05)	0.068	**1.04 (1.02-1.07)**	**0.001**
Sex	*M (ref)*	-	-	-	-	-	-
	F	0.889 (0.482-1.64)	0.705	1.4 (0.726-2.69)	0.316	0.936 (0.541-1.62)	0.812
Histological subtype	*UPS (ref)*	-	-	-	-	-	-
	DDLPS	**2.63 (1.4-4.94)**	**0.003**	**0.426 (0.205-0.889)**	**0.023**	0.814 (0.467-1.42)	0.467
Anatomical site	*Other (ref)*	-	-	-	-	-	-
	IA/RP	2 (0.482-8.33)	0.339	3.91e-08 (0-Inf)	0.997	2.56 (0.796-8.26)	0.115
FNCLCC grade	*3 (ref)*	-	-	-	-	-	-
	2	1.39 (0.726-2.65)	0.322	**0.284 (0.1-0.809)**	**0.018**	**0.435 (0.204-0.926)**	**0.031**
Performance status	*0 (ref)*	-	-	-	-	-	-
	1	1.91 (0.944-3.88)	0.072	1.6 (0.746-3.42)	0.228	**2.66 (1.35-5.24)**	**0.005**
	2-3	1.23 (0.355-4.24)	0.746	1.3 (0.371-4.57)	0.681	**4.21 (1.76-10.1)**	**0.001**
	unknown	1.98 (0.845-4.65)	0.116	1.48 (0.573-3.84)	0.416	**2.7 (1.21-6.02)**	**0.015**
Tumour depth	*Deep (ref)*	-	-	-	-	-	-
	Superficial	0.432 (0.133-1.4)	0.162	0.752 (0.265-2.13)	0.591	0.673 (0.268-1.69)	0.4
Tumour margin	*R1 & R2 (ref)*	-	-	-	-	-	-
	R0	0.778 (0.408-1.49)	0.447	1.46 (0.745-2.86)	0.27	0.815 (0.464-1.43)	0.476
	unknown	1.29 (0.387-4.29)	0.68	1.09 (0.252-4.73)	0.906	0.28 (0.038-2.05)	0.21
Log(Tumour size [mm])	4-5 (ref)	-	-	-	-	-	-
	< 4	0.927 (0.352-2.44)	0.877	0.679 (0.267-1.73)	0.417	0.553 (0.237-1.29)	0.171
	> 5	**2.14 (1.08-4.22)**	**0.029**	0.793 (0.384-1.64)	0.532	1.02 (0.57-1.84)	0.938

Supplemental Table 5.8 Multivariable Cox regression for dedifferentiated liposarcoma (DDLPS) and undifferentiated pleomorphic sarcoma (UPS) patients.
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. Anatomical site of 'Other' indicates extremity, trunk wall, and head/neck cases. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval; IA = Intra-abdominal; RP = retroperitoneal

		LRFS		MFS		OS	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
Age at excision (years)		1.03 (0.997-1.06)	0.077	1.02 (0.984-1.06)	0.263	**1.04 (1-1.07)**	**1**
Sex	*M (ref)*	-	-	-	-	-	-
	F	1.11 (0.551-2.23)	0.77	1.63 (0.719-3.68)	0.243	1.22 (0.65-2.29)	0.538
Histological subtype	*UPS (ref)*	-	-	-	-	-	-
	DDLPS	0.747 (0.147-3.8)	0.725	0.595 (0.12-2.96)	0.526	0.568 (0.118-2.73)	0.481
Anatomical site	*Other (ref)*	-	-	-	-	-	-
	IA/RP	**6.42 (1.07-38.7)**	**0.042**	0.248 (0.043-1.44)	0.12	1.4 (0.264-7.45)	0.691
FNCLCC grade	*3 (ref)*	-	-	-	-	-	-
	2	0.836 (0.367-1.9)	0.669	0.411 (0.129-1.31)	0.133	0.518 (0.222-1.21)	0.128
Performance status	*0 (ref)*	-	-	-	-	-	-
	1	1.99 (0.901-4.38)	0.089	1.78 (0.741-4.3)	0.196	**2.57 (1.22-5.42)**	**0.013**
	2-3	1.28 (0.33-4.94)	0.724	0.485 (0.101-2.32)	0.365	2.12 (0.715-6.29)	0.175
	unknown	1.29 (0.491-3.39)	0.606	1.02 (0.299-3.5)	0.971	1.95 (0.743-5.14)	0.174
Tumour depth	*Deep (ref)*	-	-	-	-	-	-
	Superficial	0.35 (0.066-1.84)	0.215	0.314 (0.081-1.22)	0.095	0.556 (0.166-1.86)	0.34
Tumour margin	*R1 & R2 (ref)*	-	-	-	-	-	-
	R0	0.824 (0.383-1.77)	0.62	1.16 (0.53-2.54)	0.708	0.888 (0.446-1.77)	0.735
	unknown	1.46 (0.364-5.88)	0.592	1.22 (0.175-8.43)	0.844	0.424 (0.047-3.81)	0.444
Log(Tumour size [mm])	4-5 (ref)	-	-	-	-	-	-
	< 4	1.17 (0.341-4)	0.805	0.472 (0.159-1.4)	0.177	0.424 (0.152-1.19)	0.102
	> 5	0.821 (0.296-2.28)	0.706	**4.12 (1.29-13.1)**	**0.017**	1.7 (0.721-4.03)	0.224

Supplemental Table 5.9 Statistical associations between dedifferentiated liposarcoma and undifferentiated pleomorphic sarcoma clinicopathological features and tumour infiltrating lymphocyte (TIL) burden.
CD3+/4+/8+ TIL = categorical variable dichotomised at median.

Variable 1	Variable 2	Test performed	Test statistic	Degrees of freedom	p	FDR
CD3	Anatomical site	Chi-squared	1.314	1	0.252	0.412
CD3	Grade	Chi-squared	0.000	1	1.000	1.000
CD3	Histological subtype	Chi-squared	1.281	1	0.258	0.412
CD3	Performance status	Chi-squared	2.977	3	0.395	0.527
CD3	Sex	Chi-squared	4.011	1	0.045	0.181
CD3	Tumour depth	Chi-squared	2.847	1	0.092	0.244
CD3	Tumour margin	Chi-squared	0.054	2	0.974	1.000
CD3	Tumour size	Chi-squared	8.071	2	0.018	0.141
CD4	Anatomical site	Chi-squared	1.710	1	0.191	0.319
CD4	Grade	Chi-squared	0.000	1	1.000	1.000
CD4	Histological subtype	Chi-squared	1.649	1	0.199	0.319
CD4	Performance status	Chi-squared	3.596	3	0.309	0.411
CD4	Sex	Chi-squared	4.461	1	0.035	0.139
CD4	Tumour depth	Chi-squared	2.995	1	0.084	0.223
CD4	Tumour margin	Chi-squared	0.020	2	0.990	1.000
CD4	Tumour size	Chi-squared	9.012	2	0.011	0.088
CD8	Anatomical site	Chi-squared	1.403	1	0.236	0.378
CD8	Grade	Chi-squared	0.005	1	0.942	0.942
CD8	Histological subtype	Chi-squared	1.974	1	0.160	0.320
CD8	Performance status	Chi-squared	2.416	3	0.491	0.654
CD8	Sex	Chi-squared	3.766	1	0.052	0.177
CD8	Tumour depth	Chi-squared	3.370	1	0.066	0.177
CD8	Tumour margin	Chi-squared	0.877	2	0.645	0.737
CD8	Tumour size	Chi-squared	7.428	2	0.024	0.177

Supplemental Table 5.10 Multivariable Cox regression assessing CD4+ tumour infiltrating lymphocyte (TIL) burden in dedifferentiated liposarcoma (DDLPS) and undifferentiated pleomorphic sarcoma (UPS) patients.
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. Anatomical site of 'Other' indicates extremity, trunk wall, and head/neck cases. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval; IA = Intra-abdominal; RP = retroperitoneal

		LRFS		MFS		OS	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
Age at excision (years)		1.03 (0.996-1.07)	0.083	1.03 (0.989-1.08)	0.145	**1.04 (1-1.08)**	**0.037**
Sex	*M (ref)*	-	-	-	-	-	-
	F	0.984 (0.465-2.08)	0.965	1.79 (0.767-4.16)	0.179	1.09 (0.569-2.1)	0.79
Histological subtype	*UPS (ref)*	-	-	-	-	-	-
	DDLPS	0.802 (0.186-3.45)	0.767	0.613 (0.142-2.65)	0.512	0.619 (0.147-2.61)	0.514
Anatomical site	*Other (ref)*	-	-	-	-	-	-
	IA/RP	**6.82 (1.32-35.1)**	**0.022**	0.295 (0.059-1.47)	0.136	1.67 (0.365-7.63)	0.509
FNCLCC grade	*3 (ref)*	-	-	-	-	-	-
	2	0.867 (0.36-2.09)	0.751	0.462 (0.146-1.46)	0.19	0.525 (0.209-1.32)	0.172
Performance status	*0 (ref)*	-	-	-	-	-	-
	1	**2.59 (1.02-6.55)**	**0.045**	2.09 (0.797-5.46)	0.134	**3.2 (1.4-7.29)**	**0.006**
	2-3	1.34 (0.335-5.39)	0.677	0.394 (0.078-1.98)	0.259	2.32 (0.733-7.37)	0.152
	unknown	1.16 (0.402-3.36)	0.782	0.68 (0.176-2.62)	0.576	1.72 (0.583-5.09)	0.325
Tumour depth	*Deep (ref)*	-	-	-	-	-	-
	Superficial	0.529 (0.096-2.91)	0.464	0.313 (0.077-1.28)	0.106	0.684 (0.199-2.35)	0.546
Tumour margin	*R1 & R2 (ref)*	-	-	-	-	-	-
	R0	0.866 (0.376-1.99)	0.735	1.05 (0.464-2.39)	0.899	1 (0.473-2.12)	0.998
	unknown	1.37 (0.333-5.61)	0.664	0.934 (0.141-6.17)	0.943	0.38 (0.043-3.39)	0.386
Log(Tumour size [mm])	4-5 (ref)	-	-	-	-	-	-
	< 4	0.812 (0.212-3.11)	0.761	0.352 (0.108-1.15)	0.083	0.337 (0.111-1.03)	0.055
	> 5	0.931 (0.316-2.74)	0.897	5.13 (1.5-17.5)	0.009	1.86 (0.759-4.55)	0.175
CD4	*low (ref)*	-	-	-	-	-	-
	high	0.798 (0.362-1.76)	0.575	1.04 (0.419-2.58)	0.933	0.838 (0.401-1.75)	0.638

Supplemental Table 5.11 Multivariable Cox regression assessing CD8+ tumour infiltrating lymphocyte (TIL) burden in dedifferentiated liposarcoma (DDLPS) and undifferentiated pleomorphic sarcoma (UPS) patients.
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. Anatomical site of 'Other' indicates extremity, trunk wall, and head/neck cases. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval; IA = Intra-abdominal; RP = retroperitoneal

		LRFS		MFS		OS	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
Age at excision (years)		1.02 (0.985-1.06)	0.258	1.03 (0.982-1.07)	0.255	1.03 (0.993-1.07)	0.114
Sex	*M (ref)*	-	-	-	-	-	-
	F	1.07 (0.499-2.28)	0.869	1.95 (0.804-4.74)	0.14	1.15 (0.596-2.2)	0.682
Histological subtype	*UPS (ref)*	-	-	-	-	-	-
	DDLPS	1.65 (0.271-9.99)	0.588	0.661 (0.108-4.05)	0.654	0.809 (0.148-4.41)	0.806
Anatomical site	*Other (ref)*	-	-	-	-	-	-
	IA/RP	3.16 (0.497-20.1)	0.222	0.266 (0.042-1.68)	0.159	1.22 (0.217-6.82)	0.823
FNCLCC grade	*3 (ref)*	-	-	-	-	-	-
	2	0.55 (0.208-1.46)	0.23	0.284 (0.059-1.36)	0.115	0.353 (0.12-1.04)	0.058
Performance status	*0 (ref)*	-	-	-	-	-	-
	1	**2.64 (1.09-6.43)**	**0.032**	1.74 (0.661-4.6)	0.262	**3.06 (1.3-7.19)**	**0.01**
	2-3	1.61 (0.406-6.39)	0.497	0.551 (0.117-2.59)	0.451	2.9 (0.902-9.34)	0.074
	unknown	1.3 (0.411-4.1)	0.656	0.677 (0.17-2.69)	0.579	1.97 (0.627-6.21)	0.245
Tumour depth	*Deep (ref)*	-	-	-	-	-	-
	Superficial	0.49 (0.09-2.68)	0.41	0.328 (0.078-1.38)	0.129	0.705 (0.203-2.45)	0.582
Tumour margin	*R1 & R2 (ref)*	-	-	-	-	-	-
	R0	0.885 (0.38-2.06)	0.777	0.949 (0.406-2.22)	0.905	0.982 (0.457-2.11)	0.964
	unknown	1.9 (0.413-8.78)	0.409	1.53 (0.126-18.7)	0.737	0.779 (0.082-7.42)	0.828
Log(Tumour size [mm])	4-5 (ref)	-	-	-	-	-	-
	< 4	0.942 (0.24-3.71)	0.932	0.423 (0.124-1.45)	0.171	0.382 (0.122-1.19)	0.098
	> 5	1.09 (0.392-3.04)	0.868	**5.27 (1.59-17.5)**	**0.006**	1.97 (0.788-4.91)	0.147
CD8	*low (ref)*	-	-	-	-	-	-
	high	0.61 (0.297-1.25)	0.178	0.629 (0.268-1.47)	0.286	0.625 (0.322-1.21)	0.164

Chapter 6 Unbiased characterisation of the pan-STS proteome

6.1 Background and objectives

So far, this thesis has provided an overview of the STS proteome (**Chapter 4**), investigated intra-subtype heterogeneity and immune heterogeneity within select histological subtypes (**Chapter 5**). To do this, publicly available gene sets and protein databases have been utilised alongside the complete proteomic data. These analyses have focused on a histological subtype point of view and utilised the current understanding of STS biology to direct investigations. However, the STS proteome is yet to be defined from an unbiased protein-centric perspective.

In this Chapter, network analysis is used to modularly defined the STS proteome as groups of co-expressed proteins. STS share a common mesenchymal origin[3]. It therefore follows that shared mesenchymal features may exist across histological subtypes. As such, these groups of proteins are hypothesised to be co-functioning and map to key biological activities underlying multiple subtypes of STS (i.e., 'pan-subtype'). By extension, these common STS features may correspond to clinical differences across the STS population, and therefore the modular proteome was assessed for its relationship with clinicopathological features and patient outcomes.

Accordingly, the objectives of this chapter are:

1) To define the unbiased STS proteome network composition and structure
2) To determine whether the STS proteome can provide clinical utility that is complementary to histological subtype information

6.1.1 Results

6.1.2 Weighted gene correlation network analysis of the proteomic dataset

To characterise the pan-subtype STS proteomic network of the MS cohort, weighted gene correlation network analysis (WGCNA) was applied to the normalised expression values of 3,290 proteins across all profiled samples (n = 321). WGCNA utilises correlation networks to identify clusters (named 'modules') of highly correlated genes/proteins[521]. This method offers superiority over hierarchical clustering by providing improved sensitivity to correlations between proteins and considering inversely correlated patterns across the cohort which would not be identified otherwise. For clarity,

all references to the WGCNA method herein will refer to proteins, however WGCNA was developed on gene expression data, and thus WGCNA literature refers predominantly to genes. Biological networks comprise nodes (e.g., proteins) and edges connecting nodes (e.g., indicating an 'interaction'). Networks can be described by many different measures. The measures primarily used by WGCNA include node degree, the number of edges connected to each protein, and topological overlap, a similarity measure based on how many neighbours are shared between protein pairs (a high topological overlap value indicates many common neighbours)[521,655]. The WGCNA method assumes a scale-free network is present in the data. A scale-free network has 'hub-and-spoke' architecture preserved throughout its structure (**Figure 6.1A**)[656]. This means that many nodes have low degrees, and few nodes ('hubs') have high degrees. This scale-free degree distribution is known as power law distribution; a continuous positive distribution where the degree distribution decreases as a power of its magnitude[657]. To contrast this, other common networks include random networks that show no structure or hierarchical pattern and small world networks which show high local clustering (**Figure 6.1B-C**)[658,659]. Both random and small world networks have a unimodal and approximate normal degree distribution. Scale-free networks are typical of protein interactions in biological systems, where hubs correspond to key proteins with wide ranging roles an high influence[656]. To identify WGCNA modules within the data, the scale free topology model fit was optimised to ensure the final network built had high similarity

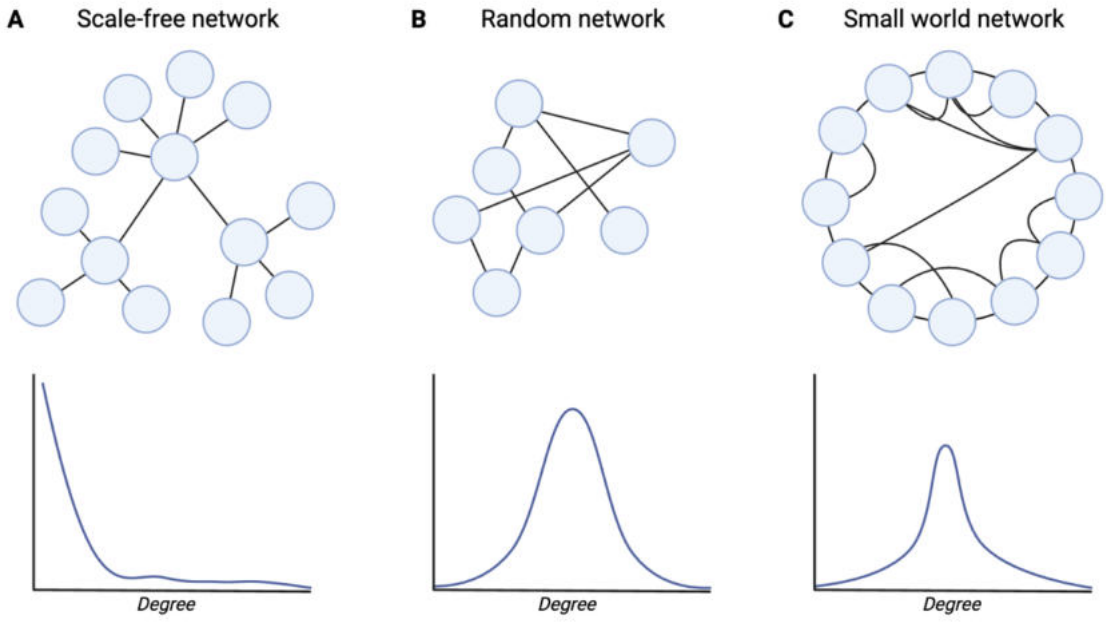


Figure 6.1 Types of networks
Diagrammatic representation of a scale-free network **(A)**, random network **(B)**, and small world network **(C)**, with approximated degree distributions plotted below.

to a 'perfect' scale-free network fit. This was achieved by raising gene correlations to a certain power to reduce background correlation noise and amplify the stronger correlations. The optimal power value can be defined as the point beyond which only minimal improvements in model fit are seen. As model fit can be measured as an R^2 value, the optimal value can also be selected based on a certain R^2 threshold.

Herein, when the WGCNA method was applied, an optimal power value of 5 was identified; close to the inflection point on the scale independence plot and where $R^2 > 0.9$ (**Figure 6.2A**). A value of 5 was also shown as appropriate based on mean connectivity. Considering the degree (i.e., connectivity) of a scale free network, mean connectivity is expected to be low (**Figure 6.2A**). The correlations were therefore raised

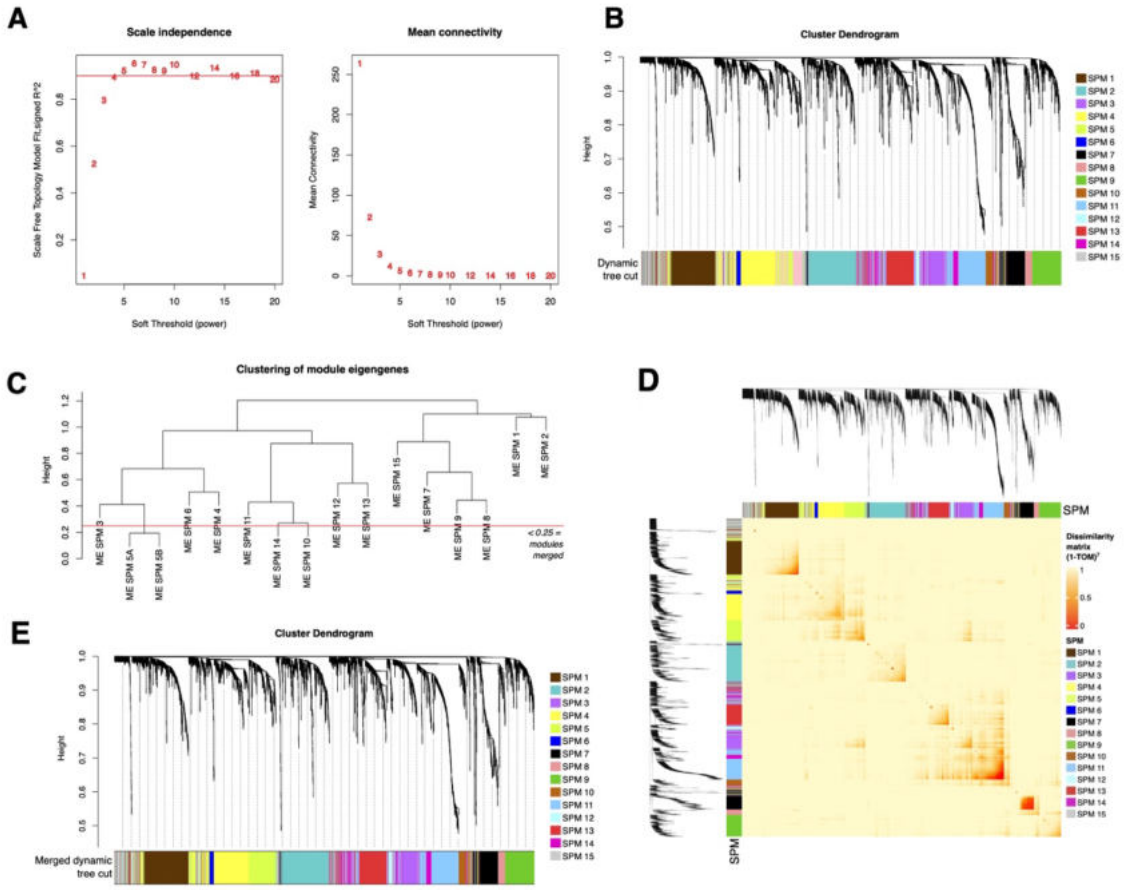


Figure 6.2 Weighted gene correlation network analysis (WGCNA) for the identification of sarcoma proteome modules (SPM)
(A) Scale free topology model fit and mean connectivity of model at Soft Threshold (power) values up to 20. Red line indicates R^2 of 0.9. **(B)** Cluster dendrogram of all proteins where height indicates 1-Pearson's correlation. SPM identification and protein assignments by dynamic tree cut annotated in colour. **(C)** Dendrogram of SPM module eigengenes (ME) where height indicates 1-Pearson's correlation. Modules with height < 0.25 (red line) were merged. **(D)** Co-expression heatmap showing correlation of protein expression based on Topological Overlap Matrix (TOM) dissimilarity ($(1\text{-TOM})^7$). Cluster dendrogram height indicates 1-Pearson's correlation**. (E)** Cluster dendrogram of all proteins where height indicates 1-Pearson's correlation. SPM identification and protein assignments by merged dynamic tree cut annotated in colour.

to the power of 5 to create an adjacency matrix. A topology overlap matrix (TOM) was then generated from the adjacency matrix, and modules identified by the dynamic tree cut algorithm applied to the dissimilarity TOM (further details in **section X**; **Figure 6.2B**). In this dataset, WGCNA and the use of a power value of 5, identified 16 modules which we termed as sarcoma proteome modules (SPM), 1 of which comprised proteins which could not be assigned to a highly correlated SPM. To assess the robustness of these SPMs as individual networks, the SPM eigengene values were hierarchically clustered (**Figure 6.2D**). This revealed 2 SPMs to be of particularly high similarity, diverging on the dendrogram at a height < 0.25, where height indicates 1 minus the Pearson correlation (1 – r). These SPMs were therefore merged, resulting in 14 robust SPMs and 1 SPM comprising the remaining proteins (named 'SPM 15'; **Figure 6.2D-E**).

6.1.3 Biological characterisation of the SPMs

Across the 15 SPMs, 3,290 proteins with 10,820,810 interactions were identified. Interactions between protein pairs are quantified as the WGCNA 'edge weight' or 'co-expression score'. These values are derived from the adjacency matrix and can be interpreted much like a correlation measure: where a high value indicates a stronger correlation. For interpretation purposes it is important to remember that a power of 5 was used to scale the data (**section 6.1.2**). Therefore, a co-expression score of 0.5 is effectively equal to a correlation of 0.87 ($0.87^5 = 0.5$). SPMs comprised between 41 and 420 proteins, with a median within-SPM co-expression score of 0.025, and a median between-SPM co-expression score of 0.002. Many interactions were weak. Therefore, to reduce noise within the network for visualisation purposes, the co-expression score between protein pairs was restricted to ≥ 0.05. This left 3,290 proteins and 168,574 interactions, 32% of which were present in the STRINGdb[602,603]. To visualise this representative STS proteome, a protein co-expression network was constructed, revealing evident SPM-based structure (**Figure 6.3**).

To identify SPM-specific biology, the proteins present in each SPM were assessed by over-representation analysis using the GO BP and Hallmarks of MSigDB[506–508,512]. This found no significant enrichment of any gene set in any SPM. Each SPM was therefore manually inspected as a PPI network, and each was revealed to comprise groups of proteins describing specific functional biology. Briefly, SPM 1 comprised 389 proteins with a median co-expression weight of 0.044 and most weights ranging from 0.01 - 0.04 (**Supplemental Figure 6.1A-B**). Of all interactions measured by WGCNA, ~ 6% were present in the STRINGdb (**Supplemental Figure 6.1C**). The STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ >

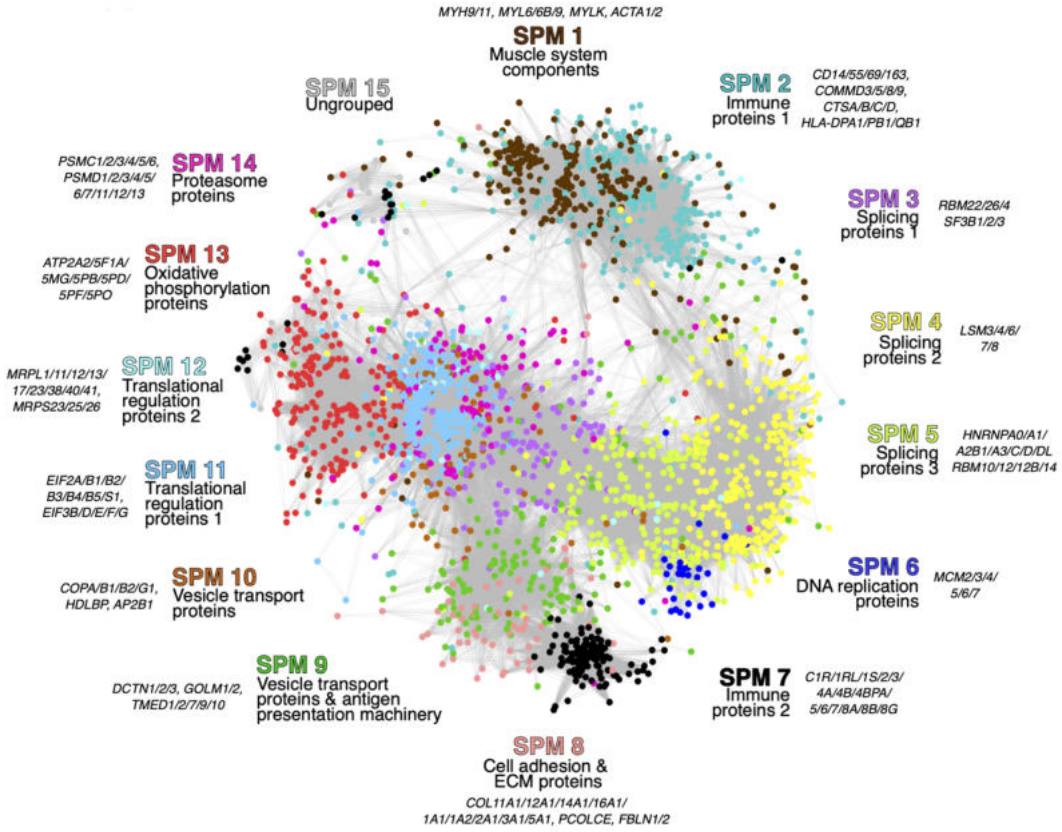


Figure 6.3 The STS proteome network defined as sarcoma proteome modules (SPM)
Protein co-expression network comprising 3290 nodes and 168,574 edges. Nodes indicate proteins and are coloured based on SPM membership. Edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. Representative biological features and selected proteins are annotated for each SPM.

0.26). Key proteins present within SPM 1 included those related to muscle activity, heat shock proteins, and ECM components. (**Supplemental Figure 6.1D**). SPM 2 comprised 420 proteins with a median co-expression weight of 0.035 (**Supplemental Figure 6.2A-B**). Of all interactions measured by WGCNA, ~ 9% were present in the STRINGdb (**Supplemental Figure 6.2C**). As before, the STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ > 0.3). Key proteins present within SPM 2 included those with immune roles (cathepsins, human leukocyte antigens (HLA), and S100 proteins) and cell surface markers (**Supplemental Figure 6.2D**). SPM 3 comprised 141 proteins with a median co-expression weight of 0.046 (**Supplemental Figure 6.3A-B**). Of all interactions measured by WGCNA, ~ 17% were present in the STRINGdb (**Supplemental Figure 6.3C**), and the STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ > 0.15). SPM 3 contained mostly splicing proteins (**Supplemental Figure *6.3*D**). SPM 4 comprised 356 proteins and had a median co-expression weight of 0.035

(**Supplemental Figure *6.4*A-B**). Of all interactions measured by WGCNA, ~ 8% were present in the STRINGdb (**Supplemental Figure *6.4*C**). The STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ > 0.2). Key proteins present within SPM 4 included those involved in splicing and ubiquitination (**Supplemental Figure 6.4D**). SPM 5 comprised 314 proteins with a median co-expression weight of 0.039 (**Supplemental Figure 6.5A-B**). Of all interactions measured by WGCNA, ~ 21% were present in the STRINGdb (**Supplemental Figure 6.5C**). The STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ > 0.15). Key proteins present within SPM 5 included vesicle trafficking proteins: Rab proteins, Sec machinery, and TMED proteins (**Supplemental Figure 6.5D**). SPM 6 comprised 41 proteins with a median co-expression weight of 0.047, and a weight distribution showing 1 peak at 0.03 and 2 small peaks at higher values (0.13 and 0.35; **Supplemental Figure 6.6A-B**). Of all interactions measured by WGCNA, a relatively high proportion (~ 38%) were present in the STRINGdb (**Supplemental Figure 6.6C**). The STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ > 0.3). Key proteins present within SPM 6 included those in DNA replication (**Supplemental Figure 6.6D**). SPM 7 comprised 185 proteins with a median co-expression weight of 0.167, and most weights ranging from 0.03 – 0.3 (**Supplemental Figure 6.7A-B**). Of all interactions measured by WGCNA, ~ 21% were present in the STRINGdb (**Supplemental Figure *6.7*C**). The STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ > 0.35). Key proteins present within SPM 7 included those involved in the immune response: immunoglobulins and complement components. (**Supplemental Figure 6.7D**). SPM 8 comprised 66 proteins with a median co-expression weight of 0.042 (**Supplemental Figure 6.8A-B**). Of all interactions measured by WGCNA, ~ 41% were present in the STRINGdb (**Supplemental Figure 6.8C**). The STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ > 0.15). Key proteins present within SPM 8 included ECM components, most notably several of collagen chains (**Supplemental Figure 6.8D**). SPM 9 comprised 254 proteins with a median co-expression weight of 0.036 (**Supplemental Figure 6.9A-B**). Of all interactions measured by WGCNA, ~ 8% were present in the STRINGdb (**Supplemental Figure 6.9C**). The STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ > 0.15). Key proteins present within SPM 9 included splicing proteins, heterogenous nuclear ribonucleoproteins (HNRNP) and histone proteins (**Supplemental Figure 6.9D**). SPM 10 comprised 94 proteins with a median co-expression weight of 0.041 (**Supplemental Figure *6.10*A-B**). Of all interactions measured by WGCNA, ~ 11% were

present in the STRINGdb (**Supplemental Figure *6.10*C**). The STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ > 0.2). Proteins of SPM 10 primarily harboured vesicle trafficking roles, such as the adaptor and COPI coat proteins (**Supplemental Figure *6.10*D**). SPM 11 comprised 409 proteins with a median co-expression weight of 0.065, and most weights between 0.02 - 0.03 (**Supplemental Figure *6.11*A-B**). Of all interactions measured by WGCNA, ~ 25% were present in the STRINGdb (**Supplemental Figure *6.11*C**). The STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ > 0.45). Key proteins present within SPM 11 included translation initiation factors, ribosome, and proteasome components (**Supplemental Figure *6.11*D**). SPM 12 comprised 44 proteins with a median co-expression weight of 0.052, most weights between 0.02 – 0.06, and a relatively high proportion of high co-expression weights (> 0.1; **Supplemental Figure *6.12*A-B**). Of all interactions measured by WGCNA, ~ 60% were present in the STRINGdb (the highest of any SPM; **Supplemental Figure *6.12*C**). Key proteins present within SPM 12 were primarily those of the mitochondrial ribosomes (mitoribosomes; **Supplemental Figure *6.12*D**). SPM 13 comprised 239 proteins with a median co-expression weight of 0.042 (range = 0.025 – 0.05; **Supplemental Figure *6.13*A-B**). Of all interactions measured by WGCNA, ~ 28% were present in the STRINGdb (**Supplemental Figure *6.13*C**). The STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ > 0.2). Key proteins present within SPM 13 included those involved in oxidative phosphorylation such as the NDUF proteins which form complex I and II of the electron transport chain, as well as mitochondrial ATP synthase subunits (**Supplemental Figure *6.13*D**). SPM 14 comprised 177 proteins with a median co-expression weight of 0.036 (**Supplemental Figure *6.14*A-B**). Of all interactions measured by WGCNA, ~ 20% were present in the STRINGdb (**Supplemental Figure *6.14*C**). The STRINGdb and WGCNA overlap was highest for WGCNA interactions with a higher co-expression weight (~ > 0.2). Key proteins present within SPM 14 included the cullin molecular scaffold proteins and proteasome subunits (**Supplemental Figure *6.14*D**). Overall, unbiased analysis of the STS proteome has identified SPMs which span a range of key functional biological activities.

6.1.4 Clinical characterisation of the SPMs

Next, we sought to investigate the clinical feature inherent to different SPMs. To achieve this, the associations between SPMs and clinicopathological variables were interrogated. The median SPM expression was used to summarise each SPM for each patient. This

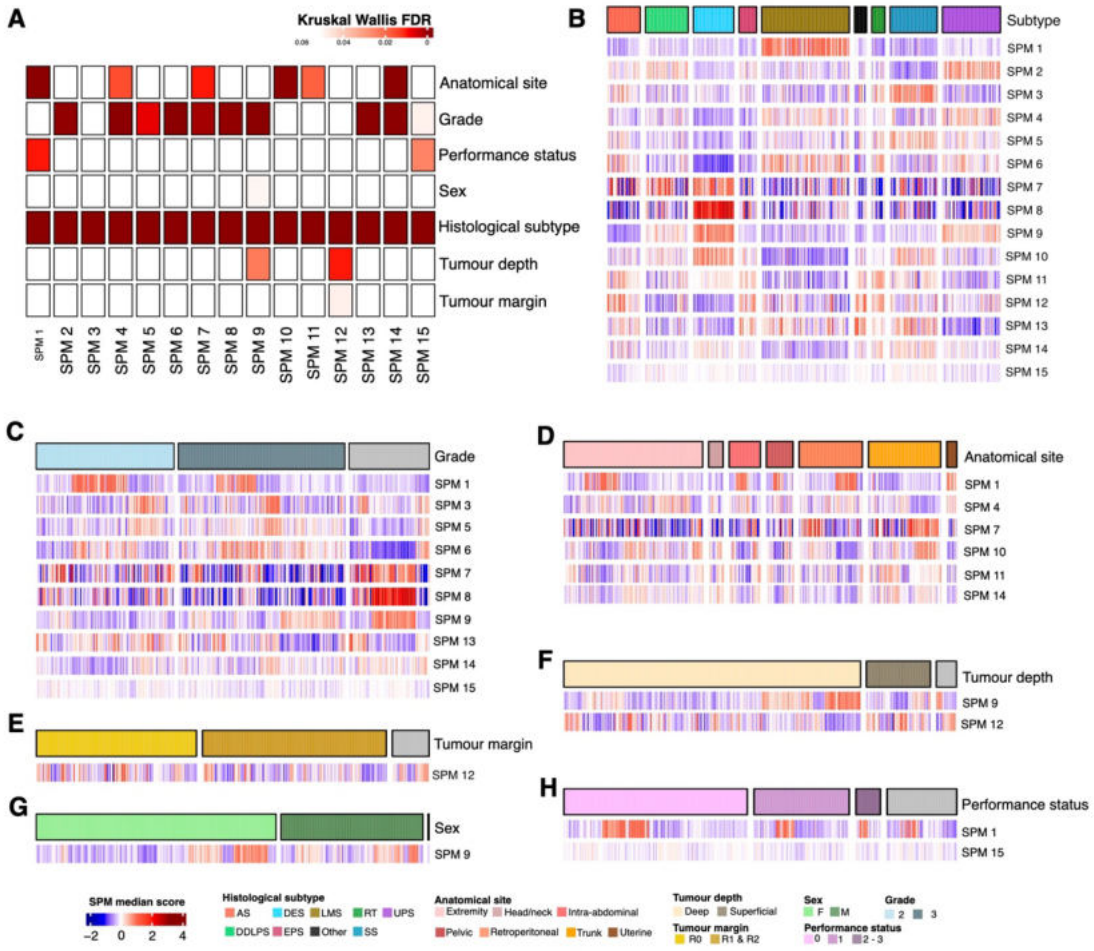


Figure 6.4 Associations between sarcoma proteome modules (SPMs) and clinicopathological variables
(A) Overview of Kruskal Wallis tests assessing the statistical association between SPMs and clinicopathological variables. Colour indicates false discovery rate (FDR). **(B-H)** Supervised heatmaps showing the SPM median expression score for each case. SPMs included where significantly associated with variables: histological subtype **(B)**, grade **(C)**, anatomical site **(D)**, tumour margin **(E)**, tumour depth **(F)**, sex **(G)**, performance status **(H)**.

revealed a significant association between all SPMs and histological subtype (**Figure 6.4A** and **Supplemental Table 6.1**). Given the dominance of histology in proteomic features revealed throughout this thesis, this is unsurprising. Specifically, SPM 1 and to a lesser extent SPM 6 were enriched in LMS compared to all other subtypes. SPM 2 and SPM 4 were enriched in UPS, and SPM 3 was enriched in SS and RT (**Figure 6.4B**). SPMs 7, 8, 9, and 10 were enriched in DES, with SPM 9 also enriched in UPS, and SPM 10 also enriched in SS. Most SPMs (SPM 1, SPM 3, SPM 5, SPM 6, SPM 7, SPM 8, SPM 9, SPM 13, SPM 14, and SPM 15) were also significantly associated with grade (**Figure 6.4A** and **Supplemental Table 6.1**). However, this was driven by tumours where grading information was not available or applicable; high SPM 7, SPM 8, and SPM 9

expression was seen in these tumours (**Figure 6.4C**). Beyond this, SPM 5, SPM 6, and SPM 9 showed higher expression in grade 3 than grade 2 tumours. Additionally, SPM 1, SPM 4, SPM 7, SPM 10, SPM 11, and SPM 14 were significantly associated with anatomical site (**Figure 6.4A** and **Supplemental Table 6.1**). Whilst the anatomical site heatmap association showed extensive heterogeneity, a significant enrichment of SPM 1 was consistently observed in uterine tumours, and in subsets of intra-abdominal, extremity, and retroperitoneal tumours (**Figure 6.4D**). Subsets of trunk wall and retroperitoneal tumours showed high SPM 7 expression. Other significant associations included SPM 12 with tumour margin, SPM 9 and SPM 12 tumour depth, SPM 9 with sex, and SPM 1 and SPM 15 with performance status (**Figure 6.4A** and **Supplemental Table 6.1**). Yet, inspection of these heatmaps revealed high heterogeneity in expression levels across groups of clinicopathological variables (**Figure 6.4E-H**).

Further to assessing SPM associations with clinicopathological variables, we also interrogated the relationship between SPM and patient outcome. Due to the clinical differences between DES and RT, and the typical adult STS population, patients of these diagnoses were excluded from survival analysis[557,660]. Median SPM scores were assessed by univariable Cox regression (summarised in **Figure *6.5*** and detailed in **Supplemental Table *6.2***). This illustrated high SPM 1 expression as significantly associated with a superior LRFS (HR = 0.554, 95% CI = 0.372-0.825, FDR = 0.041), high SPM 6 as significantly associated with a poorer MFS (HR = 2.19, 95% CI = 1.52-3.15, FDR = 0.001), and high SPM 15 as associated with a superior MFS (HR = 0.092, 95% CI = 0.026-0.321, FDR = 0.003) and OS (HR = 0.087, 95% CI = 0.025-0.302, FDR = 0.003). Given several relationships between SPMs and clinicopathological variables were identified, survival analyses were reperformed using the multivariable Cox regression to adjust for such variables. To avoid inflating the type I error rate, only SPMs where a significant univariable relationship with outcome was seen were assessed. However, a more lenient significance cut off was used for selection (FDR < 0.1). This led to the inclusion of an SPM 10 MFS model (univariable: HR = 0.563, 95% CI = 0.367-0.863, FDR = 0.063), SPM 4 OS model (univariable: HR = 1.85, 95% CI = 1.16-2.95, FDR = 0.065), and SPM 13 OS model (univariable: HR = 0.61, 95% CI = 0.427-0.871, FDR = 0.059) in analyses (**Figure 6.5** and **Supplemental Table 6.2**). Following multivariable adjustment, SPM 6 remained associated with a significantly superior MFS (HR = 1.96, 95% CI = 1.19-3.25, FDR = 0.009; **Supplemental Table 6.3**), and SPM 10 remained associated with a significantly superior MFS (HR = 0.466, 95% CI = 0.247-0.879, FDR = 0.018; **Supplemental Table *6.4***). All PH assumptions were met within these models.

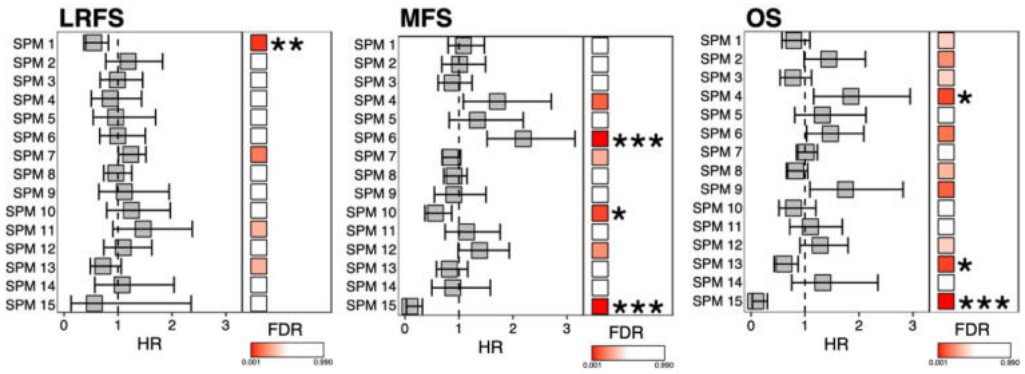


Figure 6.5 Associations between sarcoma proteome modules (SPMs) and clinical outcome
Forest plots for local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) illustrating hazard ratio (HR), 95% confidence intervals (as bars), and false discovery rate (FDR). Significance indicated by * = $p < 0.1$, ** = $p < 0.05$, *** = $p < 0.001$

6.1.5 SPM 6

SPM 6 was found to have a significant and independent prognostic value for outcome. Namely, a high expression of SPM 6 associated with a poorer MFS. To better understand this relationship, SPM 6 was analysed further.

SPM 6 was one of the smallest SPMs identified, comprised of 41 proteins, which mostly possess roles in DNA replication. Leveraging on the network-basis of SPMs, the influence each protein had within the SPM 6 network was quantified. For each protein, the eigengene-based connectivity was calculated as a correlation between each protein expression profile and the SPM eigengene. Additionally, network measures of degree, closeness centrality, and betweenness centrality were extracted[661,662]. The closeness centrality describes the distance from each node to other nodes, where a high value indicates shorter distances and thus a more central node. Betweenness centrality describes the proportion of shortest paths between all pairs of nodes that pass through a specific node, where a high value indicates involvement in many shortest path and thus a more central node. Visualisation of these measures together highlighted 6 proteins to consistently have the highest values of all measures (**Figure 6.6**). These included 5 MCM complex components (MCM2/3/4/6/7) and FEN1, an endonuclease involved in base excision DNA repair[663,664].

SPM 6, as with all other SPMs showed a significant association with histology; with higher expression of SPM 6 seen in LMS and AS, and lower expression seen in DES compared to all other subtypes. SPM 6 was also significantly associated with grade, which was driven by low expression in tumours where grading was not available or

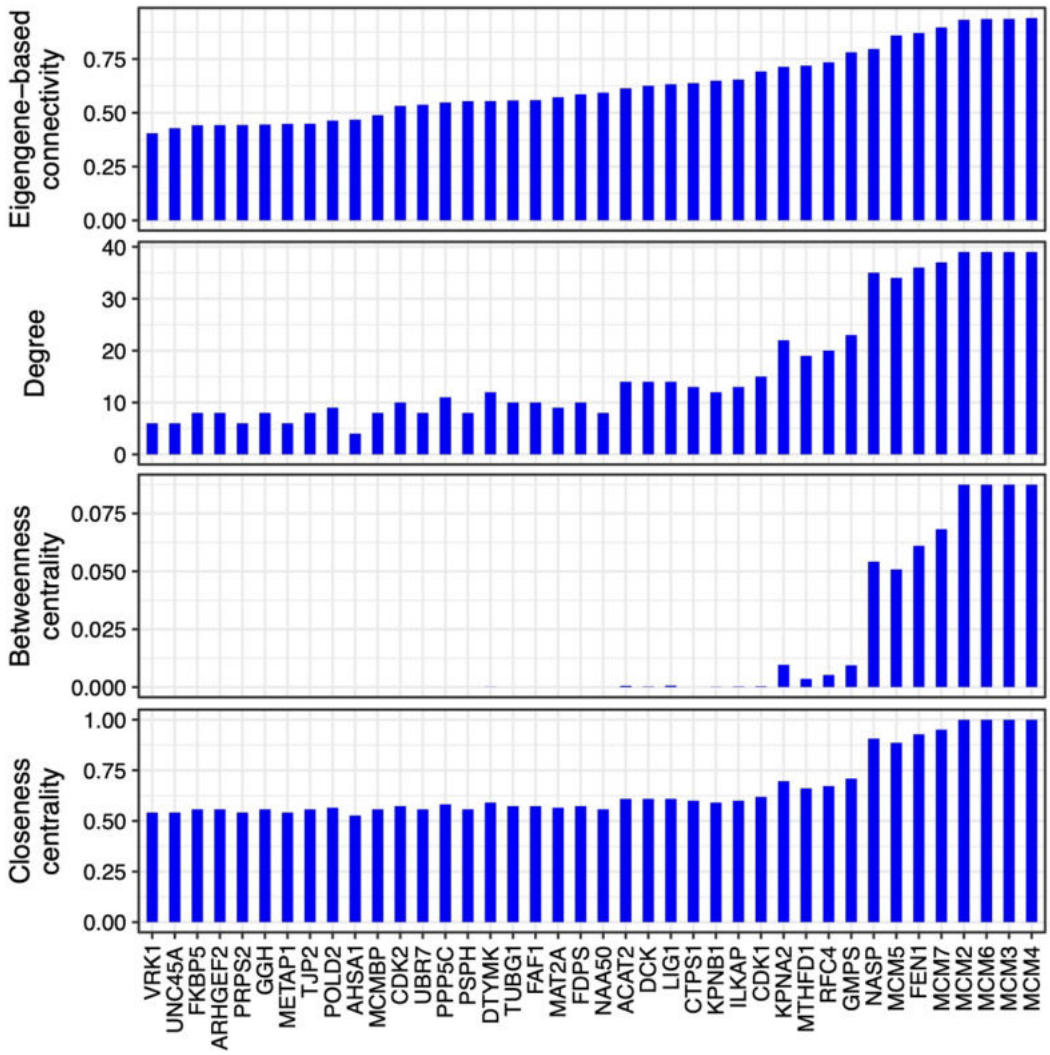


Figure 6.6 Network analysis of sarcoma proteome module 6
From top to bottom, plot shows eigengene-based connectivity, degree, betweenness centrality, and closeness centrality for all proteins within sarcoma proteome module 6 (SPM 6).

applicable. However, despite these associations, stratification of the SPM 6 median scores into tertiles showed representation of all histologies and grades within low, intermediate, and high expression groups (**Figure 6.7A-B**). When split into tertiles, SPM 6 also showed a significant association with MFS in univariable and multivariable analyses, reiterating the previously highlighted prognostic value of this set of proteins (**Figure 6.7C, Table 6.1** and **Supplemental Table 6.5**). No statistical association was revealed between SPM 6 expression and LRFS or OS (**Supplemental Table 6.5**), although a minor trend in OS was observed on the Kaplan Meier plot (**Figure 6.7C**). This was in agreement with MFS observations, suggesting low SPM 6 expression may be associated with a superior OS. The PH assumption was met for all variables of all models except histological subtype in multivariable MFS analyses (Schoenfeld p = 0.02).

However, inspection of the Schoenfeld residuals illustrated no severe violation of the model assumption, and therefore interpretation was valid (**Supplemental Figure 6.15**).

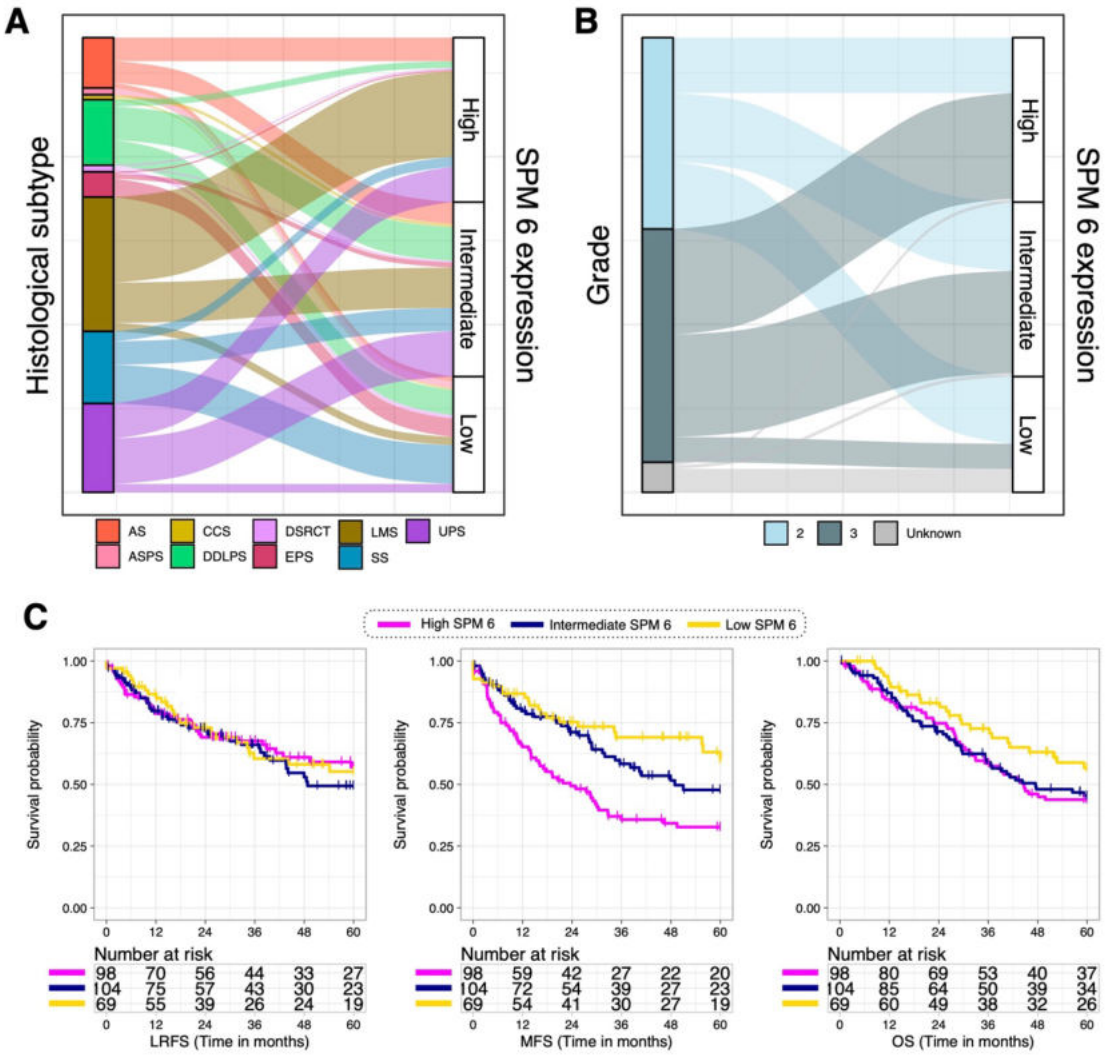


Figure 6.7 Clinical characterisation of sarcoma proteome module 6
(A-B) Alluvial plots illustrating the distribution of **(A)** histological subtype (excluding desmoid tumours (DES) and rhabdoid tumours (RT)) and **(B)** grade across three SPM 6 subgroups. Subgroups were identified by tertile stratification based on median SPM 6 expression across the full cohort. **(C)** Kaplan Meier plots of local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) across the three SPM 6 subgroups. Corresponding univariable Cox regression detailed in **Table 6.1**. Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma.

Table 6.1 Univariable Cox regression for sarcoma proteome module (SPM) 6
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. SPM subgroups identified by tertile stratification based on median expression across the full cohort. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval

		LRFS		MFS		OS	
		HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
SPM6	*low (ref)*	-	-	-	-	-	-
	intermediate	1.15 (0.701-1.9)	0.573	1.4 (0.838-2.35)	0.197	1.48 (0.915-2.39)	0.11
	high	1 (0.6-1.68)	0.992	**2.42 (1.48-3.95)**	**<0.001**	1.52 (0.946-2.46)	0.04

6.1.6 SPM 10

SPM 10 was also found to have a significant and independent prognostic value for MFS. However, in contrast to SPM 6, higher SPM 10 expression was associated with a superior MFS. To better understand this relationship, SPM 10 was analysed further.

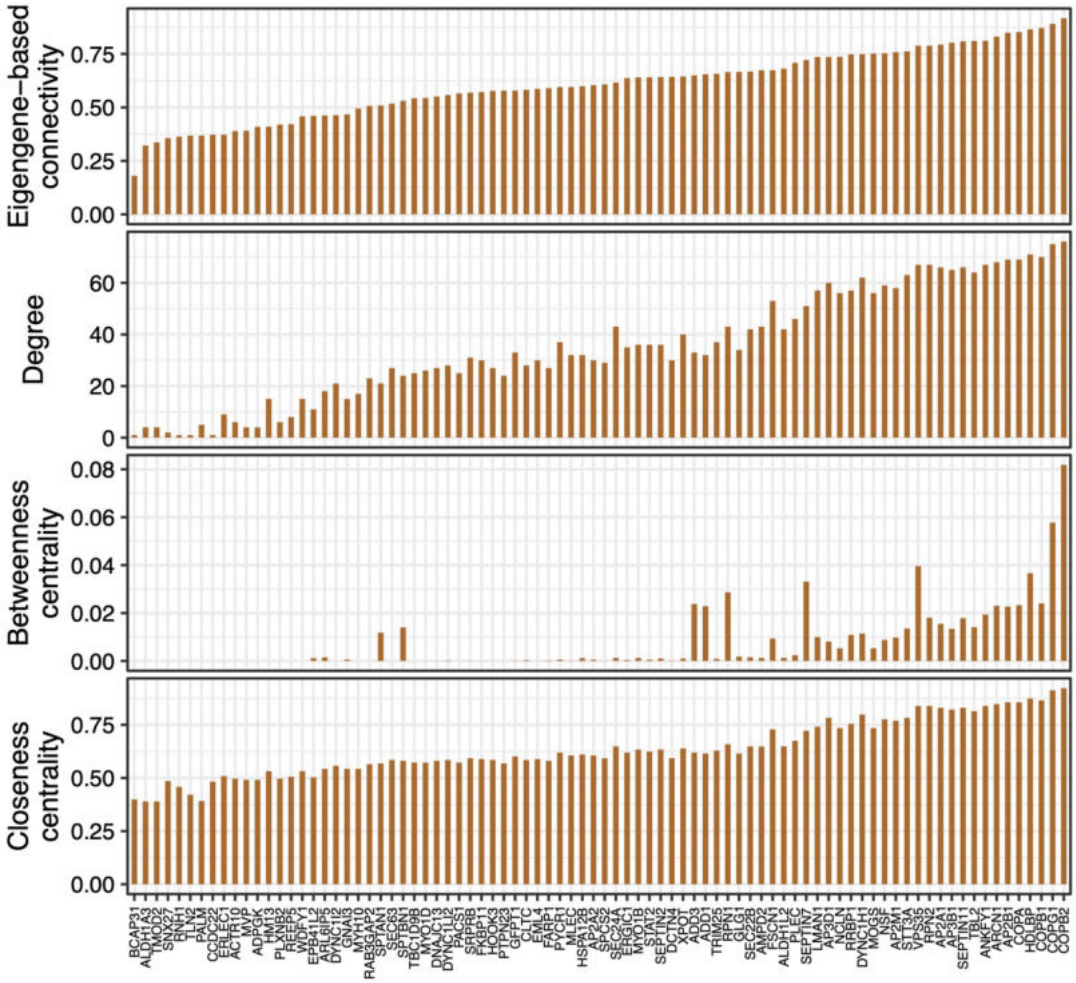


Figure 6.8 Network analysis of sarcoma proteome module 10
From top to bottom, plot shows eigengene-based connectivity, degree, betweenness centrality, and closeness centrality for all proteins within sarcoma proteome module 10 (SPM 10).

SPM 10 contained 94 proteins comprised primarily of vesicle transport proteins. Analysis of the SPM 10 network characteristics revealed 2 proteins to consistently have the highest influence (COPB2/G1; **Figure 6.8**). Other highly influential proteins illustrated by the network measures included COPB1, COPA, AP2B1, and ARCN1. COPB2/G1/A/B1 are COPI coat complex subunit which coat vesicles budding from the Golgi complex[665]. ARCN1 is a coatomer protein, and AP2B1 links clathrin to coated vesicles[666].

SPM10 was shown to be associated with histological subtype and anatomical site. However, despite this, when cases were stratified by tertiles, all histological subtypes in the cohort were represented in each SPM 10 expression group (**Figure 6.9A**). Additionally, all anatomical sites except uterine were present in each of the three SPM 10 expression group (**Figure 6.9B**). Uterine tumours were split between the low and intermediate SPM 10 groups. As for SPM 6, when split in to tertiles SPM 10 retained prognostic significance for MFS. High SPM 10 expression was significantly associated with a superior MFS in both univariable and multivariable analyses (**Figure 6.9C, Table 6.2** and **Supplemental Table 6.6**). Additionally, whilst SPM 10 showed no association with OS in univariable analysis, significance was seen in multivariable analysis (**Supplemental Table 6.6**). Specifically, as for MFS, a high SPM 10 expression was associated with a significantly superior OS (HR = 0.432, 95% CI = 0.238 – 0.782, p = 0.006). No statistical association was revealed between SPM 10 expression and LRFS (**Supplemental Table *6.6***). The PH assumption of these models was met for all variables.

6.1.7 Validation of the prognostic SPMs

Next, we assessed whether the prognostic findings revealed for SPM 6 and SPM 10 were reproducible in an independent dataset. There is no other publicly available MS STS dataset corresponding to an independent, multi-subtype cohort. Therefore, the TCGA STS RNAseq data was explored[36]. This data is derived from an independent patient cohort; however, it corresponds to gene expression measures as opposed to MS data. This introduces challenges as to the interpretation of any results; differentiating between the impact of different cohorts, and different methods is not possible. Nevertheless, the expression of genes in the TCGA dataset reflecting the proteins found in SPM 6 and SPM 10 was assessed. To facilitate appropriate comparisons, TCGA data for only those subtypes present in the proteomic data was assessed (LMS, DDLPS, UPS, and SS). Additionally, TCGA outcome data was censored at 5 years post-surgery, as was the case for the MS cohort. In line with associations revealed in the proteomic

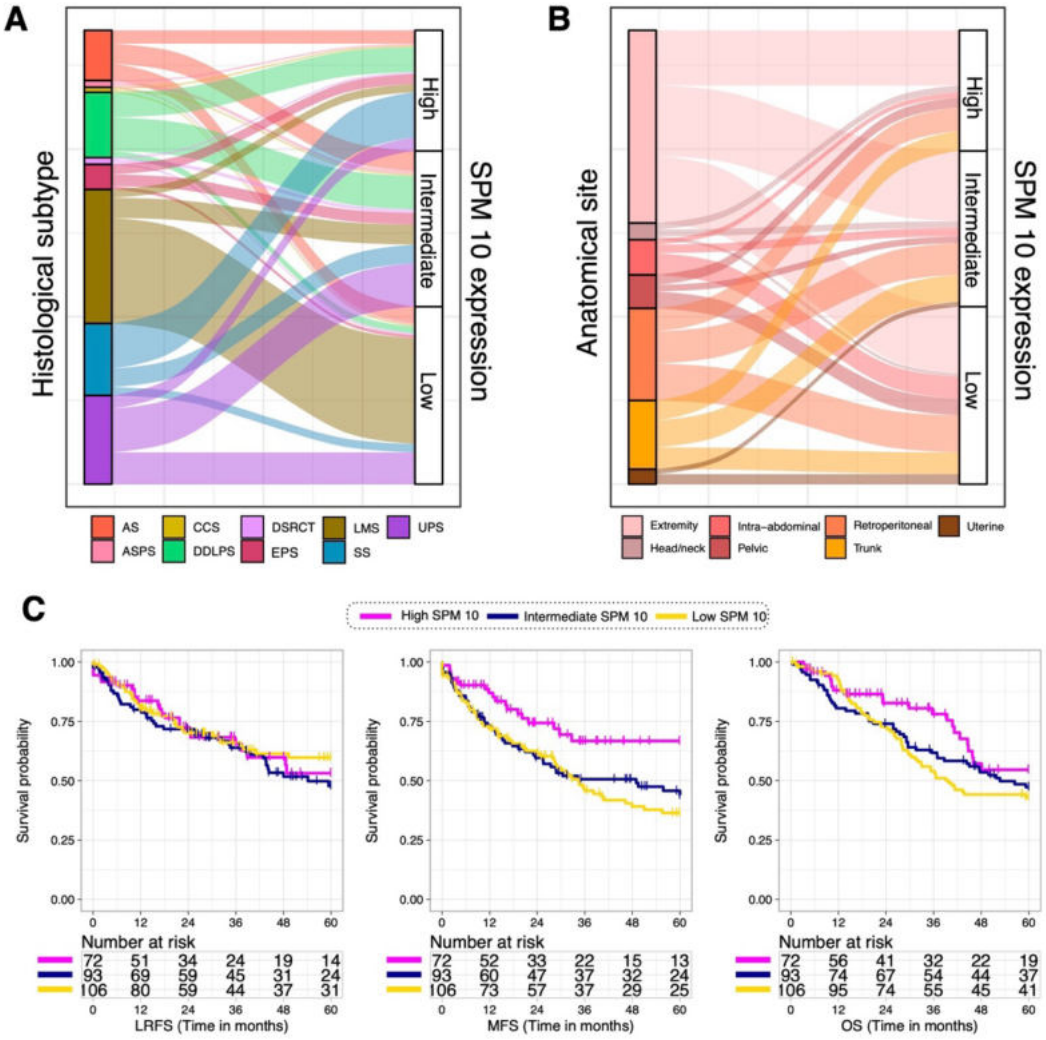


Figure 6.9 Clinical characterisation of sarcoma proteome module 10
(A-B) Alluvial plots illustrating the distribution of **(A)** histological subtype (excluding desmoid tumours (DES) and rhabdoid tumours (RT)) and **(B)** anatomical site across three SPM 10 subgroups. Subgroups were identified by tertile stratification based on median SPM 10 expression across the full cohort. **(C)** Kaplan Meier plots of local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) across the 3 SPM 6 subgroups. Corresponding univariable Cox regression detailed in **Table 6.2**. Abbreviations: AS = angiosarcoma; ASPS = alveolar soft part sarcoma; CCS = clear cell sarcoma; DDLPS = dedifferentiated liposarcoma; DSRCT = desmoplastic small round cell tumour; EPS = epithelioid sarcoma; LMS = leiomyosarcoma; SS = synovial sarcoma; UPS = undifferentiated pleomorphic sarcoma.

Table 6.2 Univariable Cox regression for sarcoma proteome module (SPM) 10
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. SPM subgroups identified by tertile stratification based on median expression across the full cohort. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval

		LRFS		MFS		OS	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
SPM 10	*low (ref)*	-	-	-	-	-	-
	intermediate	1.28 (0.815-2)	0.287	0.882 (0.6-1.3)	0.523	0.889 (0.605-1.31)	0.551
	high	1.09 (0.649-1.83)	0.747	**0.48 (0.285-0.803)**	**0.005**	0.635 (0.389-1.04)	0.07

dataset, SPM 6 and SPM 10 were descriptively analysed for associations with histology, grade (SPM 6 only), and anatomical site (SPM 10 only).

The TCGA data reduced to those genes/proteins identified in SPM 6 and unsupervised clustering was performed. Clustering annotations showed no association with grade, but a strong association with histological subtype (**Figure 6.10A**). Specifically, expression of SPM 6 tended to be highest in LMS cases. The univariable Cox regression showed a high median SPM 6 score to be significantly associated with a poorer MFS (**Figure 6.10C** and **Supplemental Table *6.7***), recapitulating proteomic observations. Application of the SPM 10 proteins to the TCGA data and subsequent unsupervised clustering again highlighted histology-based expression differences (**Figure 6.10B**). Most notably, a clear LMS specific cluster was observed, seemingly driven by the expression of approximately ~ 4 proteins (ALDH1A3, HSPA12B, FSCN1, PYCR1). Additionally, within LMS, uterine and other tumours were separated showing anatomical site-based differences in SPM 10 expression. Use of the univariable Cox regression and median SPM 10 score revealed no significant association with MFS or LRFS (**Figure 6.10D** and **Supplemental Table *6.7***). However, high SPM 10 was illustrated as significantly associated with a poorer OS (HR = 2.93, 95% CI = 1.27 – 6.76, p = 0.012). Notably, this is the inverse of the relationship shown in the proteomic data with MFS.

6.1.8 Discussion and summary

Using the comprehensive proteomic data from all MS-profiled cases, this Chapter defined the landscape of the STS proteome. By leveraging on the inherent network structure of protein systems, 14 SPMs were defined. The network model built showed impressive scale-free topology fit ($R^2 = 0.93$). The WGCNA authors themselves note that the scale-free topology assumption can be challenging to meet, particularly where different tissue types are analysed[521]. However, despite this cohort comprising many different tumour types, a robust proteome network could be derived. Overall STRINGdb coverage of WGCNA interactions was 32% and ranged from 8% to 41% within SPMs[602,603]. The highest overlap between STRINGdb and WGCNA observations were seen in WGCNA interactions with high co-expression weights. This illustrates WGCNA to capture STRINGdb-described biology with high confidence, as well as revealing novel interactions based on *de novo* analysis of proteomic expression.

Assessment of the functional biology of SPMs by overrepresentation analysis failed to yield any significant results. This was likely due to the small number of proteins present

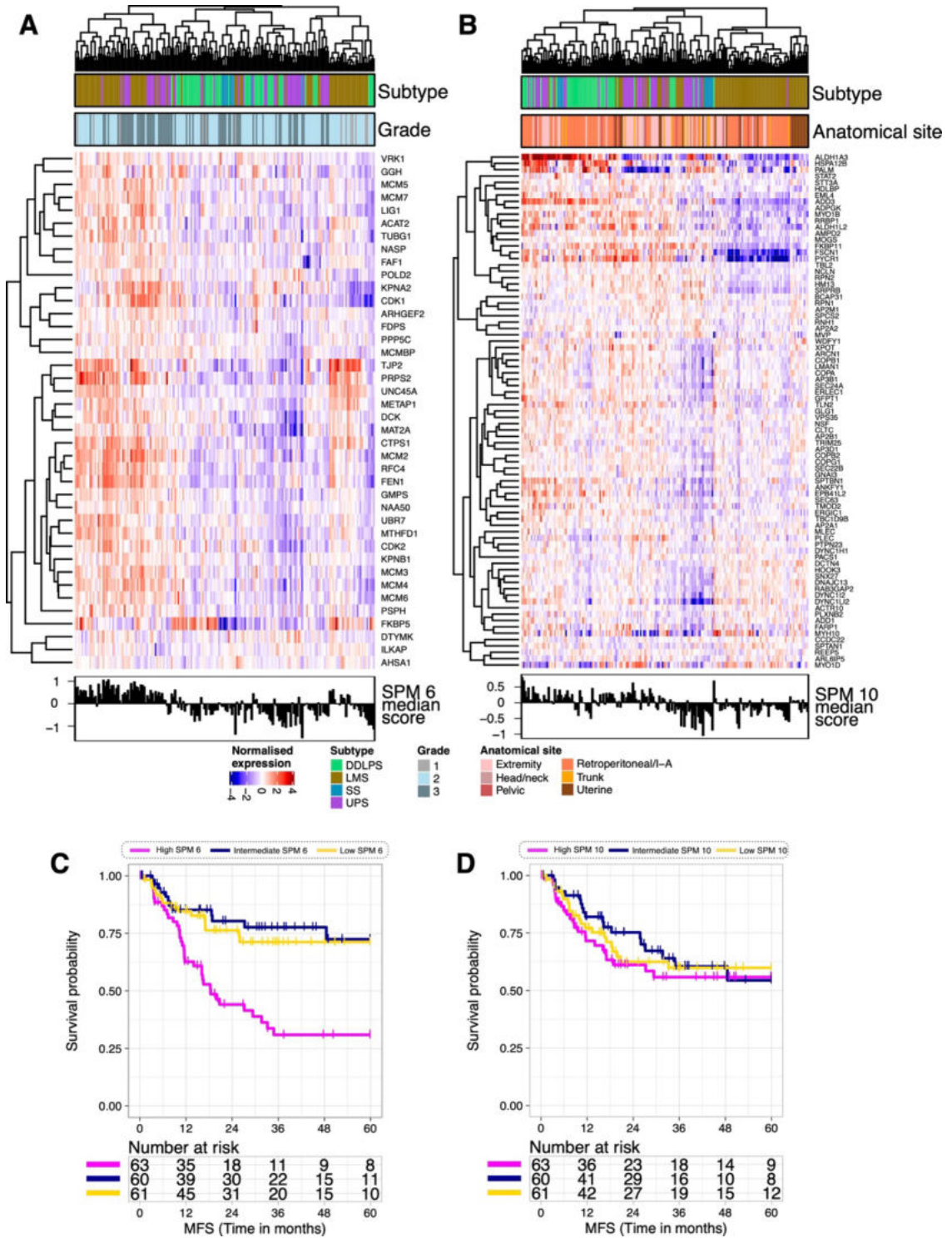


Figure 6.10 Assessment of sarcoma proteome modules 6 and 10 in The Cancer Genome Atlas (TCGA) RNAseq data

(A-B) Unsupervised clustering (Pearson's correlation distance) of **(A)** 41 SPM 6 and **(B)** 94 SPM 10 proteins across leiomyosarcoma (LMS), dedifferentiated liposarcoma (DDLPS), undifferentiated pleiomorphic sarcoma (UPS), and synovial sarcoma (SS) cases of the TCGA cohort (n = 184). Top annotations show **(A-B)** histological subtype, **(A)** grade and **(B)** and anatomical site. Bottom annotation shows median score for each SPM for each case. **(C-D)** Kaplan Meier plots of metastasis free survival (MFS) across the 3 SPM 6 subgroups **(C)** and 3 SPM 10 subgroups. Subgroups were identified by tertile stratification based on median SPM 10 expression across the full cohort. **(D)**. Corresponding univariable Cox regression detailed in **Supplemental Table 6.7**. Abbreviations: I-A = Intra-abdominal

within each SPM and the reduced coverage of genome-wide gene sets in the MS data. However, inspection of each SPM PPI illustrated SPMs to capture a range of STS biological activity. Biology spanned the regulation of DNA, RNA, and proteins (replication, splicing, translation, and proteasomal degradation), vesicle trafficking, and matrisomal processes (cell adhesion, ECM, and immune proteins). The SPMs identified were, at least in part, reflective of cohort composition; evidenced by the identification of a muscle-related SPM (SPM 1) and ECM-related SPM (SPM 8). SPM 1 showed high expression in nearly all LMS tumours and showed low expression in all other histological subtypes, consistent with LMS being smooth muscle derived[4]. Whilst SPM 8 showed high expression in DES and low expression in almost all tumours of other histological subtypes, consistent with the fibrotic nature of DES[557]. In fact, the expression of all SPMs identified was associated with histology. Observations of interest include the high expression of SPM 9 in both DES and UPS, two tumours with very different biological and clinical features[557,621]. SPM 9 was enriched in vesicle transport proteins and antigen presentation machinery. The biological implications of the shared enrichment of these proteins in DES and UPS is unclear. Also of note is the unique co-upregulation of SPM 2, SPM 5, and SPM 13 in SS tumours. SS are driven by the *SS18-SSX1/2/4* fusion and show a distinctive profile at the proteomic level (**Chapter 4**), and transcriptomic level[36,165]. High SPM 2 in SS suggests an active immune component; however, this in contrast to the current literature which reports low immune activity in SS[36,667]. High expression of splicing proteins (SPM 5) in SS may be reflective of the underlying fusion gene characteristic of SS. Previous data has suggested alternative splicing to play a role in fusion transcript formation[668]. SPM 13 comprises oxidative phosphorylation proteins, and therefore high SPM 13 in SS indicates active mitochondrial respiration. The downstream consequences of this are unknown. However, oxidative phosphorylation has been reported as enriched in some STS subtypes relative to carcinomas, and hypoxia signatures have been reported to hold prognostic value in STS[384,385,387,388,669]. Finally, it is notable that of the 2 immune-related SPMs (SPM 2 and SPM 7), 1 (SPM 7) shows high heterogeneity within DDLPS and UPS. This may be reflective of observations made in **Chapter 5** which showed distinctive immune processes to be active in subsets of DDLPS an UPS cases. A high number of SPMs were also associated with grade; however, this was driven by tumours where grading information was not available, and therefore was likely driven by histology. Indeed, DES are not graded and SPMs notably high in DES (7, 8, and 9) and low in DES (6) were amongst those identified in SPM-grade analyses[4,53,54,71]. The association between SPM and anatomical site is also hypothesised to be driven by histological subtype. For example, SPM 1 showed high expression in all uterine tumours, yet all uterine tumours profiled were LMS. Overall, the

SPMs illustrated a recapitulation of known tumour biology, highlighted commonalities between distinct STS histological subtypes, and reproduced findings of heterogeneity investigated elsewhere in this project. To build on this work, future analyses include those to differentiate between SPMs which are noted to harbour similar biology. For example, 3 SPMs enriched in splicing activity are reported (SPM 3, SPM 4, and SPM 5). However, these SPMs show different expression profiles across tumours and therefore likely describe different functional biology. Important next steps involve attempts to delineate the proteins central to each SPM to better describe their function.

As a proof-of-principle for the clinical utility of these SPMs, their association with outcome was assessed. This revealed SPM 6 and SPM 10 to harbour significant and independent prognostic value for MFS in multivariable analyses. SPM 6 comprised DNA replication machinery, and network analysis specifically revealed the MCM complex to have high influence within this group of proteins. This may indicate active and/or aberrant DNA replication in these tumours which can have consequences on cell cycle and proliferative activity[670,671]. As such, high MCM protein expression may be reflective of genomic instability and may represent a surrogate measure for such instability in these tumours. Irrespective of the underlying mechanism, this chapter showed a high expression of SPM 6 to be prognostic for poor MFS. Moreover, application of SPM 6 to the TCGA cohort illustrated a recapitulation of its prognostic value in an independent cohort. This is in support of the hypothesis that high MCM expression indicates genomic instability, as tumours with high instability have been reported to have an increased metastasis risk (as measured by CINSARC; discussed in **section 1.5.2.1**)

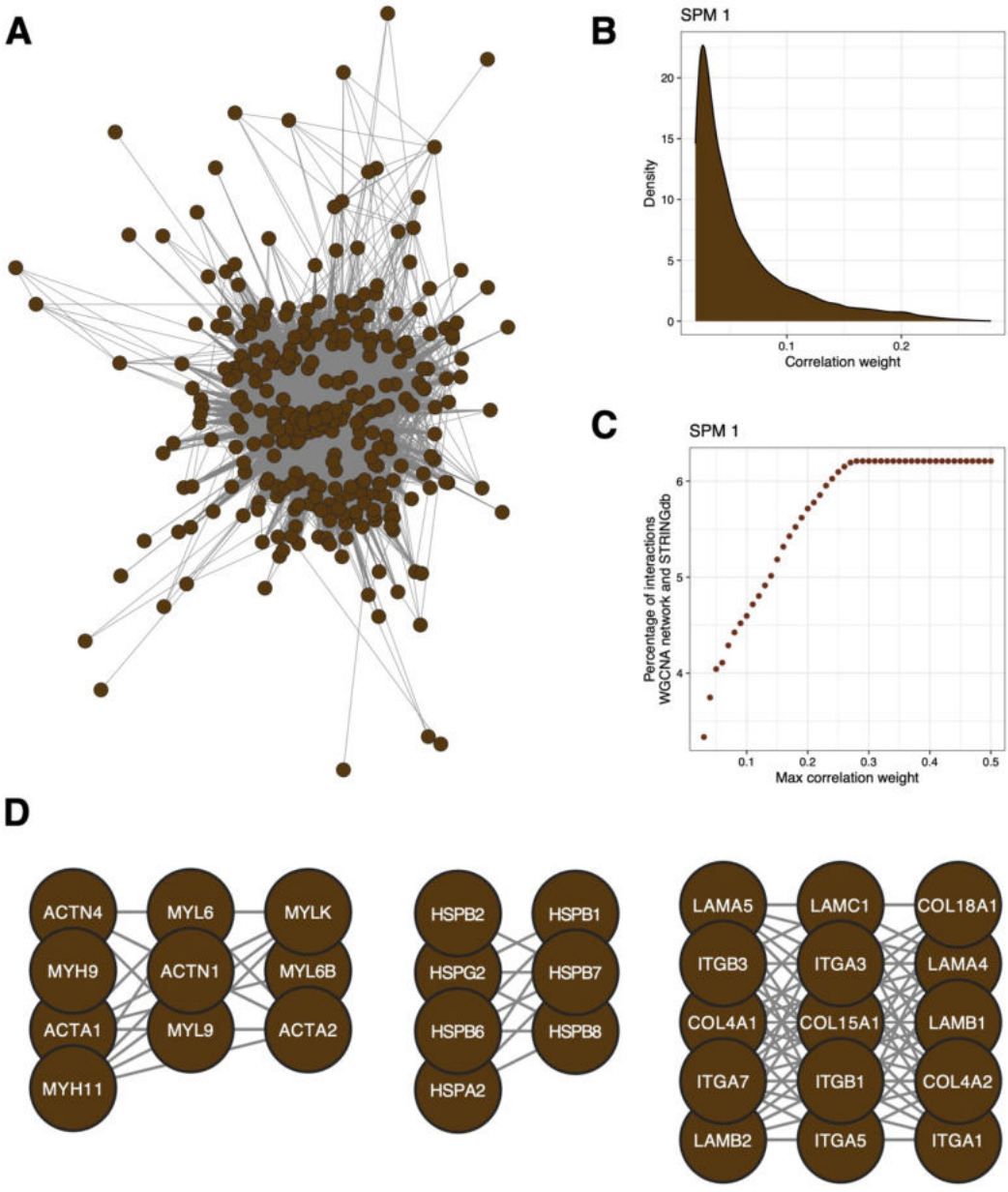
Whilst the prognostic value of SPM 6 can be rationalised based on the current literature, the prognostic value of vesicle trafficking proteins (i.e., SPM 10) in STS has not been reported before. Vesicular trafficking can correspond to the intracellular localisation, secretion, or endosomal trafficking of biological molecules such as proteins[672]. These processes cover a broad range of functions which can impact tumour behaviour in many different ways. For example, vesicle transport proteins such as the Rab GTPases have been shown to harbour both oncogenic and tumour suppressive effects, dependent on the context[673]. With the current data in this project, it was not possible to determine whether SPM 10 describes tumour suppressive and/or oncogenic activity. Although high SPM 10 expression was shown to be associated with a superior MFS in this cohort, thus suggesting the proteins in SPM 10 confer tumour suppressive effects in STS. Application of SPM 10 to the TCGA data did not recapitulate this prognostic significance. In fact, the inverse relationship was identified with OS. In TCGA high expression of SPM conferred

a significantly poorer OS. This may be reflective of the multiple and contrasting roles of vesicle trafficking proteins, or alternatively may indicate that the expression of SPM 10 and it's relation to a superior MFS is an attribute of the proteome not the transcriptome. This stresses the importance of protein-level characterisation of tumours. Additionally, it may be the case that the TCGA cohort composition itself is restricting recapitulation. In **Chapter 5**, clinicopathological features of the LMS cases were shown to significantly differ between the MS- and TCGA-profiled cohorts. Moreover, SPMs were derived using data from 11 different histological subtypes, of which TCGA profiled only 4. Despite not being recapitulated in the TCGA cohort, the revelation that vesicle trafficking proteins are associated with outcome illustrates one advantage of the *de novo* WGCNA approach used. Such unsupervised methods can identify novel findings without the use of prior biological knowledge.

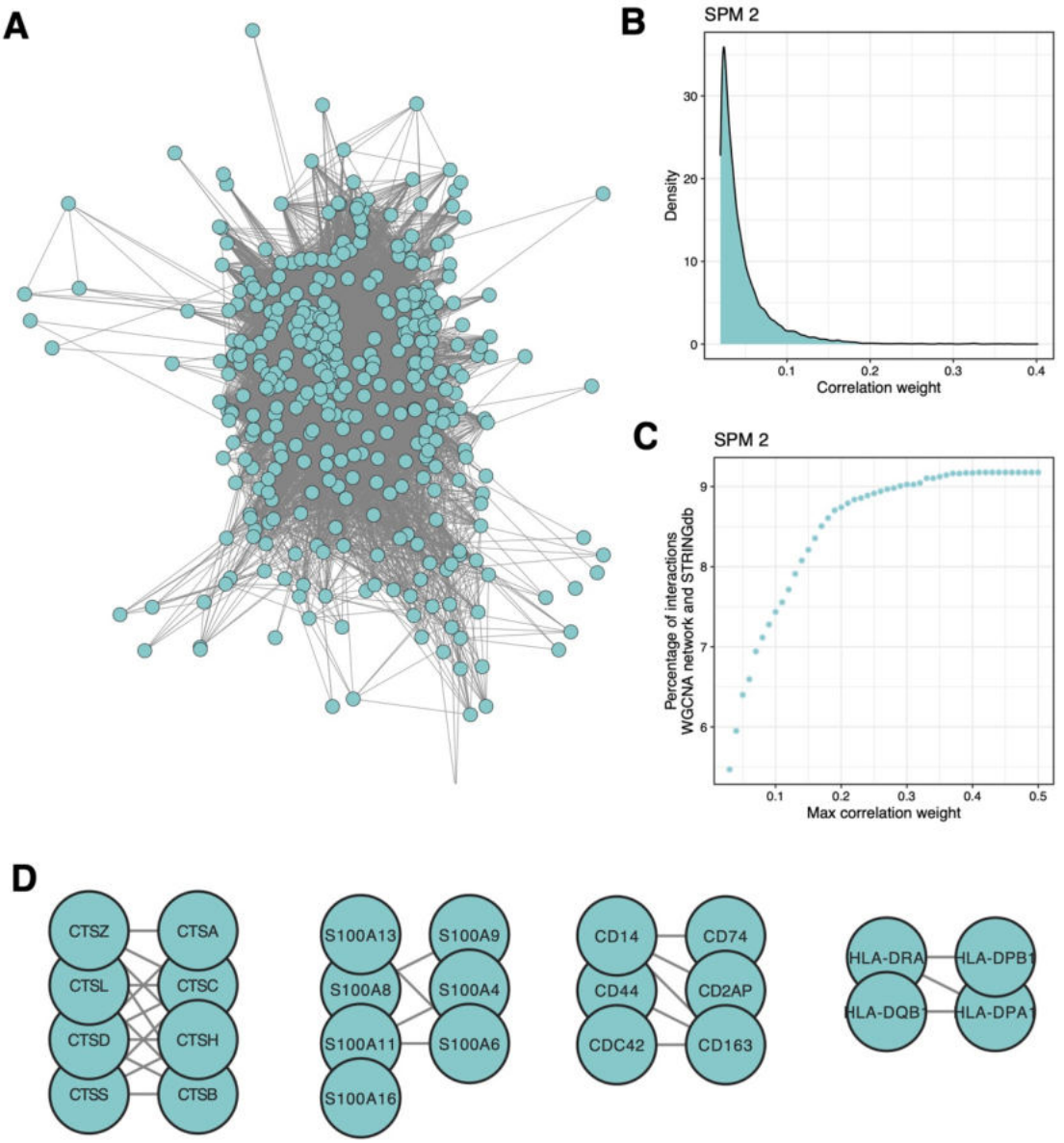
There are several future research avenues leading on from this work. Specifically, it would be interesting to assess SPM prognostication in an independent MS dataset, and benchmark its performance to current risk stratification tools within STS clinical practice (such as the nomograms; **section 1.2.2.2**). Whilst such analysis of an independent dataset could involve comprehensive MS as was performed for this project, given the SPMs of interest contain relatively low numbers of proteins, targeted MS may be more appropriate. Furthermore, it would be of interest to identify whether the proteins of influence revealed by network analysis of SPM 6 and SPM 10 capture sufficient biology to confer a prognostic value themselves. This would reduce the number of proteins needing assessment and may facilitate risk stratification by IHC. Another avenue for exploration includes the assessment of the remaining SPMs not focused on herein. This project demonstrated prognostic utility in 2 SPMs by use of a median summary score. However, there are 12 other SPMs which could be further explored using other approaches. As well as further analyses directed at prognostication, this proteome network could also be assessed for drug targets. Utilising the network structure of the proteome may be a promising approach to reveal candidate therapeutic choices. Though the network structure, drug targets can be assessed for their influence within SPMs and across the proteome network, which could highlight the pathways most vulnerable to therapeutic intervention.

6.2 Supplemental material

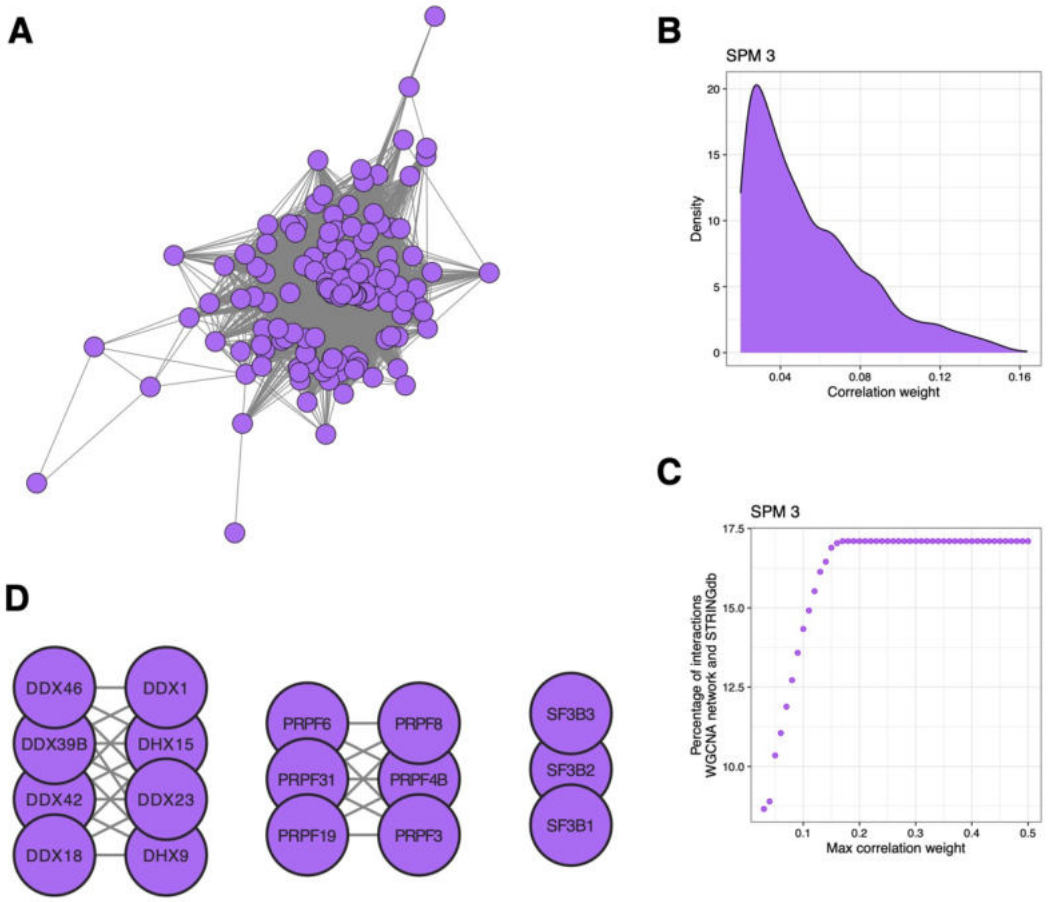
6.2.1 Supplemental figures



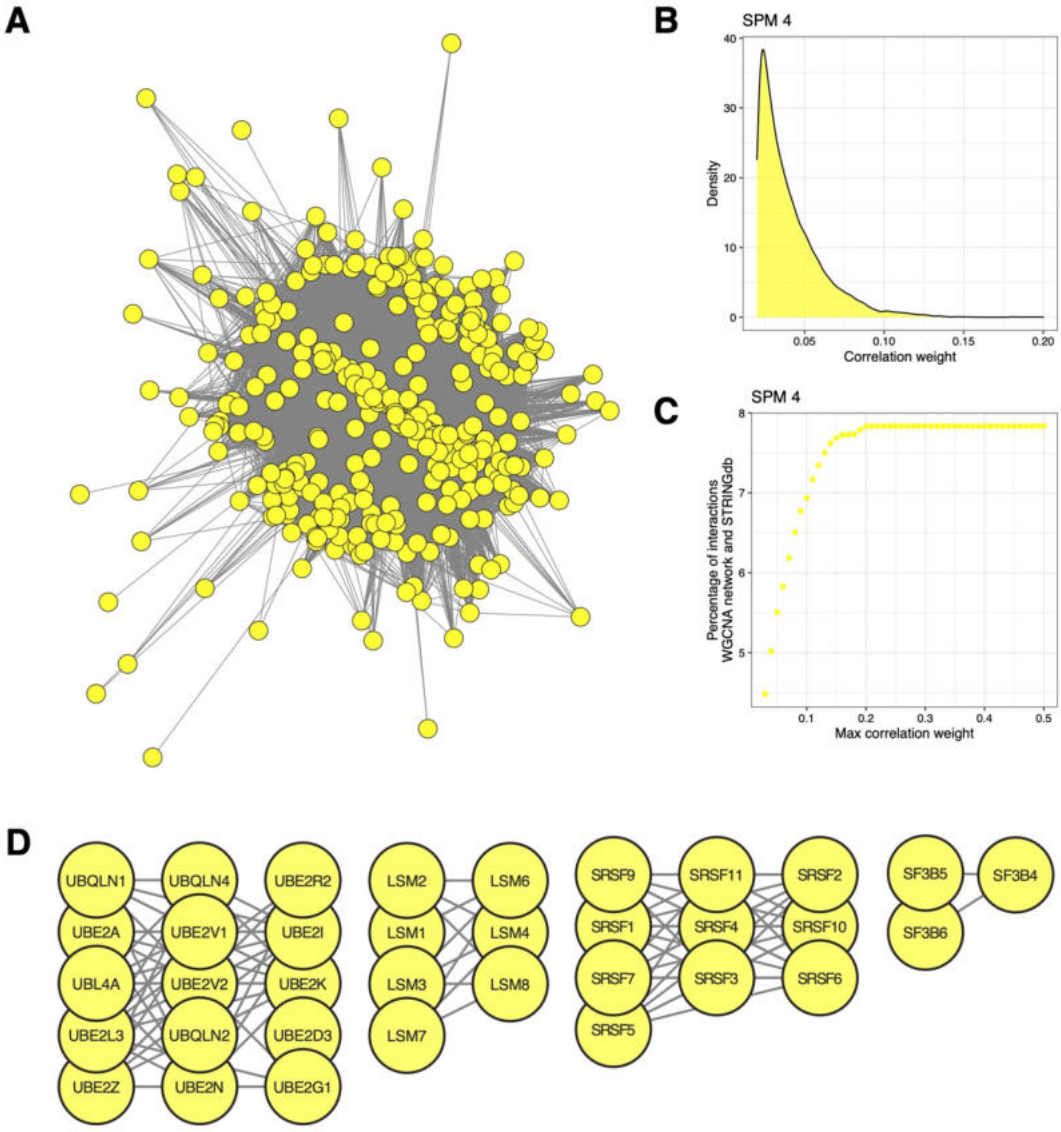
Supplemental Figure 6.1 Sarcoma proteome module (SPM) 1
(A) SPM 1 protein co-expression network comprising 354 nodes and 17,475 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 1 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



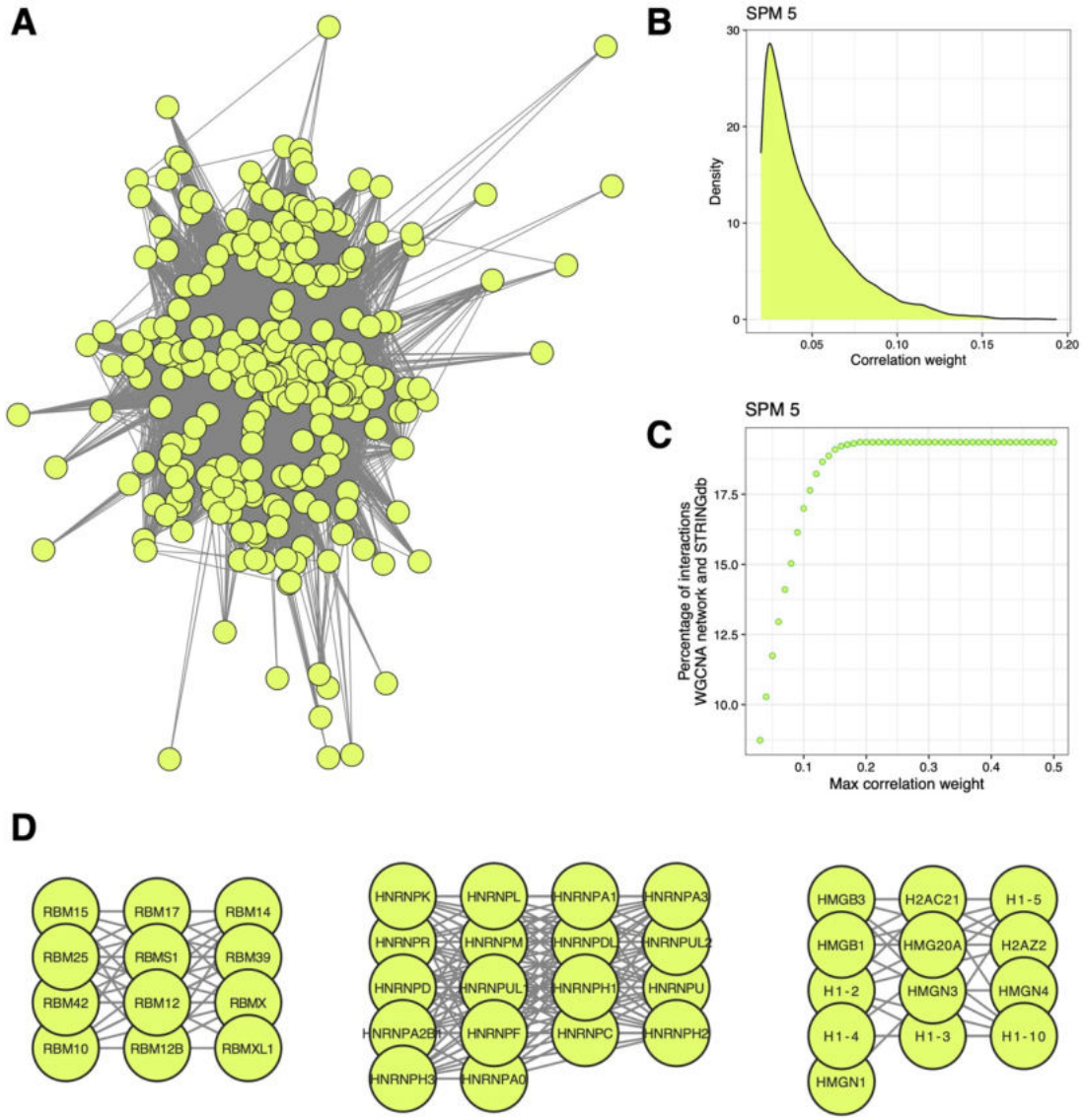
Supplemental Figure 6.2 Sarcoma proteome module (SPM) 2
(A) SPM 2 protein co-expression network comprising 383 nodes and 13,642 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 2 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



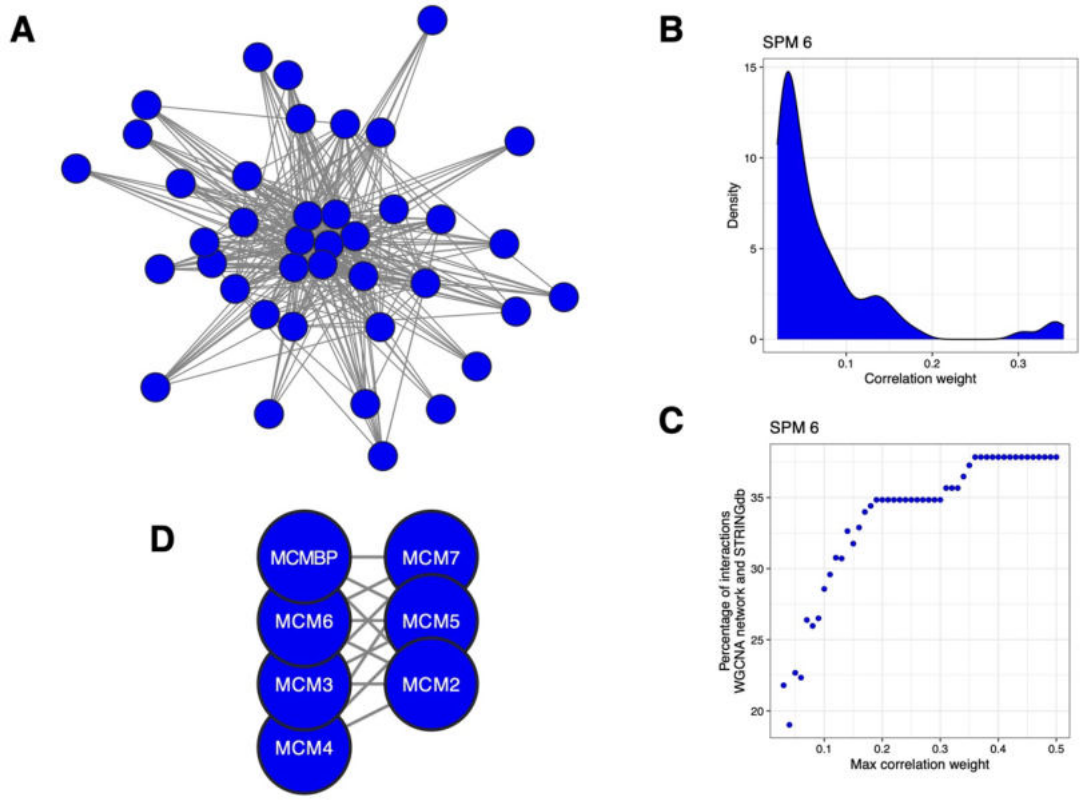
Supplemental Figure 6.3 Sarcoma proteome module (SPM) 3
(A) SPM 3 protein co-expression network comprising 136 nodes and 5,186 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 3 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



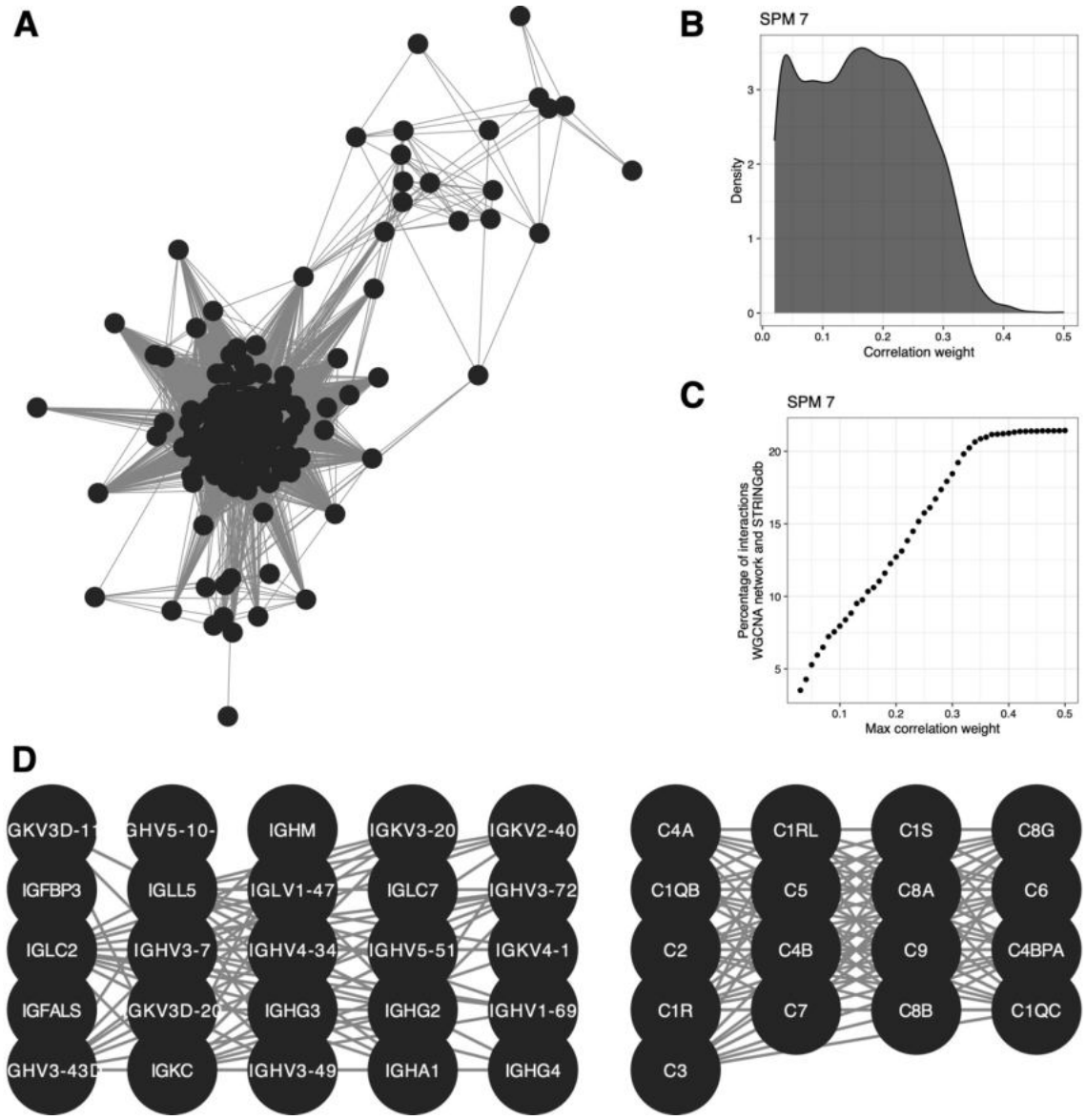
Supplemental Figure 6.4 Sarcoma proteome module (SPM) 4
(A) SPM 4 protein co-expression network comprising 342 nodes and 18,828 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 4 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



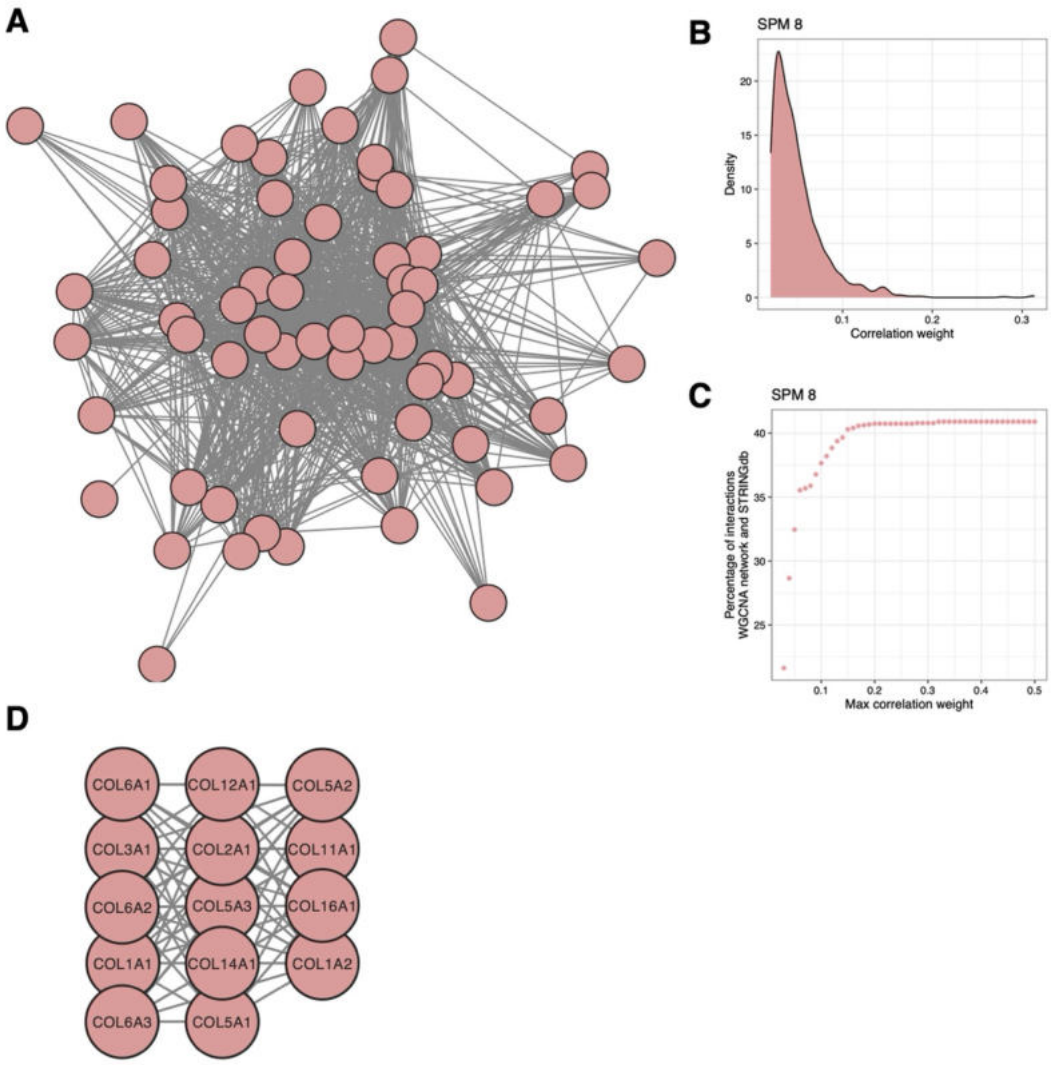
Supplemental Figure 6.5 Sarcoma proteome module (SPM) 5
(A) SPM 5 protein co-expression network comprising 275 nodes and 13,853 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 5 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



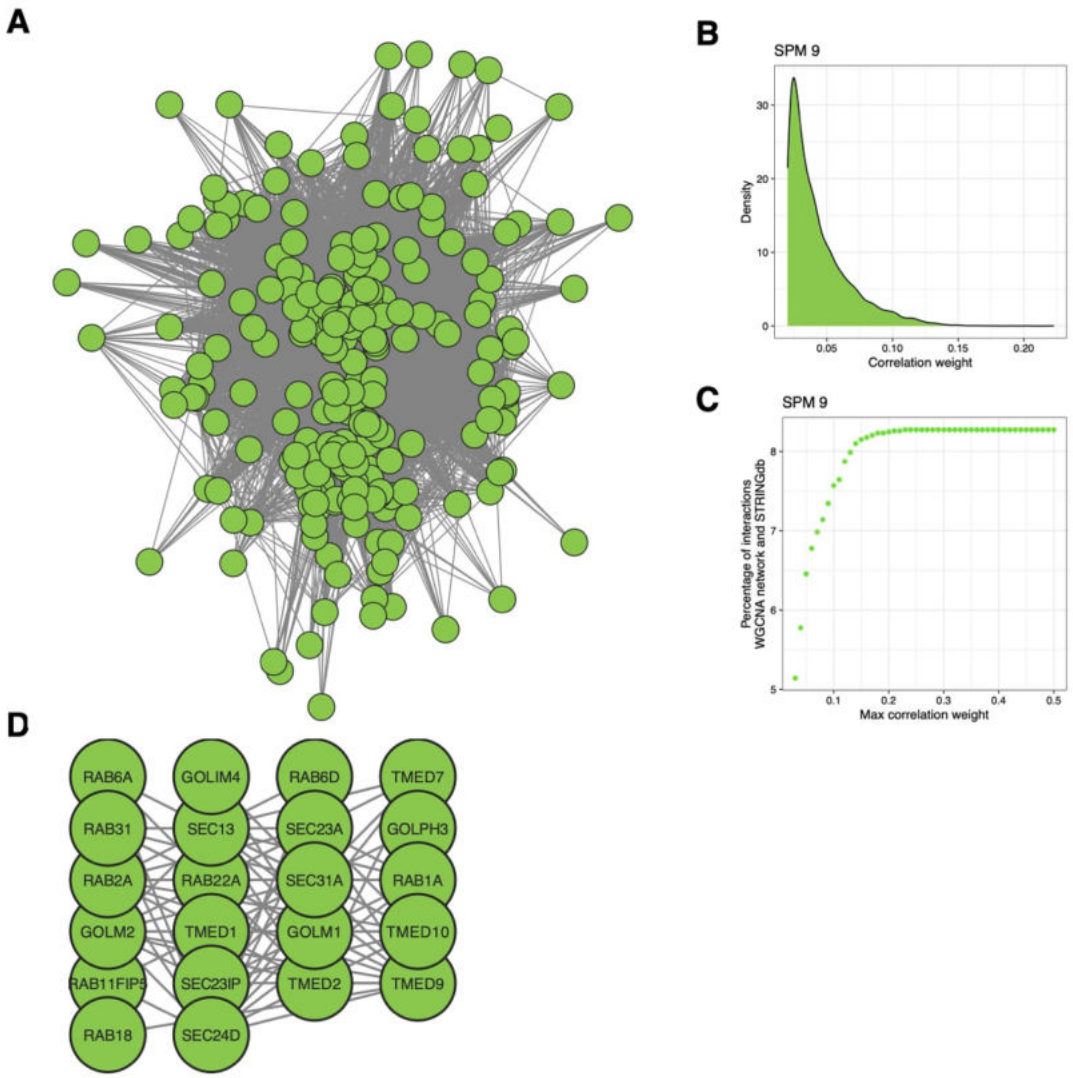
Supplemental Figure 6.6 Sarcoma proteome module (SPM) 6
(A) SPM 6 protein co-expression network comprising 41 nodes and 325 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 6 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



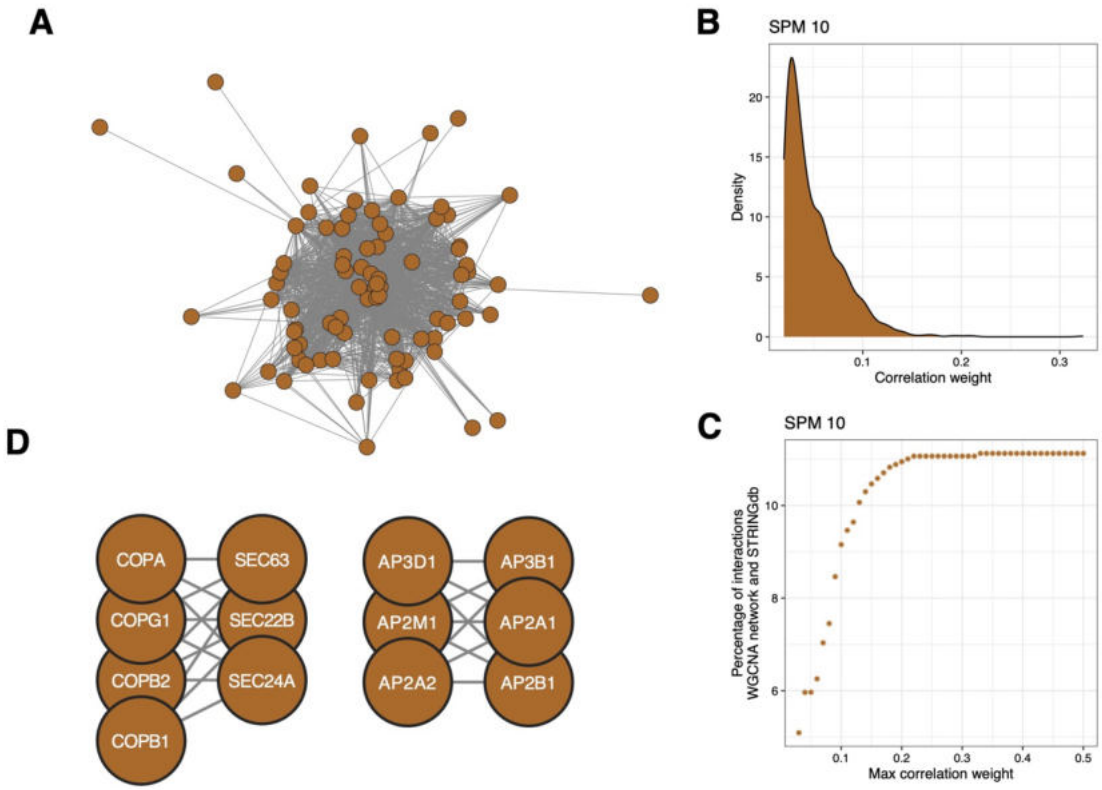
Supplemental Figure 6.7 Sarcoma proteome module (SPM) 7
(A) SPM 7 protein co-expression network comprising 176 nodes and 9,809 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 7 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



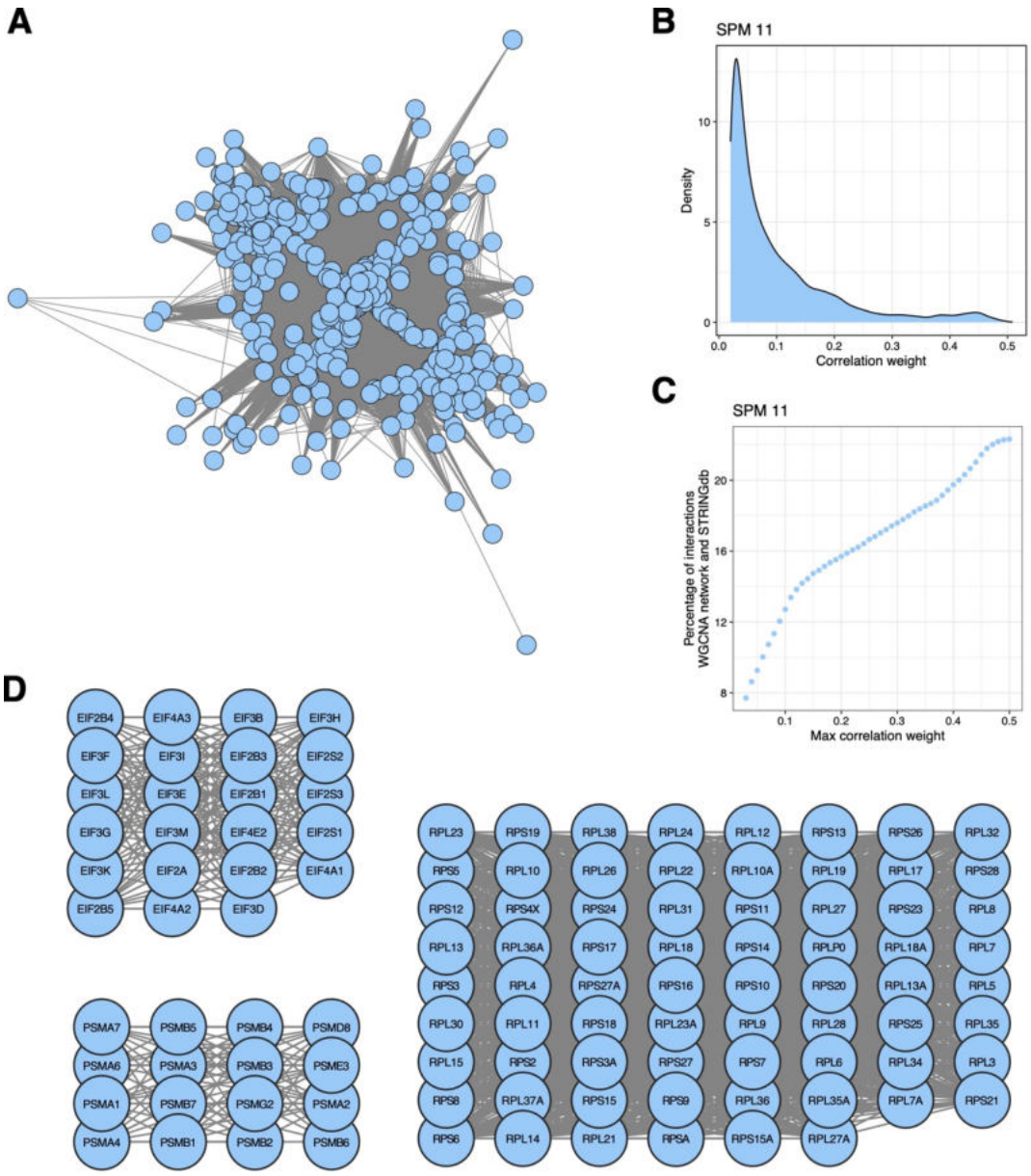
Supplemental Figure 6.8 Sarcoma proteome module (SPM) 8
(A) SPM 8 protein co-expression network comprising 63 nodes and 1,088 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 8 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



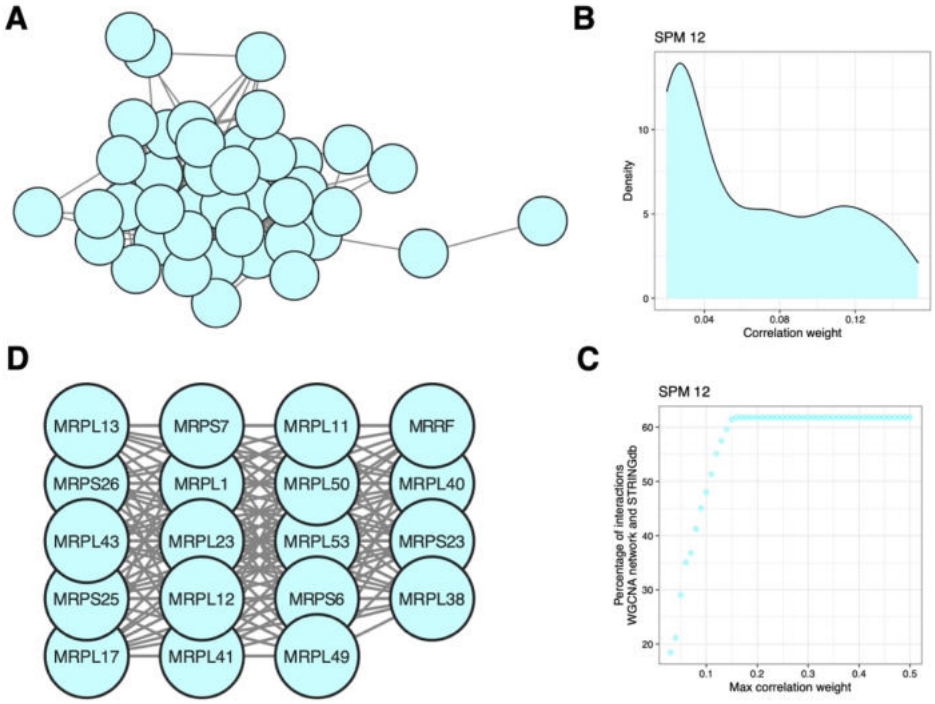
Supplemental Figure 6.9 Sarcoma proteome module (SPM) 9
(A) SPM 9 protein co-expression network comprising 231 nodes and 6,527 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 9 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



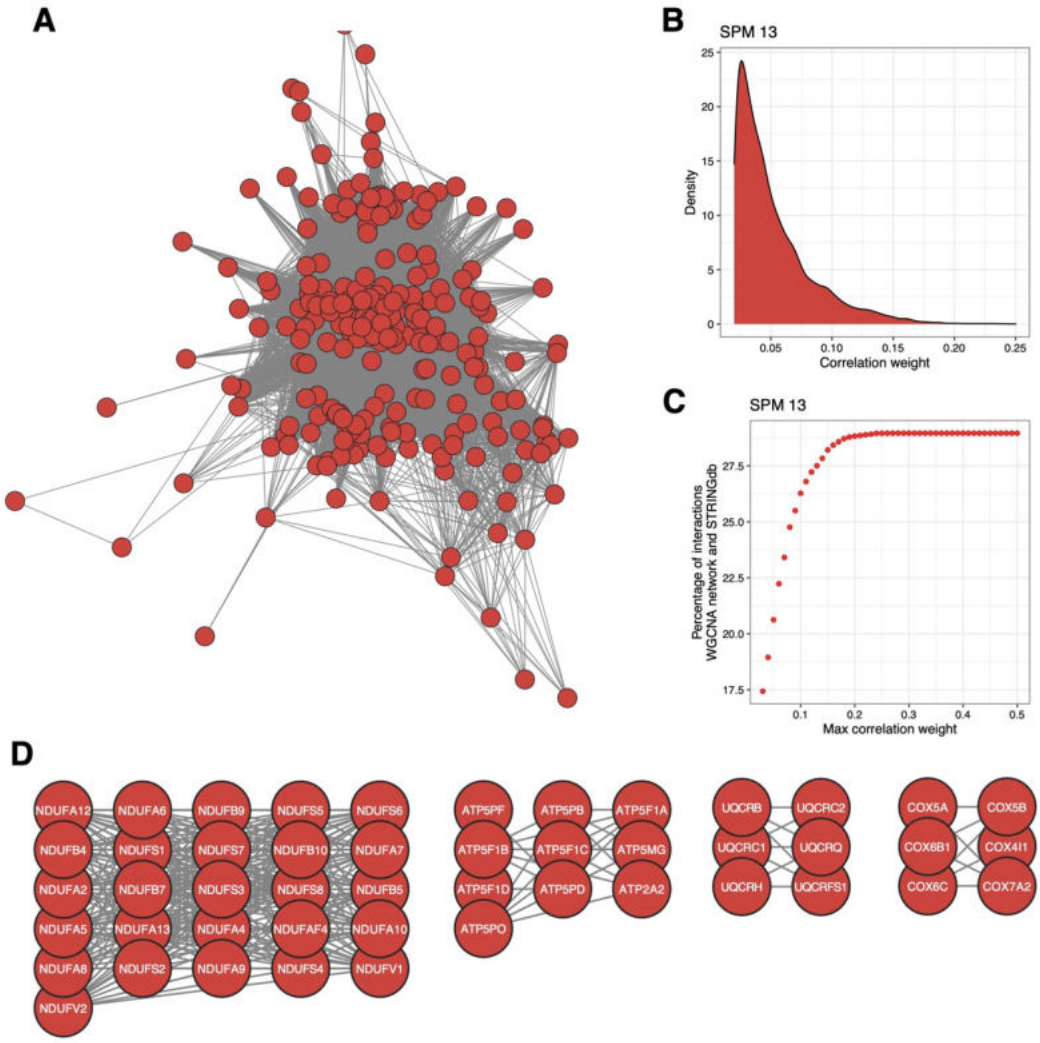
Supplemental Figure 6.10 Sarcoma proteome module (SPM) 10
(A) SPM 10 protein co-expression network comprising 84 nodes and 1,492 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 10 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



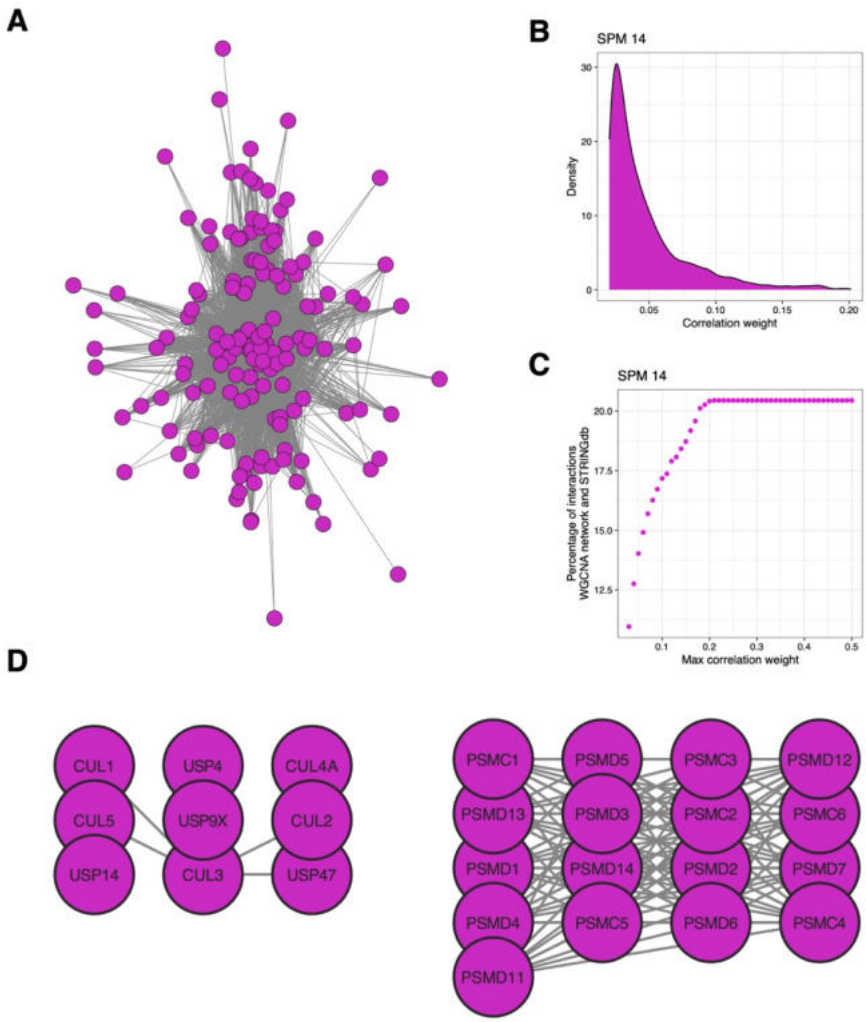
Supplemental Figure 6.11 Sarcoma proteome module (SPM) 11
(A) SPM 11 protein co-expression network comprising 391 nodes and 36,419 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 11 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



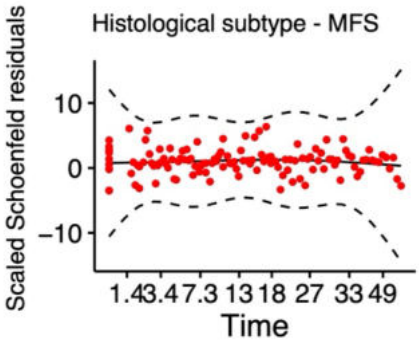
Supplemental Figure 6.12 Sarcoma proteome module (SPM) 12
(A) SPM 12 protein co-expression network comprising 41 nodes and 335 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 12 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



Supplemental Figure 6.13 Sarcoma proteome module (SPM) 13
(A) SPM 13 protein co-expression network comprising 231 nodes and 10,929 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 13 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



Supplemental Figure 6.14 Sarcoma proteome module (SPM) 14
(A) SPM 14 protein co-expression network comprising 151 nodes and 2,690 edges (restricted to co-expression weight ≥ 0.05). Nodes indicate proteins, edges show co-expression between protein expression, where a thicker line indicates a stronger correlation. **(B)** Distribution of co-expression weights within SPM 14 **(C)** Dotplot showing the percentage overlap between STRINGdb interactions and WGCNA-revealed interactions, where different maximum WGCNA correlation weights are applied. **(D)** Subnetworks of interest manually selected from **(A).**



Supplemental Figure 6.15 Assessment of the proportional hazards (PH) assumption in the multivariable Cox model inclusive of sarcoma proteome module (SPM) 6
Plot shown for variable-model combination where a minor violation of the PH assumption was identified. Scaled Schoenfeld residuals plotted for histological subtype in the metastasis free survival (MFS) model. Solid black line indicates a smoothed spline fit of residuals and dashed black lines indicate +/- 2-standard error.

6.2.2 Supplemental tables

Supplemental Table 6.1 Statistical associations between clinicopathological features and sarcoma proteome modules (SPM)
Significant results in bold. Abbreviations: d.f = degrees of freedom; FDR = false discovery rate

Variable 1	Variable 2	X2	d.f	p	FDR	Variable 1	Variable 2	X2	d.f	p	FDR
SPM 1	**Anatomical site**	**38.073**	**6**	**<0.001**	**<0.001**	SPM 6	Anatomical site	14.586	6	0.024	0.062
SPM 1	Grade	5.409	1	0.020	0.054	**SPM 6**	**Grade**	**19.817**	**1**	**<0.001**	**<0.001**
SPM 1	**Performance status**	**11.722**	**2**	**0.003**	**0.011**	SPM 6	Performance status	6.028	2	0.049	0.117
SPM 1	Sex	0.027	1	0.870	0.895	SPM 6	Sex	0.509	1	0.476	0.617
SPM 1	**Histological subtype**	**182.105**	**8**	**<0.001**	**<0.001**	**SPM 6**	**Histological subtype**	**160.058**	**8**	**<0.001**	**<0.001**
SPM 1	Tumour depth	1.095	1	0.295	0.450	SPM 6	Tumour depth	1.519	1	0.218	0.363
SPM 1	Tumour margin	1.185	2	0.553	0.691	SPM 6	Tumour margin	4.709	2	0.095	0.190
SPM 2	Anatomical site	9.142	6	0.166	0.300	**SPM 7**	**Anatomical site**	**19.828**	**6**	**0.003**	**0.011**
SPM 2	**Grade**	**14.489**	**1**	**<0.001**	**0.001**	**SPM 7**	**Grade**	**11.793**	**1**	**0.001**	**0.002**
SPM 2	Performance status	0.946	2	0.623	0.752	SPM 7	Performance status	0.438	2	0.803	0.870
SPM 2	Sex	5.410	1	0.020	0.054	SPM 7	Sex	1.117	1	0.291	0.450
SPM 2	**Histological subtype**	**177.846**	**8**	**<0.001**	**<0.001**	**SPM 7**	**Histological subtype**	**63.291**	**8**	**<0.001**	**<0.001**
SPM 2	Tumour depth	2.828	1	0.093	0.190	SPM 7	Tumour depth	3.868	1	0.049	0.117
SPM 2	Tumour margin	0.569	2	0.752	0.846	SPM 7	Tumour margin	2.037	2	0.361	0.503
SPM 3	Anatomical site	7.260	6	0.298	0.450	SPM 8	Anatomical site	12.517	6	0.051	0.120
SPM 3	Grade	0.078	1	0.780	0.863	**SPM 8**	**Grade**	**14.074**	**1**	**<0.001**	**0.001**
SPM 3	Performance status	4.043	2	0.132	0.253	SPM 8	Performance status	6.479	2	0.039	0.098
SPM 3	Sex	0.053	1	0.818	0.876	SPM 8	Sex	1.055	1	0.304	0.450
SPM 3	**Histological subtype**	**116.064**	**8**	**<0.001**	**<0.001**	**SPM 8**	**Histological subtype**	**104.626**	**8**	**<0.001**	**<0.001**
SPM 3	Tumour depth	0.425	1	0.514	0.659	SPM 8	Tumour depth	1.993	1	0.158	0.291
SPM 3	Tumour margin	0.341	2	0.843	0.877	SPM 8	Tumour margin	1.646	2	0.439	0.583
SPM 4	**Anatomical site**	**18.669**	**6**	**0.005**	**0.016**	SPM 9	Anatomical site	12.224	6	0.057	0.128
SPM 4	**Grade**	**18.304**	**1**	**<0.001**	**<0.001**	**SPM 9**	**Grade**	**27.683**	**1**	**<0.001**	**<0.001**
SPM 4	Performance status	5.407	2	0.067	0.147	SPM 9	Performance status	0.616	2	0.735	0.846
SPM 4	Sex	0.396	1	0.529	0.669	**SPM 9**	**Sex**	**5.699**	**1**	**0.017**	**0.048**
SPM 4	**Histological subtype**	**116.164**	**8**	**<0.001**	**<0.001**	**SPM 9**	**Histological subtype**	**232.628**	**8**	**<0.001**	**<0.001**
SPM 4	Tumour depth	1.077	1	0.299	0.450	**SPM 9**	**Tumour depth**	**7.184**	**1**	**0.007**	**0.023**
SPM 4	Tumour margin	2.392	2	0.302	0.450	SPM 9	Tumour margin	4.495	2	0.106	0.205

continuation of table from previous page

Variable 1	Variable 2	X2	d.f	p	FDR
SPM 5	Anatomical site	4.580	6	0.599	0.740
SPM 5	**Grade**	**9.491**	**1**	**0.002**	**0.008**
SPM 5	Performance status	5.197	2	0.074	0.159
SPM 5	Sex	0.108	1	0.743	0.846
SPM 5	**Histological subtype**	**104.620**	**8**	**<0.001**	**<0.001**
SPM 5	Tumour depth	0.072	1	0.789	0.863
SPM 5	Tumour margin	1.622	2	0.444	0.583
SPM 11	**Anatomical site**	**18.179**	**6**	**0.006**	**0.019**
SPM 11	Grade	0.049	1	0.825	0.876
SPM 11	Performance status	3.424	2	0.180	0.321
SPM 11	Sex	0.823	1	0.364	0.503
SPM 11	**Histological subtype**	**124.326**	**8**	**<0.001**	**<0.001**
SPM 11	Tumour depth	1.631	1	0.202	0.345
SPM 11	Tumour margin	0.997	2	0.607	0.741
SPM 12	Anatomical site	6.832	6	0.337	0.491
SPM 12	Grade	0.182	1	0.670	0.790
SPM 12	Performance status	0.557	2	0.757	0.846
SPM 12	Sex	4.648	1	0.031	0.080
SPM 12	**Histological subtype**	**114.231**	**8**	**<0.001**	**<0.001**
SPM 12	**Tumour depth**	**8.763**	**1**	**0.003**	**0.011**
SPM 12	Tumour margin	8.318	2	0.016	0.047
SPM 13	Anatomical site	6.560	6	0.363	0.503
SPM 13	**Grade**	**21.026**	**1**	**<0.001**	**<0.001**
SPM 13	Performance status	0.354	2	0.838	0.877
SPM 13	Sex	1.663	1	0.197	0.345
SPM 13	**Histological subtype**	**131.092**	**8**	**<0.001**	**<0.001**
SPM 13	Tumour depth	0.024	1	0.878	0.895
SPM 13	Tumour margin	2.999	2	0.223	0.366
SPM 10	**Anatomical site**	**29.210**	**6**	**<0.001**	**<0.001**
SPM 10	Grade	0.905	1	0.341	0.491
SPM 10	Performance status	4.689	2	0.096	0.190
SPM 10	Sex	0.782	1	0.376	0.513
SPM 10	**Histological subtype**	**166.794**	**8**	**<0.001**	**<0.001**
SPM 10	Tumour depth	2.037	1	0.154	0.288
SPM 10	Tumour margin	1.928	2	0.381	0.513
SPM 14	**Anatomical site**	**38.286**	**6**	**<0.001**	**<0.001**
SPM 14	**Grade**	**16.823**	**1**	**<0.001**	**<0.001**
SPM 14	Performance status	2.880	2	0.237	0.383
SPM 14	Sex	0.165	1	0.685	0.799
SPM 14	**Histological subtype**	**110.637**	**8**	**<0.001**	**<0.001**
SPM 14	Tumour depth	0.004	1	0.948	0.957
SPM 14	Tumour margin	3.181	2	0.204	0.345
SPM 15	Anatomical site	1.435	6	0.964	0.964
SPM 15	**Grade**	**5.782**	**1**	**0.016**	**0.047**
SPM 15	**Performance status**	**9.637**	**2**	**0.008**	**0.025**
SPM 15	Sex	0.196	1	0.658	0.785
SPM 15	**Histological subtype**	**64.364**	**8**	**<0.001**	**<0.001**
SPM 15	Tumour depth	3.721	1	0.054	0.123
SPM 15	Tumour margin	4.837	2	0.089	0.187

Supplemental Table 6.2 Univariable Cox regression for sarcoma proteome modules (SPM)
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. SPM measures are median scores for all proteins in the SPM. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval

	LRFS			MFS			OS		
	HR (95% CI)	**p**	**FDR**	**HR (95% CI)**	**p**	**FDR**	**HR (95% CI)**	**p**	**FDR**
SPM 1	**0.554 (0.372-0.825)**	**0.004**	**0.041**	1.08 (0.794-1.47)	0.625	0.821	0.789 (0.571-1.09)	0.153	0.377
SPM 2	1.19 (0.775-1.83)	0.426	0.673	1.01 (0.681-1.49)	0.968	0.990	1.44 (0.984-2.12)	0.061	0.215
SPM 3	0.992 (0.672-1.46)	0.966	0.990	0.872 (0.612-1.24)	0.449	0.674	0.774 (0.536-1.12)	0.171	0.385
SPM 4	0.857 (0.509-1.44)	0.561	0.789	1.71 (1.08-2.71)	0.023	0.114	**1.85 (1.16-2.95)**	**0.010**	**0.065**
SPM 5	0.961 (0.544-1.7)	0.890	0.954	1.34 (0.817-2.19)	0.247	0.493	1.32 (0.813-2.13)	0.263	0.493
SPM 6	1 (0.664-1.51)	0.990	0.990	**2.19 (1.52-3.15)**	**< 0.001**	**0.001**	1.47 (1.03-2.09)	0.034	0.153
SPM 7	1.24 (1.01-1.52)	0.042	0.171	0.842 (0.687-1.03)	0.099	0.298	1.02 (0.843-1.23)	0.846	0.929
SPM 8	0.965 (0.739-1.26)	0.794	0.910	0.908 (0.715-1.15)	0.430	0.673	0.828 (0.654-1.05)	0.116	0.310
SPM 9	1.12 (0.648-1.95)	0.676	0.821	0.903 (0.543-1.5)	0.693	0.821	1.75 (1.09-2.82)	0.020	0.112
SPM 10	1.25 (0.793-1.97)	0.337	0.583	**0.563 (0.367-0.863)**	**0.008**	**0.063**	0.786 (0.516-1.2)	0.263	0.493
SPM 11	1.47 (0.908-2.38)	0.117	0.310	1.14 (0.741-1.76)	0.549	0.789	1.1 (0.718-1.69)	0.658	0.821
SPM 12	1.1 (0.74-1.63)	0.645	0.821	1.38 (0.984-1.93)	0.062	0.215	1.28 (0.908-1.8)	0.159	0.377
SPM 13	0.721 (0.491-1.06)	0.095	0.298	0.821 (0.582-1.16)	0.261	0.493	**0.61 (0.427-0.871)**	**0.007**	**0.059**
SPM 14	1.08 (0.573-2.04)	0.809	0.910	0.884 (0.494-1.58)	0.679	0.821	1.33 (0.755-2.35)	0.323	0.581
SPM 15	0.565 (0.135-2.36)	0.434	0.673	**0.092 (0.026-0.321)**	**< 0.001**	**0.003**	**0.087 (0.025-0.302)**	**< 0.001**	**0.003**

Supplemental Table 6.3 Multivariable Cox regression for sarcoma proteome module (SPM) 6
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. SPM measure is the median score for all proteins in the SPM. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval

		MFS	
		HR (95% CI)	**p**
Age at excision (years)		0.998 (0.981-1.01)	0.792
Sex	*F (ref)*	-	-
	M	1.24 (0.8-1.91)	0.339
Histological subtype	*LMS (ref)*	-	-
	AS	**3.38 (1.42-8.05)**	**0.006**
	DDLPS	0.523 (0.215-1.27)	0.153
	EPS	**4.47 (1.64-12.2)**	**0.003**
	SS	0.792 (0.357-1.76)	0.567
	UPS	1.08 (0.585-1.99)	0.809
	Other	2.33 (0.718-7.54)	0.159
Anatomical site	*Extremity (ref)*	-	-
	Pelvic	1.19 (0.547-2.57)	0.665
	Trunk	0.659 (0.303-1.43)	0.292
	Intra-abdominal	1.28 (0.634-2.59)	0.488
	Retroperitoneal	0.741 (0.363-1.51)	0.411
	Uterine	1.64 (0.518-5.21)	0.4
	Head/neck	0.973 (0.279-3.4)	0.966
FNCLCC grade	*2 (ref)*	-	-
	3	**1.64 (1.03-2.61)**	**0.035**
	unknown	0.814 (0.3-2.21)	0.686
Performance status	*0 (ref)*	-	-
	1	**1.62 (1-2.61)**	**0.048**
	2-3	1.47 (0.639-3.39)	0.363
	unknown	1.58 (0.897-2.78)	0.114
Tumour depth	*Deep (ref)*	-	-
	Superficial	0.581 (0.32-1.05)	0.074
Tumour margin	*R1 & R2 (ref)*	-	-
	R0	1.06 (0.709-1.6)	0.763
	unknown	1.66 (0.582-4.74)	0.343
Log(Tumour size [mm])	4-5 (ref)	-	-
	< 4	**0.421 (0.228-0.779)**	**0.006**
	> 5	1.01 (0.583-1.74)	0.979
	SPM 6	**1.96 (1.19-3.25)**	**0.009**

Supplemental Table 6.4 Multivariable Cox regression for sarcoma proteome module (SPM) 10
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. SPM measure is the median score for all proteins in the SPM. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval

		MFS	
		HR (95% CI)	p
	Age at excision (years)	0.999 (0.983-1.02)	0.94
Sex	*F (ref)*	-	-
	M	1.24 (0.796-1.92)	0.345
Histological subtype	*LMS (ref)*	-	-
	AS	**3.54 (1.51-8.26)**	**0.004**
	DDLPS	0.521 (0.215-1.27)	0.15
	EPS	**4.29 (1.58-11.7)**	**0.004**
	SS	0.882 (0.391-1.99)	0.762
	UPS	1.01 (0.556-1.84)	0.971
	Other	1.85 (0.601-5.69)	0.284
Anatomical site	*Extremity (ref)*	-	-
	Pelvic	0.947 (0.432-2.08)	0.892
	Trunk	0.725 (0.342-1.54)	0.403
	Intra-abdominal	1.26 (0.628-2.53)	0.515
	Retroperitoneal	0.683 (0.335-1.39)	0.296
	Uterine	1.24 (0.396-3.91)	0.708
	Head/neck	1.03 (0.3-3.56)	0.957
FNCLCC grade	*2 (ref)*	-	-
	3	**2.12 (1.36-3.32)**	**<0.001**
	unknown	0.848 (0.31-2.32)	0.749
Performance status	*0 (ref)*	-	-
	1	**1.64 (1.02-2.63)**	**0.041**
	2-3	1.2 (0.522-2.74)	0.673
	unknown	1.44 (0.825-2.52)	0.199
Tumour depth	*Deep (ref)*	-	-
	Superficial	**0.487 (0.266-0.89)**	**0.019**
Tumour margin	*R1 & R2 (ref)*	-	-
	R0	1.15 (0.764-1.73)	0.502
	unknown	1.43 (0.498-4.08)	0.508
Log(Tumour size [mm])	4-5 (ref)	-	-
	< 4	**0.42 (0.228-0.773)**	**0.005**
	> 5	1.12 (0.649-1.93)	0.683
	SPM 10	**0.466 (0.247-0.879)**	**0.018**

Supplemental Table 6.5 Multivariable Cox regression for sarcoma proteome module (SPM) 6
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. SPM subgroups identified by tertile stratification based on median expression across the full cohort. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval

		LRFS		MFS		OS	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
Age at excision (years)		1 (0.984-1.02)	0.95	0.998 (0.982-1.01)	0.761	1.01 (0.996-1.03)	0.121
Sex	*F (ref)*	-	-	-	-	-	-
	M	1.44 (0.912-2.29)	0.117	1.3 (0.854-1.98)	0.222	**1.55 (1.03-2.34)**	**0.036**
Histological subtype	*LMS (ref)*	-	-	-	-	-	-
	AS	**6.53 (3.12-13.6)**	**<0.001**	**2.6 (1.36-4.98)**	**0.004**	**3.58 (1.85-6.91)**	**<0.001**
	DDLPS	1.74 (0.81-3.75)	0.155	**0.399 (0.175-0.909)**	**0.029**	0.732 (0.367-1.46)	0.376
	EPS	3.55 (1.07-11.8)	0.038	**4.91 (1.8-13.4)**	**0.002**	2.35 (0.739-7.44)	0.148
	SS	1.67 (0.716-3.88)	0.235	0.732 (0.347-1.54)	0.412	1.02 (0.489-2.14)	0.95
	UPS	1.28 (0.608-2.71)	0.513	0.984 (0.56-1.73)	0.955	1.09 (0.613-1.92)	0.777
	Other	1.61 (0.377-6.91)	0.518	2.23 (0.736-6.77)	0.156	1.62 (0.419-6.23)	0.486
FNCLCC grade	*2 (ref)*	-	-	-	-	-	-
	3	1.08 (0.669-1.75)	0.748	**1.72 (1.11-2.65)**	**0.014**	**1.76 (1.13-2.74)**	**0.012**
	unknown	0.945 (0.362-2.46)	0.908	0.896 (0.326-2.46)	0.832	0.941 (0.309-2.87)	0.916
Performance status	*0 (ref)*	-	-	-	-	-	-
	1	**1.81 (1.08-3.05)**	**0.025**	**1.68 (1.04-2.72)**	**0.034**	**2.11 (1.33-3.36)**	**0.002**
	2-3	1.23 (0.485-3.09)	0.668	1.37 (0.581-3.25)	0.469	**3.66 (1.86-7.18)**	**<0.001**
	unknown	1.03 (0.542-1.95)	0.933	1.55 (0.897-2.68)	0.117	1.46 (0.834-2.56)	0.185
Tumour depth	*Deep (ref)*	-	-	-	-	-	-
	Superficial	0.967 (0.521-1.79)	0.914	**0.539 (0.302-0.964)**	**0.037**	0.799 (0.457-1.4)	0.432
Tumour margin	*R1 & R2 (ref)*	-	-	-	-	-	-
	R0	0.725 (0.467-1.13)	0.153	1.13 (0.758-1.68)	0.551	1.06 (0.719-1.57)	0.766
	Rx	1.35 (0.602-3.01)	0.469	1.68 (0.656-4.31)	0.279	1.03 (0.391-2.71)	0.952
Log[tumour size(mm)]	*<4 (ref)*	-	-	-	-	-	-
	4 - 5	**2.23 (1.18-4.2)**	**0.014**	**2.32 (1.32-4.07)**	**0.003**	**2.06 (1.17-3.64)**	**0.012**
	> 5	**4.15 (1.88-9.15)**	**<0.001**	**2.24 (1.08-4.62)**	**0.029**	**3.37 (1.68-6.77)**	**<0.001**
SPM 6	*Low (ref)*	-	-	-	-	-	-
	Intermediate	1.48 (0.794-2.76)	0.217	1.75 (0.924-3.3)	0.086	1.59 (0.9-2.82)	0.11
	High	1.34 (0.659-2.74)	0.415	**2.42 (1.23-4.77)**	**0.011**	1.15 (0.621-2.14)	0.653

Supplemental Table 6.6 Multivariable Cox regression for sarcoma proteome module (SPM) 10
Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. SPM subgroups identified by tertile stratification based on median expression across the full cohort. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval

		LRFS		MFS		OS	
		HR (95% CI)	**p**	**HR (95% CI)**	**p**	**HR (95% CI)**	**p**
Age at excision (years)		0.998 (0.981-1.01)	0.781	0.996 (0.981-1.01)	0.581	1.01 (0.991-1.03)	0.341
Sex	*F (ref)*	-	-	-	-	-	-
	M	1.45 (0.919-2.3)	0.11	1.33 (0.867-2.03)	0.192	**1.65 (1.09-2.49)**	**0.019**
Histological subtype	*LMS (ref)*	-	-	-	-	-	-
	AS	**8.22 (3.88-17.4)**	**<0.001**	**2.72 (1.42-5.21)**	**0.003**	**4.16 (2.17-8)**	**<0.001**
	DDLPS	**2.28 (1.07-4.88)**	**0.033**	**0.359 (0.16-0.805)**	**0.013**	1.18 (0.593-2.34)	0.641
	EPS	**3.76 (1.15-12.2)**	**0.028**	**3.51 (1.34-9.19)**	**0.011**	2.43 (0.767-7.67)	0.131
	SS	1.93 (0.827-4.51)	0.128	0.697 (0.329-1.48)	0.347	1.21 (0.584-2.52)	0.606
	UPS	1.5 (0.704-3.21)	0.293	0.889 (0.502-1.57)	0.685	1.27 (0.715-2.25)	0.415
	Other	1.42 (0.355-5.67)	0.621	1.53 (0.514-4.55)	0.445	1.77 (0.459-6.79)	0.408
FNCLCC grade	*2 (ref)*	-	-	-	-	-	-
	3	1.27 (0.793-2.02)	0.323	**2.26 (1.47-3.47)**	**<0.001**	**2.23 (1.43-3.48)**	**<0.001**
	unknown	0.993 (0.379-2.6)	0.989	0.859 (0.317-2.33)	0.764	1.1 (0.36-3.39)	0.863
Performance status	*0 (ref)*	-	-	-	-	-	-
	1	**1.79 (1.08-2.97)**	**0.023**	**1.65 (1.02-2.66)**	**0.041**	**2.24 (1.41-3.56)**	**<0.001**
	2-3	1.09 (0.437-2.74)	0.848	1.27 (0.563-2.85)	0.567	**3.65 (1.89-7.04)**	**<0.001**
	unknown	1.01 (0.539-1.9)	0.967	1.39 (0.808-2.37)	0.236	1.56 (0.897-2.73)	0.115
Tumour depth	*Deep (ref)*	-	-	-	-	-	-
	Superficial	0.883 (0.474-1.65)	0.696	**0.471 (0.262-0.848)**	**0.012**	0.699 (0.398-1.23)	0.212
Tumour margin	*R1 & R2 (ref)*	-	-	-	-	-	-
	R0	0.75 (0.484-1.16)	0.198	1.12 (0.76-1.66)	0.557	1.08 (0.736-1.57)	0.706
	Rx	1.25 (0.532-2.92)	0.611	1.17 (0.406-3.38)	0.77	0.725 (0.213-2.47)	0.608
Log[tumour size(mm)]	*4 - 5 (ref)*	-	-	-	-	-	-
	<4	**0.488 (0.259-0.916)**	**0.026**	**0.43 (0.247-0.749)**	**0.003**	**0.564 (0.321-0.993)**	**0.047**
	> 5	**1.9 (1.1-3.27)**	**0.021**	1.02 (0.616-1.68)	0.947	1.54 (0.969-2.44)	0.068
SPM 10	*Low (ref)*	-	-	-	-	-	-
	Intermediate	0.656 (0.369-1.17)	0.151	0.79 (0.481-1.3)	0.352	0.629 (0.379-1.04)	0.073
	High	0.553 (0.29-1.05)	0.072	**0.46 (0.249-0.847)**	**0.013**	**0.432 (0.238-0.782)**	**0.006**

Supplemental Table 6.7 Univariable Cox regression for sarcoma proteome modules (SPM) in The Cancer Genome Atlas (TCGA) cohort

Local recurrence free survival (LRFS), metastasis free survival (MFS), and overall survival (OS) assessed. SPM measures are median scores for all proteins in the SPM. LMS, DDLP, UPS, and SS patients included. Significant results in bold. Abbreviations: ref = reference variable; HR = hazard ratio; CI = confidence interval

	LRFS		MFS		OS	
	HR (95% CI)	p	HR (95% CI)	p	HR (95% CI)	p
SPM 6	0.806 (0.462-1.41)	0.448	**2.94 (1.77-4.87)**	**<0.001**	1.09 (0.689-1.73)	0.713
SPM 10	2.68 (0.989-7.25)	0.053	1.32 (0.572-3.04)	0.516	**2.93 (1.27-6.76)**	**0.012**

Chapter 7 Conclusions and future directions

Our current biological understanding of STS is incomplete. This is due in part to the molecular heterogeneity observed between and within histological subtypes of STS[4,36,41,165]. This heterogeneity is reflected at the clinical level, through differential rates of disease progression, recurrence and metastasis, and disparate responses to treatment intervention[5,44]. Together, the biological and clinical heterogeneity, as well as the rarity of the STS complicates clinical management[1]. Across cancer care, the integration of molecular biology into clinical practice has been key in transforming patient outcomes[159–163]. However, this is yet to be fully realised in STS. There is a pressing need for prognostic risk stratification in STS, to identify high risk patients which may benefit from aggressive treatment regimens and/or increased monitoring. Furthermore, current treatment decisions for most adult STS patients do not integrate a molecular basis[44]. In line with this, there is a need for molecular stratification to support the use of targeted therapies. These limitations in STS care are underscored by gaps in our biological knowledge of this disease. Without comprehensive disease understanding, improvements in clinical outcomes for patients will continue to be restricted.

Whilst large-scale genomic, epigenomic, and transcriptomic studies have been performed in STS, there is no comprehensive proteomic understanding of the disease[36,41,165,207,367]. My project has aimed to tackle this gap by using MS to profile the proteome of a large, retrospective, multi-subtype STS cohort. Analyses were directed in 3 ways: 1) to compare biology between histological subtypes; 2) to compare biology within histological subtypes; and 3) to investigate biology independent of histological subtypes. At their core, these approaches aim to dissect disease heterogeneity. To reflect on the key findings of this thesis; the project aims, the extent to which they have been achieved, and the future directions associated with them are detailed below.

7.1 Aim 1: To profile the STS proteome of multiple histological subtypes

This project successfully profiled the proteome of 11 histological subtypes of STS (**Chapter 3**). For subtypes where proteomic data has previously been published, this study comprises the largest cohorts to date[483]. For other subtypes, this represents the first attempt at comprehensive proteomic profiling. Robust data acquisition was achieved through experimental optimisation and thorough data processing. Specifically, to capture a multi-subtype cohort, methodological steps were modified to handle highly vascular tumours, and samples of low tumour content. Furthermore, the inclusion of many

subtypes was facilitated by assessments of the reference sample. Quality control of the data was conducted, and appropriate normalisation procedures were implemented for the removal of batch effects. This established a robust high confidence proteomic dataset of STS.

An overview of the proteomic data was provided in **Chapter 4**. This illustrated data to capture a range of functional biology spanning key stromal components of the TME (the immune component and matrisome), as well as the adhesome and the kinome[500–503]. This covers key modalities in STS that are therapeutically targetable; the immune component with immunotherapy, and the kinome with kinase inhibitors. Furthermore, it was demonstrated that the proteome could be analysed in the context of already established gene sets to reveal broad biological signatures across the STS cohort[506–508,510,512,674]. Importantly, this work demonstrated known biology to be consistently identified, providing confidence in the ability of MS data to reflect STS biology. The top-level interpretation within **Chapter 4** also highlighted novel findings, such as the distinctive matrisome and adhesome of LMS. It therefore revealed many avenues for future research.

Further to proteomic assessments, **Chapter 4** also comprehensively profiled the clinicopathological features of the cohort. This illustrated a largely representative STS population, suggesting findings may be translated beyond this study to other patients. Key next steps in understanding the STS proteome include the profiling of advanced disease (metastatic and recurrent). Assessments of proteome stability throughout disease course will be crucial in establishing whether the proteome revelations herein (rooted in primary disease), are applicable to advanced STS patients with high clinical need. As metastatic disease is not routinely managed by surgical resection, concordant methodological developments for biopsy proteomic profiling will be required to facilitate this[44].

Additionally, in designing future analyses it is important to consider intra-tumoural heterogeneity. Herein whole tumour sections were profiled, which provided an overall proteomic profile that lacked spatial resolution. However significant intra-tumoural heterogeneity in STS is well established to exist[56,606]. The implication of not considering distinct regions of tumours is well demonstrated in LMS, where extensive heterogeneity often results in incorrect tumour grading by biopsy[56]. To move beyond bulk proteomic data, emerging single cell proteomic approaches could be deployed[675,676]. Alternatively, spatially resolved proteomic methods such as full section IHC, or matrix-assisted laser

desorption/ionization imaging MS (MALDI-IMS), an MS method whereby peptides are profiled from intact tissue sections to generate an 'image', could be used[677,678]. These methods would complement the aggregated tumour proteomic features identified by this project.

7.2 Aim 2: To investigate intra-subtype heterogeneity in LMS, DDLPS, and UPS

Within **Chapter 5**, this project conducted focused analyses on the proteome of LMS. The current literature notes molecular heterogeneity in LMS at the transcriptomic level[36,43,274,281–283]. However, no consensus molecular groups have been defined and the clinical applications of such subgrouping are currently unclear. Building on this, **Chapter 5** defined 3 proteomic subtypes of LMS, each with unique biology. Proteomic subtype heterogeneity was primarily characterised by differential immune infiltration, as well as variant expression of smooth muscle markers. Integration with clinical outcome measures illustrated the dedifferentiated subtype, characterised by low smooth muscle protein expression, to be associated with a poorer LRFS and MFS. This work suggests LMS patients may benefit from prognostic stratification based on the expression of smooth muscle proteins. Future directions for these analyses include IHC assessment of smooth muscle markers within this cohort. This would determine the feasibility of using IHC, a routine clinical diagnostic method, to identify dedifferentiated LMS patients.

Chapter 5 also investigated the immune composition of DDLPS and UPS. Clinical trials have shown potential utility for ICBs in a subset of DDLPS and UPS patients[139,140]. Moreover, many studies suggest UPS to harbour the highest level of immune infiltrate across STS subtypes[36,220]. These highly infiltrated and ICB-responsive tumours are considered 'immune hot', and therefore vulnerable to immunotherapy intervention. Notably, there is no consensus on alternative therapeutic options for the so called 'immune cold' tumours. **Chapter 5** utilised TIL IHC data, immune-targeted transcriptomic data, and the comprehensive proteomic data to reveal therapeutic options for immune cold DDLPS and UPS. This work showed CD3+ TIL low patients (immune cold) to have a poorer OS compared to CD3+ high (immune hot). Biologically, these immune cold tumours were shown to have low expression of immune checkpoint genes, and a significant enrichment of humoral immune activity, including the complement cascade. This re-frames the concept of 'immune-cold' in DDLPS and UPS, suggesting that tumour with low TIL burden still harbour an active immune component. This is immune component is simply not a TIL-mediated immune response. Inhibitors of complement are

approved and in late-stage clinical trials for non-oncology medical purposes[646–648]. This work highlights the potential for repurposing of these therapeutics for STS care. Future directions applicable to these findings include targeted proteomic profiling, for example by IHC, of complement components, and *in vitro* assessments of the response to complement inhibition in STS cell lines.

The proteome-based insights of LMS, DDLPS, and UPS described in **Chapter 5**, are an illustration of the potential clinical benefit that MS can drive. Proteomic heterogeneity was demonstrated as associated with patient outcome. Furthermore, broad proteomic features highlighted therapeutic vulnerabilities in patients with limited treatment options. The proteomic disease understanding therefore has huge potential to inform and improve the clinical management of STS. To validate these findings, the vital next steps include curation and analysis of independent LMS, DDLPS, and UPS cohorts. Currently, data herein is derived from a single institution, which may introduce bias into the cohort.

7.3 Aim 3: To assess and characterise the unbiased, protein-centric STS proteome

The pan-subtype STS proteome was characterised by an unbiased and protein-centric approach within **Chapter 6**. **Chapter 6** presented a conceptually modular proteome of STS across multiple histological subtypes, comprised of 14 SPMs. The SPMs defined and captured wide ranging functional biology, and several were identified as associated with clinical outcome. This included the novel finding that expression of vesicle transport machinery in STS held prognostic value. The revelation of this prognostic utility was made possible by the use of unbiased network analysis of the proteome, without the input of prior biological knowledge[521]. When patients were categorised based on prognostic SPMs, expression was shown to be independent of histological subtype. This suggests proteomic signatures can transcend histology and complement current strategies of patient care, which are largely directed in a histology-specific manner. The identification of groups of patients with shared tumour protein biology, can enable focussed clinical efforts to therapeutically target the molecular activity of such tumours. Alternatively, where no targeted therapies are available for the identified biological functions, this work can streamline future research efforts. Another future direction leading on from this work is to establish how these SPMs perform in the current risk stratification landscape of STS. Specifically, it would be interesting to benchmark SPM performance against other published molecular risk signatures, such as CINSARC[376].

Furthermore, the SPMs may be integrated and assessed with non-molecular risk stratification methods such as nomograms [59,73–75].

7.4 Final remarks

To conclude, my work has revealed proteome features of STS which identify histology-specific biology, characterised biological heterogeneity within and across histological subtypes, and established proteomic features which go beyond histology. Throughout this project, protein-based findings were consistently identified to be associated with clinical outcome. My project therefore demonstrates the feasibility of using proteomic biology to derive prognostic tools. In addition, this work has established an invaluable resource for the STS research community. By publicly depositing the raw proteomic data and accompanying clinicopathological annotations, this work can support orthogonal protein-level validation in a currently genomic- and transcriptomic-dominant STS research landscape. Furthermore, inherent to retrospective, large-scale profiling experiments, this project was hypothesis generating. As such, this work provides the basis for many further investigations. It is anticipated that this data will be re-mined in future studies centred on the STS proteome, and by extension will support improvements in outcomes for STS patients.

Chapter 8 References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J Clin* **68**, 7–30 (2018).
2. Ries, L. *et al.* Cancer Incidence and Survival Among Children and Adolescents - Pediatric Monograph - SEER Publications 1975-1995. *Bethesda, MD* 99–4649 (1999).
3. Lye, K. L., Nordin, N., Vidyadaran, S. & Thilakavathy, K. Mesenchymal stem cells: From stem cells to sarcomas. *Cell Biol Int* **40**, 610–618 (2016).
4. *WHO Classification of Tumours of Soft Tissue and Bone, 5th Edition*. *IARC Press* (IARC Press, 2020).
5. Katz, D., Palmerini, E. & Pollack, S. M. More Than 50 Subtypes of Soft Tissue Sarcoma: Paving the Path for Histology-Driven Treatments. *American Society of Clinical Oncology Educational Book* 925–938 (2018) doi:10.1200/edbk_205423.
6. Johnson, G. D., Smith, G., Dramis, A. & Grimer, R. J. Delays in Referral of Soft Tissue Sarcomas. *Sarcoma* **2008**, (2008).
7. Smith, G. M., Johnson, G. D., Grimer, R. J. & Wilson, S. Trends in presentation of bone and soft tissue sarcomas over 25 years: little evidence of earlier diagnosis. *Ann R Coll Surg Engl* **93**, 542 (2011).
8. Survival Rates for Soft Tissue Sarcoma (SEER). https://www.cancer.org/cancer/soft-tissue-sarcoma/detection-diagnosis-staging/survival-rates.html.
9. Survival | Soft tissue sarcoma | Cancer Research UK. https://www.cancerresearchuk.org/about-cancer/soft-tissue-sarcoma/survival.
10. Eilber, F. C. *et al.* High-grade extremity soft tissue sarcomas: factors predictive of local recurrence and its effect on morbidity and mortality. *Ann Surg* **237**, 218–226 (2003).
11. Lewis, J. J., Leung, D., Heslin, M., Woodruff, J. M. & Brennan, M. F. Association of local recurrence with subsequent survival in extremity soft tissue sarcoma. *J Clin Oncol* **15**, 646–652 (1997).
12. Trovik, C. S. *et al.* Surgical margins, local recurrence and metastasis in soft tissue sarcomas: 559 surgically-treated patients from the Scandinavian Sarcoma Group Register. *Eur J Cancer* **36**, 710–716 (2000).
13. Rodriguez, R., Rubio, R. & Menendez, P. Modeling sarcomagenesis using multipotent mesenchymal stem cells. *Cell Research 2011 22:1* **22**, 62–77 (2011).

14. Gaebler, M. *et al.* Three-Dimensional Patient-Derived In Vitro Sarcoma Models: Promising Tools for Improving Clinical Tumor Management. *Front Oncol* **7**, 1 (2017).
15. Risk Factors for Soft Tissue Sarcomas. https://www.cancer.org/cancer/soft-tissue-sarcoma/causes-risks-prevention/risk-factors.html.
16. Bhatia, K., Shiels, M. S., Berg, A. & Engels, E. A. Sarcomas other than Kaposi sarcoma occurring in immunodeficiency: interpretations from a systematic literature review. *Curr Opin Oncol* **24**, 537–546 (2012).
17. Mesri, E. A., Cesarman, E. & Boshoff, C. Kaposi's sarcoma and its associated herpesvirus. *Nat Rev Cancer* **10**, 707–719 (2010).
18. Billings, S. D., McKenney, J. K., Folpe, A. L., Hardacre, M. C. & Weiss, S. W. Cutaneous angiosarcoma following breast-conserving surgery and radiation: an analysis of 27 cases. *Am J Surg Pathol* **28**, 781–788 (2004).
19. Fodor, J. *et al.* Angiosarcoma after conservation treatment for breast carcinoma: our experience and a review of the literature. *J Am Acad Dermatol* **54**, 499–504 (2006).
20. Andersson, J. *et al.* NF1-associated gastrointestinal stromal tumors have unique clinical, phenotypic, and genotypic characteristics. *American Journal of Surgical Pathology* **29**, 1170–1176 (2005).
21. International Consensus Statement on Malignant Peripheral Nerve Sheath Tumors in Neurofibromatosis 11 | Cancer Research | American Association for Cancer Research. https://aacrjournals.org/cancerres/article/62/5/1573/509653/International-Consensus-Statement-on-Malignant.
22. Ballinger, M. L. *et al.* Monogenic and polygenic determinants of sarcoma risk: an international genetic study. *Lancet Oncol* **17**, 1261–1271 (2016).
23. Jo, V. Y. & Fletcher, C. D. M. WHO classification of soft tissue tumours: an update based on the 2013 (4th) edition. *Pathology* **46**, 95–104 (2014).
24. Helman, L. J. & Meltzer, P. Mechanisms of sarcoma development. *Nature Reviews Cancer 2003 3:9* **3**, 685–694 (2003).
25. Matushansky, I. & Maki, R. G. Mechanisms of Sarcomagenesis. *Hematology/Oncology Clinics* **19**, 427–449 (2005).
26. Damerell, V., Pepper, M. S. & Prince, S. Molecular mechanisms underpinning sarcomas and implications for current and future therapy. *Signal Transduction and Targeted Therapy 2021 6:1* **6**, 1–19 (2021).
27. Tsuji, K., Ishikawa, Y. & Imamura, T. Technique for differentiating alveolar soft part sarcoma from other tumors in paraffin-embedded tissue: comparison of

immunohistochemistry for TFE3 and CD147 and of reverse transcription polymerase chain reaction for ASPSCR1-TFE3 fusion transcript. *Hum Pathol* **43**, 356–363 (2012).

28. Kira, A. *et al.* SYT–SSX Gene Fusion as a Determinant of Morphology and Prognosis in Synovial Sarcoma. *https://doi.org/10.1056/NEJM199801153380303* **338**, 153–160 (1998).
29. Ladanyi, M. *et al.* Impact of SYT-SSX Fusion Type on the Clinical Behavior of Synovial Sarcoma: A Multi-Institutional Retrospective Study of 243 Patients 1. *Cancer Res* **62**, 135–140 (2002).
30. Skytting, B. *et al.* A Novel Fusion Gene, SYT-SSX4, in Synovial Sarcoma. *JNCI: Journal of the National Cancer Institute* **91**, 974–975 (1999).
31. Aulmann, S., Longerich, T., Schirmacher, P., Mechtersheimer, G. & Penzel, R. Detection of the ASPSCR1–TFE3 gene fusion in paraffin-embedded alveolar soft part sarcomas. *Histopathology* **50**, 881–886 (2007).
32. Mohamed, M. *et al.* Desmoplastic small round cell tumor: evaluation of reverse transcription-polymerase chain reaction and fluorescence in situ hybridization as ancillary molecular diagnostic techniques. *Virchows Archiv 2017 471:5* **471**, 631–640 (2017).
33. Noujaim, J. *et al.* The spectrum of EWSR1-rearranged neoplasms at a tertiary sarcoma centre; assessing 772 tumour specimens and the value of current ancillary molecular diagnostic modalities. *British Journal of Cancer 2017 116:5* **116**, 669–678 (2017).
34. Antonescu, C. R. *et al.* Molecular Diagnosis of Clear Cell Sarcoma: Detection of EWS-ATF1 and MITF-M Transcripts and Histopathological and Ultrastructural Analysis of 12 Cases. *The Journal of Molecular Diagnostics* **4**, 44–52 (2002).
35. Taylor, B. S. *et al.* Advances in sarcoma genomics and new therapeutic targets. *Nat Rev Cancer* **11**, 541 (2011).
36. Abeshouse, A. *et al.* Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell* **171**, 950-965.e28 (2017).
37. Persson, F. *et al.* Characterization of the 12q amplicons by high-resolution, oligonucleotide array CGH and expression analyses of a novel liposarcoma cell line. *Cancer Lett* **260**, 37–47 (2008).
38. Italiano, A. *et al.* Clinical and biological significance of CDK4 amplification in well-differentiated and dedifferentiated liposarcomas. *Clin Cancer Res* **15**, 5696–5703 (2009).

39. Nilbert, M., Rydholm, A., Mitelman, F., Meltzer, P. S. & Mandahl, N. Characterization of the 12q13-15 amplicon in soft tissue tumors. *Cancer Genet Cytogenet* **83**, 32–36 (1995).
40. Italiano, A. *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* **122**, 2233–2241 (2008).
41. Gibault, L. *et al.* New insights in sarcoma oncogenesis: a comprehensive analysis of a large series of 160 soft tissue sarcomas with complex genomics. *J Pathol* **223**, 64–71 (2011).
42. Kelleher, F. C. & Viterbo, A. Histologic and genetic advances in refining the diagnosis of 'undifferentiated pleomorphic sarcoma'. *Cancers (Basel)* **5**, 218–233 (2013).
43. Chudasama, P. *et al.* Integrative genomic and transcriptomic analysis of leiomyosarcoma. *Nat Commun* **9**, 1–15 (2018).
44. Dangoor, A. *et al.* UK guidelines for the management of soft tissue sarcomas. *Clinical Sarcoma Research 2016 6:1* **6**, 1–26 (2016).
45. Coindre, J. M. *et al.* Prognostic factors in adult patients with locally controlled soft tissue sarcoma. A study of 546 patients from the French Federation of Cancer Centers Sarcoma Group. *J Clin Oncol* **14**, 869–877 (1996).
46. Stefanovski, P. D. *et al.* Prognostic factors in soft tissue sarcomas: A study of 395 patients. *European Journal of Surgical Oncology* **28**, 153–164 (2002).
47. Park, J. O. *et al.* Predicting Outcome by Growth Rate of Locally Recurrent Retroperitoneal Liposarcoma: "The One Centimeter per Month Rule". *Ann Surg* **250**, 977 (2009).
48. Henricks, W. H., Chu, Y. C., Goldblum, J. R. & Weiss, S. W. Dedifferentiated liposarcoma: a clinicopathological analysis of 155 cases with a proposal for an expanded definition of dedifferentiation. *Am J Surg Pathol* **21**, 271–281 (1997).
49. Zhang, H. *et al.* Clinical Significance and Risk Factors of Local Recurrence in Synovial Sarcoma: A Retrospective Analysis of 171 Cases. *Front Surg* **8**, 708 (2022).
50. McCormick, D., Mentzel, T., Beham, A. & Fletcher, C. D. M. Dedifferentiated liposarcoma. Clinicopathologic analysis of 32 cases suggesting a better prognostic subgroup among pleomorphic sarcomas. *Am J Surg Pathol* **18**, 1213–1223 (1994).
51. Vos, M. *et al.* Differences in recurrence and survival of extremity liposarcoma subtypes. *Eur J Surg Oncol* **44**, 1391–1397 (2018).

52. Casali, P. G. *et al.* Soft tissue and visceral sarcomas: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology* **29**, iv51–iv67 (2018).
53. Refai, F. The Histopathological Grading Of Soft Tissue Sarcomas: A Review. *Saudi Journal of Pathology and Microbiology Abbreviated Key Title: Saudi J Pathol Microbiol* (2019) doi:10.21276/sjpm.2019.4.8.2.
54. Neuville, A., Chibon, F. & Coindre, J. M. Grading of soft tissue sarcomas: from histological to molecular assessment. *Pathology* **46**, 113–120 (2014).
55. Lin, X. *et al.* Federation Nationale des Centers de Lutte Contre le Cancer grading of soft tissue sarcomas on needle core biopsies using surrogate markers. *Hum Pathol* **56**, 147–154 (2016).
56. Schneider, N. *et al.* The Adequacy of Core Biopsy in the Assessment of Smooth Muscle Neoplasms of Soft Tissues: Implications for Treatment and Prognosis. *Am J Surg Pathol* **41**, 923–931 (2017).
57. Edge, S. B. & Compton, C. C. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann Surg Oncol* **17**, 1471–1474 (2010).
58. Amin, M. B. *et al. American Joint Committee on Cancer (AJCC). AJCC Cancer Staging Manual*. *AJCC Cancer Staging Manual* (2017).
59. Kattan, M. W., Leung, D. H. Y. & Brennan, M. F. Postoperative nomogram for 12-year sarcoma-specific death. *J Clin Oncol* **20**, 791–796 (2002).
60. Szkandera, J. *et al.* Validation of the prognostic relevance of plasma C-reactive protein levels in soft-tissue sarcoma patients. *British Journal of Cancer 2013 109:9* **109**, 2316–2322 (2013).
61. Szkandera, J. *et al.* The lymphocyte/monocyte ratio predicts poor clinical outcome and improves the predictive accuracy in patients with soft tissue sarcomas. *Int J Cancer* **135**, 362–370 (2014).
62. Szkandera, J. *et al.* The elevated pre-operative plasma fibrinogen level is an independent negative prognostic factor for cancer-specific, disease-free and overall survival in soft-tissue sarcoma patients. *J Surg Oncol* **109**, 139–144 (2014).
63. Bagaria, S. P. *et al.* Validation of a Soft Tissue Sarcoma Nomogram Using a National Cancer Registry. *Annals of Surgical Oncology 2015 22:3* **22**, 398–403 (2015).
64. Eilber, F. C. & Kattan, M. W. Sarcoma Nomogram: Validation and a Model to Evaluate Impact of Therapy. *J Am Coll Surg* **205**, (2007).

65. Mariani, L. *et al.* Validation and adaptation of a nomogram for predicting the survival of patients with extremity soft tissue sarcoma using a three-grade system. *Cancer* **103**, 402–408 (2005).
66. Shuman, A. G. *et al.* Soft tissue sarcoma of the head & neck: Nomogram validation and analysis of staging systems. *J Surg Oncol* **111**, 690–695 (2015).
67. Wong, R. X. *et al.* Applicability of the Sarculator and MSKCC nomograms to retroperitoneal sarcoma prognostication in an Asian tertiary center. *Asian J Surg* **43**, 1078–1085 (2020).
68. Ng, D. W. J. *et al.* Is the Memorial Sloan Kettering Cancer Centre (MSKCC) sarcoma nomogram useful in an Asian population? *Asia Pac J Clin Oncol* **13**, e466–e472 (2017).
69. Ferrari, A. *et al.* Adult-type soft tissue sarcomas in paediatric age: A nomogram-based prognostic comparison with adult sarcoma. *Eur J Cancer* **43**, 2691–2697 (2007).
70. Trojani, M. *et al.* Soft-tissue sarcomas of adults; study of pathological prognostic variables and definition of a histopathological grading system. *Int J Cancer* **33**, 37–42 (1984).
71. Coindre, J. M. Grading of Soft Tissue Sarcomas: Review and Update. *Arch Pathol Lab Med* **130**, 1448–1453 (2006).
72. Guillou, L. *et al.* Comparative study of the National Cancer Institute and French Federation of Cancer Centers Sarcoma Group grading systems in a population of 410 adult patients with soft tissue sarcoma. *https://doi.org/10.1200/JCO.1997.15.1.350* **15**, 350–362 (2016).
73. Callegaro, D. *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* **17**, 671–680 (2016).
74. Raut, C. P. *et al.* External validation of a multi-institutional retroperitoneal sarcoma nomogram. *Cancer* **122**, 1417–1424 (2016).
75. Gronchi, A. *et al.* Outcome prediction in primary resected retroperitoneal soft tissue sarcoma: Histology-specific overall survival and disease-free survival nomograms built on major sarcoma center data sets. *Journal of Clinical Oncology* **31**, 1649–1655 (2013).
76. SARCULATOR – Apps on Google Play. https://play.google.com/store/apps/details?id=it.digitalforest.sarculator.
77. Crago, A. M. *et al.* A prognostic nomogram for prediction of recurrence in desmoid fibromatosis. *Ann Surg* **258**, 347 (2013).

78. Alman, B. *et al.* The management of desmoid tumours: A joint global consensus-based guideline approach for adult and paediatric patients. *Eur J Cancer* **127**, 96–107 (2020).
79. Zivanovic, O. *et al.* A nomogram to predict postresection 5-year overall survival for patients with uterine leiomyosarcoma. *Cancer* **118**, 660–669 (2012).
80. Dangoor, A. *et al.* UK guidelines for the management of soft tissue sarcomas. *Clinical Sarcoma Research 2016 6:1* **6**, 1–26 (2016).
81. Coindre, J. M. *et al.* Prognostic factors in adult patients with locally controlled soft tissue sarcoma. A study of 546 patients from the French Federation of Cancer Centers Sarcoma Group. *https://doi.org/10.1200/JCO.1996.14.3.869* **14**, 869–877 (2016).
82. Stefanovski, P. D. *et al.* Prognostic factors in soft tissue sarcomas: a study of 395 patients. *Eur J Surg Oncol* **28**, 153–164 (2002).
83. Bhangu, A. A., Beard, J. A. S. & Grimer, R. J. Should Soft Tissue Sarcomas be Treated at a Specialist Centre? *Sarcoma* **8**, 1–6 (2004).
84. Broto, J. M. Advancing towards Better Cooperation for Better Sarcoma Prognoses. *Oncology* **95**, 5–10 (2018).
85. Blay, J. Y. *et al.* Surgery in reference centers improves survival of sarcoma patients: a nationwide study. *Annals of Oncology* **30**, 1143 (2019).
86. Derbel, O. *et al.* Survival impact of centralization and clinical guidelines for soft tissue sarcoma (A prospective and exhaustive population-based cohort). *PLoS One* **12**, (2017).
87. Bonvalot, S. *et al.* Preoperative radiotherapy plus surgery versus surgery alone for patients with primary retroperitoneal sarcoma (EORTC-62092: STRASS): a multicentre, open-label, randomised, phase 3 trial. *Lancet Oncol* **21**, 1366–1377 (2020).
88. Beane, J. D. *et al.* Efficacy of Adjuvant Radiation Therapy in the Treatment of Soft Tissue Sarcoma of the Extremity: 20-year Follow-Up of a Randomized Prospective Trial. *Ann Surg Oncol* **21**, 2484 (2014).
89. Alektiar, K. M. *et al.* Adjuvant radiotherapy for margin-positive high-grade soft tissue sarcoma of the extremity. *International Journal of Radiation Oncology*Biology*Physics* **48**, 1051–1058 (2000).
90. O'Sullivan, B. *et al.* Preoperative versus postoperative radiotherapy in soft-tissue sarcoma of the limbs: a randomised trial. *Lancet* **359**, 2235–2241 (2002).
91. Borden, E. C. *et al.* Randomized comparison of three adriamycin regimens for metastatic soft tissue sarcomas. *J Clin Oncol* **5**, 840–850 (1987).

92. Chang, P. & Wiernik, P. H. Combination chemotherapy with adriamycin and streptozotocin. I. Clinical results in patients with advanced sarcoma. *Clin Pharmacol Ther* **20**, 605–610 (1976).
93. Edmonson, J. H. *et al.* Randomized comparison of doxorubicin alone versus ifosfamide plus doxorubicin or mitomycin, doxorubicin, and cisplatin against advanced soft tissue sarcomas. *J Clin Oncol* **11**, 1269–1275 (1993).
94. Santoro, A. *et al.* Doxorubicin versus CYVADIC versus doxorubicin plus ifosfamide in first-line treatment of advanced soft tissue sarcomas: a randomized study of the European Organization for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group. *J Clin Oncol* **13**, 1537–1545 (1995).
95. Bramwell, V., Anderson, D. & Charette, M. Doxorubicin-based chemotherapy for the palliative treatment of adult patients with locally advanced or metastatic soft tissue sarcoma. *Cochrane Database Syst Rev* **2003**, (2003).
96. Hensley, M. L. *et al.* Gemcitabine and docetaxel in patients with unresectable leiomyosarcoma: Results of a phase II trial. *Journal of Clinical Oncology* **20**, 2824–2831 (2002).
97. Maki, R. G. *et al.* Randomized phase II study of gemcitabine and docetaxel compared with gemcitabine alone in patients with metastatic soft tissue sarcomas: results of sarcoma alliance for research through collaboration study 002 [corrected]. *J Clin Oncol* **25**, 2755–2763 (2007).
98. Seddon, B. *et al.* Gemcitabine and docetaxel versus doxorubicin as first-line treatment in previously untreated advanced unresectable or metastatic soft-tissue sarcomas (GeDDiS): a randomised controlled phase 3 trial. *Lancet Oncol* **18**, 1397–1410 (2017).
99. Linch, M., Miah, A. B., Thway, K., Judson, I. R. & Benson, C. Systemic treatment of soft-tissue sarcoma—gold standard and novel therapies. *Nat Rev Clin Oncol* **11**, 187–202 (2014).
100. le Cesne, A. *et al.* Phase II Study of ET-743 in Advanced Soft Tissue Sarcomas: A European Organisation for the Research and Treatment of Cancer (EORTC) Soft Tissue and Bone Sarcoma Group Trial. (2005) doi:10.1200/JCO.2005.01.180.
101. le Cesne, A. *et al.* A randomized phase III trial comparing trabectedin to best supportive care in patients with pre-treated soft tissue sarcoma: T-SAR, a French Sarcoma Group trial. *Ann Oncol* **32**, 1034–1044 (2021).
102. Demetri, G. D. *et al.* Efficacy and safety of trabectedin or dacarbazine for metastatic liposarcoma or leiomyosarcoma after failure of conventional

chemotherapy: Results of a phase III randomized multicenter clinical trial. *Journal of Clinical Oncology* **34**, 786–793 (2016).

103. Schöffski, P. *et al.* Eribulin versus dacarbazine in previously treated patients with advanced liposarcoma or leiomyosarcoma: a randomised, open-label, multicentre, phase 3 trial. *The Lancet* **387**, 1629–1637 (2016).
104. Italiano, A. *et al.* Comparison of doxorubicin and weekly paclitaxel efficacy in metastatic angiosarcomas. *Cancer* **118**, 3330–3336 (2012).
105. Wilding, C. P. *et al.* The landscape of tyrosine kinase inhibitors in sarcomas: looking beyond pazopanib. *Expert Rev Anticancer Ther* **19**, 971 (2019).
106. Pottier, C. *et al.* Tyrosine Kinase Inhibitors in Cancer: Breakthrough and Challenges of Targeted Therapy. *Cancers (Basel)* **12**, (2020).
107. Cohen, M. H. *et al.* Approval Summary for Imatinib Mesylate Capsules in the Treatment of Chronic Myelogenous Leukemia | Clinical Cancer Research | American Association for Cancer Research. *Clinical Cancer Research* **8**, 935–942 (2002).
108. Buchdunger, E., O'Reilly, T. & Wood, J. Pharmacology of imatinib (STI571). *Eur J Cancer* **38**, S28–S36 (2002).
109. Oppelt, P. J., Hirbe, A. C. & van Tine, B. A. Gastrointestinal stromal tumors (GISTs): point mutations matter in management, a review. *J Gastrointest Oncol* **8**, 466 (2017).
110. van Oosterom, A. T. *et al.* Safety and efficacy of imatinib (STI571) in metastatic gastrointestinal stromal tumours: a phase I study. *The Lancet* **358**, 1421–1423 (2001).
111. Eorge D Emetri, G. D. *et al.* Efficacy and Safety of Imatinib Mesylate in Advanced Gastrointestinal Stromal Tumors. *https://doi.org/10.1056/NEJMoa020461* **347**, 472–480 (2002).
112. Kumar, R. *et al.* Pharmacokinetic-pharmacodynamic correlation from mouse to human with pazopanib, a multikinase angiogenesis inhibitor with potent antitumor and antiangiogenic activity. *Mol Cancer Ther* **6**, 2012–2021 (2007).
113. van der Graaf, W. T. A. *et al.* Pazopanib for metastatic soft-tissue sarcoma (PALETTE): a randomised, double-blind, placebo-controlled phase 3 trial. *The Lancet* **379**, 1879–1886 (2012).
114. Graaf, W. T. A. van der *et al.* PALETTE: Final overall survival (OS) data and predictive factors for OS of EORTC 62072/GSK VEG110727, a randomized double-blind phase III trial of pazopanib versus placebo in advanced soft tissue sarcoma (STS) patients. *https://doi.org/10.1200/jco.2012.30.15_suppl.10009* **30**, 10009–10009 (2012).

115. Sleijfer, S. *et al.* Pazopanib, a multikinase angiogenesis inhibitor, in patients with relapsed or refractory advanced soft tissue sarcoma: a phase II study from the European organisation for research and treatment of cancer-soft tissue and bone sarcoma group (EORTC study 62043). *J Clin Oncol* **27**, 3126–3132 (2009).
116. Kasper, B. *et al.* Long-term responders and survivors on pazopanib for advanced soft tissue sarcomas: subanalysis of two European Organisation for Research and Treatment of Cancer (EORTC) clinical trials 62043 and 62072. *Annals of Oncology* **25**, 719–724 (2014).
117. Hong, D. S. *et al.* Larotrectinib in adult patients with solid tumours: a multi-centre, open-label, phase I dose-escalation study. *Ann Oncol* **30**, 325–331 (2019).
118. Laetsch, T. W. *et al.* Larotrectinib for paediatric solid tumours harbouring NTRK gene fusions: phase 1 results from a multicentre, open-label, phase 1/2 study. *Lancet Oncol* **19**, 705–714 (2018).
119. Hong, D. S. *et al.* Larotrectinib in patients with TRK fusion-positive solid tumours: a pooled analysis of three phase 1/2 clinical trials. *Lancet Oncol* **21**, 531–540 (2020).
120. Westphalen, C. B. *et al.* Genomic context of NTRK1/2/3 fusion-positive tumours from a large real-world population. *npj Precision Oncology 2021 5:1* **5**, 1–9 (2021).
121. Doebele, R. C. *et al.* Entrectinib in patients with advanced or metastatic NTRK fusion-positive solid tumours: integrated analysis of three phase 1–2 trials. *Lancet Oncol* **21**, 271 (2020).
122. A Study to Test the Effect of the Drug Larotrectinib in Adults and Children With NTRK-fusion Positive Solid Tumors. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT02576431.
123. Basket Study of Entrectinib (RXDX-101) for the Treatment of Patients With Solid Tumors Harboring NTRK 1/2/3 (Trk A/B/C), ROS1, or ALK Gene Rearrangements (Fusions). *ClinicalTrials.gov* https://www.clinicaltrials.gov/ct2/show/NCT02568267.
124. Carvajal, R. D. *et al.* Trivalent ganglioside vaccine and immunologic adjuvant versus adjuvant alone in metastatic sarcoma patients rendered disease-free by surgery: A randomized phase 2 trial. *https://doi.org/10.1200/jco.2014.32.15_suppl.10520* **32**, 10520–10520 (2014).
125. Phase I Trial of Universal Donor NK Cell Therapy in Combination With ALT803. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT02890758.
126. A Study of Emactuzumab and Atezolizumab Administered in Combination in Participants With Advanced Solid Tumors. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT02323191.

127. Trial of Intratumoral Injections of TTI-621 in Subjects With Relapsed and Refractory Solid Tumors and Mycosis Fungoides . *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT02890368.
128. Talimogene Laherparepvec and Radiation Therapy in Treating Patients With Newly Diagnosed Soft Tissue Sarcoma That Can Be Removed by Surgery. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT02923778.
129. Birdi, H. K. *et al.* Immunotherapy for sarcomas: new frontiers and unveiled opportunities. *J Immunother Cancer* **9**, e001580 (2021).
130. Kerrison, W. G. J., Lee, A. T. J., Thway, K., Jones, R. L. & Huang, P. H. Current Status and Future Directions of Immunotherapies in Soft Tissue Sarcomas. *Biomedicines* **10**, (2022).
131. Lin, Z., Wu, Z. & Luo, W. A Novel Treatment for Ewing's Sarcoma: Chimeric Antigen Receptor-T Cell Therapy. *Front Immunol* **12**, (2021).
132. Pollack, S. M. *et al.* T-cell infiltration and clonality correlate with programmed cell death protein 1 and programmed death-ligand 1 expression in patients with soft tissue sarcomas. *Cancer* **123**, 3291–3304 (2017).
133. Yan, L. *et al.* Comprehensive immune characterization and T-cell receptor repertoire heterogeneity of retroperitoneal liposarcoma. *Cancer Sci* **110**, 3038–3048 (2019).
134. van Erp, A. E. M. *et al.* Expression and clinical association of programmed cell death-1, programmed death-ligand-1 and CD8+ lymphocytes in primary sarcomas is subtype dependent. *Oncotarget* **8**, 71371 (2017).
135. Dancsok, A. R. *et al.* Expression of lymphocyte immunoregulatory biomarkers in bone and soft-tissue sarcomas. *Modern Pathology 2019 32:12* **32**, 1772–1785 (2019).
136. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830.e14 (2018).
137. Huang, A. C. & Zappasodi, R. A decade of checkpoint blockade immunotherapy in melanoma: understanding the molecular basis for immune sensitivity and resistance. *Nature Immunology 2022 23:5* **23**, 660–670 (2022).
138. Combes, A. J. *et al.* Discovering dominant tumor immune archetypes in a pan-cancer census. *Cell* **185**, 184-203.e19 (2022).
139. Tawbi, H. A. *et al.* Pembrolizumab in advanced soft-tissue sarcoma and bone sarcoma (SARC028): a multicentre, two-cohort, single-arm, open-label, phase 2 trial. *Lancet Oncol* **18**, 1493–1501 (2017).
140. Burgess, M. A. *et al.* Clinical activity of pembrolizumab (P) in undifferentiated pleomorphic sarcoma (UPS) and dedifferentiated/pleomorphic liposarcoma

(LPS): Final results of SARC028 expansion cohorts. *https://doi.org/10.1200/JCO.2019.37.15_suppl.11015* **37**, 11015–11015 (2019).

141. Wilky, B. A. *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* **20**, 837–848 (2019).
142. Liu, J. *et al.* Phase II Study of TQB2450, a Novel PD-L1 Antibody, in Combination with Anlotinib in Patients with Locally Advanced or Metastatic Soft Tissue Sarcoma. *Clinical Cancer Research* OF1–OF7 (2022) doi:10.1158/1078-0432.CCR-22-0871.
143. D'Angelo, S. P. *et al.* Nivolumab with or without ipilimumab treatment for metastatic sarcoma (Alliance A091401): two open-label, non-comparative, randomised, phase 2 trials. *Lancet Oncol* **19**, 416–426 (2018).
144. Chen, J. L. *et al.* A multicenter phase II study of nivolumab +/- ipilimumab for patients with metastatic sarcoma (Alliance A091401): Results of expansion cohorts. *https://doi.org/10.1200/JCO.2020.38.15_suppl.11511* **38**, 11511–11511 (2020).
145. Keung, E. Z. *et al.* Phase II study of neoadjuvant checkpoint blockade in patients with surgically resectable undifferentiated pleomorphic sarcoma and dedifferentiated liposarcoma. *BMC Cancer* **18**, 1–7 (2018).
146. Broto, J. M. *et al.* IMMUNOSARC: A collaborative Spanish (GEIS) and Italian (ISG) sarcoma groups phase I/II trial of sunitinib plus nivolumab in advanced soft tissue and bone sarcomas: Results of the phase II- soft-tissue sarcoma cohort. *Annals of Oncology* **30**, v684 (2019).
147. Gounder, M. *et al.* Tazemetostat in advanced epithelioid sarcoma with loss of INI1/SMARCB1: an international, open-label, phase 2 basket study. *Lancet Oncol* **21**, 1423–1432 (2020).
148. Thomas, S. *et al.* A phase I trial of panobinostat and epirubicin in solid tumors with a dose expansion in patients with sarcoma. *Ann Oncol* **27**, 947–952 (2016).
149. Forrest, S. J. *et al.* Phase II trial of olaparib in combination with ceralasertib in patients with recurrent osteosarcoma. *https://doi.org/10.1200/JCO.2021.39.15_suppl.TPS11575* **39**, TPS11575–TPS11575 (2021).
150. Ingham, M. *et al.* NCI protocol 10250: A phase II study of temozolomide and olaparib for the treatment of advanced uterine leiomyosarcoma. *https://doi.org/10.1200/JCO.2021.39.15_suppl.11506* **39**, 11506–11506 (2021).
151. Hua, H. *et al.* Targeting mTOR for cancer therapy. *Journal of Hematology & Oncology 2019 12:1* **12**, 1–19 (2019).

152. Wagner, A. J. *et al.* nab-Sirolimus for Patients With Malignant Perivascular Epithelioid Cell Tumors. *J Clin Oncol* **39**, 3660–3670 (2021).
153. The human genome. *Science* vol. 291 Preprint at https://doi.org/10.1126/SCIENCE.291.5507.1218 (2001).
154. The human genome. *Nature* vol. 409 Preprint at https://doi.org/10.1038/35057454 (2001).
155. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics 2016 17:6* **17**, 333–351 (2016).
156. The Cancer Genome Atlas Program - National Cancer Institute. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.
157. Clinical Proteomic Tumor Analysis Consortium (CPTAC). https://gdc.cancer.gov/about-gdc/contributed-genomic-data-cancer-research/clinical-proteomic-tumor-analysis-consortium-cptac.
158. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature 2010 464:7291* **464**, 993–998 (2010).
159. Lung cancer: diagnosis and management. *NICE guideline* https://www.nice.org.uk/guidance/ng122.
160. Tumour profiling tests to guide adjuvant chemotherapy decisions in early breast cancer. *NICE guideline* https://www.nice.org.uk/guidance/dg34.
161. Müller, B. M. *et al.* Quantitative determination of estrogen receptor, progesterone receptor, and HER2 mRNA in formalin-fixed paraffin-embedded tissue - a new option for predictive biomarker assessment in breast cancer. *Diagn Mol Pathol* **20**, 1–10 (2011).
162. Paik, S. *et al.* A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *New England Journal of Medicine* **351**, 2817–2826 (2004).
163. Wallden, B. *et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med Genomics* **8**, 1–14 (2015).
164. NHS England - National Genomic Test Directory. https://www.england.nhs.uk/publication/national-genomic-test-directories/.
165. Baird, K. *et al.* Gene Expression Profiling of Human Sarcomas: Insights into Sarcoma Biology. *Cancer Res* **65**, 9226–9235 (2005).
166. Bovée, J. V. M. G. & Hogendoorn, P. C. W. Molecular pathology of sarcomas: concepts and clinical implications. *Virchows Arch* **456**, 193–199 (2010).

167. Jo, V. Y. EWSR1 fusions: Ewing sarcoma and beyond. *Cancer Cytopathol* **128**, 229–231 (2020).
168. Kadoch, C. & Crabtree, G. R. Reversible disruption of mSWI/SNF (BAF) complexes by the SS18-SSX oncogenic fusion in synovial sarcoma. *Cell* **153**, 71–85 (2013).
169. McBride, M. J. *et al.* The SS18-SSX Fusion Oncoprotein Hijacks BAF Complex Targeting and Function to Drive Synovial Sarcoma. *Cancer Cell* **33**, 1128 (2018).
170. Naka, N. *et al.* Synovial Sarcoma Is a Stem Cell Malignancy. *Stem Cells* **28**, 1119–1131 (2010).
171. Nakayama, R. T. *et al.* SMARCB1 is required for widespread BAF complex-mediated activation of enhancers and bivalent promoters. *Nat Genet* **49**, 1613–1623 (2017).
172. Kadoch, C. *et al.* Proteomic and Bioinformatic Analysis of mSWI/SNF (BAF) Complexes Reveals Extensive Roles in Human Malignancy. *Nat Genet* **45**, 592 (2013).
173. Wilson, B. G. *et al.* Epigenetic antagonism between Polycomb and SWI/SNF complexes during oncogenic transformation. *Cancer Cell* **18**, 316 (2010).
174. Subramaniam, M. M. *et al.* p16INK4A (CDKN2A) gene deletion is a frequent genetic event in synovial sarcomas. *Am J Clin Pathol* **126**, 866–874 (2006).
175. Bui, N. Q. *et al.* A clinico-genomic analysis of soft tissue sarcoma patients reveals CDKN2A deletion as a biomarker for poor prognosis. *Clinical Sarcoma Research 2019 9:1* **9**, 1–11 (2019).
176. Mertens, F., Antonescu, C. R. & Mitelman, F. Gene Fusions in Soft Tissue Tumors: Recurrent and Overlapping Pathogenetic Themes. *Genes Chromosomes Cancer* **55**, 291 (2016).
177. Delespaul, L. *et al.* Recurrent TRIO fusion in nontranslocation-related sarcomas. *Clinical Cancer Research* **23**, 857–867 (2017).
178. Hall, A. Rho GTPases and the Actin Cytoskeleton. *Science (1979)* **279**, 509–514 (1998).
179. Autexier, C. & Lue, N. F. The structure and function of telomerase reverse transcriptase. *Annu Rev Biochem* **75**, 493–517 (2006).
180. Cesare, A. J. & Reddel, R. R. Alternative lengthening of telomeres: models, mechanisms and implications. *Nature Reviews Genetics 2010 11:5* **11**, 319–330 (2010).
181. Versteege, I. *et al.* Truncating mutations of hSNF5/INI1 in aggressive paediatric cancer. *Nature* **394**, 203–206 (1998).

182. Biegel, J. A. *et al.* Germ-Line and Acquired Mutations of INI1 in Atypical Teratoid and Rhabdoid Tumors1 | Cancer Research | American Association for Cancer Research. *Cancer Res* **59**, 74–79 (1999).
183. Modena, P. *et al.* SMARCB1/INI1 tumor suppressor gene is frequently inactivated in epithelioid sarcomas. *Cancer Res* **65**, 4012–4019 (2005).
184. Jamshidi, F. *et al.* The genomic landscape of epithelioid sarcoma cell lines and tumours. *J Pathol* **238**, 63–73 (2016).
185. Kalimuthu, S. N. & Chetty, R. Gene of the month: SMARCB1. *J Clin Pathol* **69**, 484 (2016).
186. Sullivan, L. M., Folpe, A. L., Pawel, B. R., Judkins, A. R. & Biegel, J. A. Epithelioid sarcoma is associated with a high percentage of SMARCB1 deletions. *Mod Pathol* **26**, 385–392 (2013).
187. le Loarer, F. *et al.* Consistent SMARCB1 Homozygous Deletions in Epithelioid Sarcoma and in a Subset of Myoepithelial Carcinomas can be Reliably Detected by FISH in Archival Material. *Genes Chromosomes Cancer* **53**, 475 (2014).
188. Italiano, A. Targeting epigenetics in sarcomas through EZH2 inhibition. *J Hematol Oncol* **13**, (2020).
189. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med* **9**, 1–14 (2017).
190. Wu, H.-X. *et al.* Tumor mutational and indel burden: a systematic pan-cancer evaluation as prognostic biomarkers. *Ann Transl Med* **7**, 640–640 (2019).
191. Cote, G. M., He, J. & Choy, E. Next-Generation Sequencing for Patients with Sarcoma: A Single Center Experience. *Oncologist* **23**, 234–242 (2018).
192. He, M. *et al.* Tumor mutation burden and checkpoint immunotherapy markers in primary and metastatic synovial sarcoma. *Hum Pathol* **100**, 15–23 (2020).
193. Zhu, N. *et al.* Genomic alterations, tumour mutation burden and prognosis of chinese cardiac sarcoma patients. *Annals of Oncology* **30**, v706 (2019).
194. Ranjan, A. & Iwakuma, T. Non-Canonical Cell Death Induced by p53. *Int J Mol Sci* **17**, (2016).
195. Lane, D. & Levine, A. P53 Research: The Past Thirty Years and the Next Thirty Years. *Cold Spring Harb Perspect Biol* **2**, (2010).
196. Leroy, B. *et al.* The TP53 website: an integrative resource centre for the TP53 mutation database and TP53 mutant analysis. *Nucleic Acids Res* **41**, D962–D969 (2013).
197. Soussi, T. & Wiman, K. G. Shaping Genetic Alterations in Human Cancer: The p53 Mutation Paradigm. *Cancer Cell* **12**, 303–312 (2007).

198. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature 2013 502:7471* **502**, 333–339 (2013).
199. Barretina, J. *et al.* Subtype-specific genomic alterations define new targets for soft tissue sarcoma therapy. *Nat Genet* **42**, 715 (2010).
200. Kanojia, D. *et al.* Genomic landscape of liposarcoma. *Oncotarget* **6**, 42429 (2015).
201. Nacev, B. A. *et al.* Clinical sequencing of soft tissue and bone sarcomas delineates diverse genomic landscapes and potential therapeutic targets. *Nature Communications 2022 13:1* **13**, 1–15 (2022).
202. Gounder, M. M. *et al.* Clinical genomic profiling in the management of patients with soft tissue and bone sarcoma. *Nature Communications 2022 13:1* **13**, 1–15 (2022).
203. Dyson, N. J. RB1: a prototype tumor suppressor and an enigma. *Genes Dev* **30**, 1492 (2016).
204. Libbrecht, S., van Dorpe, J. & Creytens, D. The Rapidly Expanding Group of RB1-Deleted Soft Tissue Tumors: An Updated Review. *Diagnostics* **11**, (2021).
205. Ogura, K. *et al.* Integrated genetic and epigenetic analysis of myxofibrosarcoma. *Nature Communications 2018 9:1* **9**, 1–11 (2018).
206. Li, G. Z. *et al.* Rb and p53-deficient Myxofibrosarcoma and Undifferentiated Pleomorphic Sarcoma Require Skp2 for Survival. *Cancer Res* **80**, 2461 (2020).
207. Steele, C. D. *et al.* Undifferentiated Sarcomas Develop through Distinct Evolutionary Pathways. *Cancer Cell* **35**, 441-456.e8 (2019).
208. Hames-Fathi, S., Nottley, S. W. G. & Pillay, N. Unravelling undifferentiated soft tissue sarcomas: insights from genomics. *Histopathology* **80**, 109–121 (2022).
209. Kiuru, M. & Busam, K. J. The NF1 gene in tumor syndromes and melanoma. *Laboratory Investigation 2017 97:2* **97**, 146–157 (2017).
210. Juhász, S., Elbakry, A., Mathes, A. & Löbrich, M. ATRX Promotes DNA Repair Synthesis and Sister Chromatid Exchange during Homologous Recombination. *Mol Cell* **71**, 11-24.e7 (2018).
211. Sarma, K. *et al.* ATRX Directs Binding of PRC2 to Xist RNA and Polycomb Targets. *Cell* **159**, 869 (2014).
212. Amorim, J. P., Santos, G., Vinagre, J. & Soares, P. The Role of ATRX in the Alternative Lengthening of Telomeres (ALT) Phenotype. *Genes (Basel)* **7**, (2016).
213. Liau, J. Y. *et al.* Comprehensive screening of alternative lengthening of telomeres phenotype and loss of ATRX expression in sarcomas. *Mod Pathol* **28**, 1545–1554 (2015).

214. Italiano, A. *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* **122**, 2233–2241 (2008).
215. Pedeutour, F. *et al.* Structure of the supernumerary ring and giant rod chromosomes in adipose tissue tumors - Pedeutour - 1999 - Genes, Chromosomes and Cancer - Wiley Online Library. *Genes Chromosomes Cancer* **24**, 30–41 (1999).
216. Somaiah, N. *et al.* Targeted next generation sequencing of well-differentiated/dedifferentiated liposarcoma reveals novel gene amplifications and mutations. *Oncotarget* **9**, 19891 (2018).
217. Koczkowska, M. *et al.* Application of high-resolution genomic profiling in the differential diagnosis of liposarcoma. *Mol Cytogenet* **10**, 1–9 (2017).
218. Mi, J.-L., Xu, M., Liu, C. & Wang, R.-S. Interactions between tumor mutation burden and immune infiltration in ovarian cancer. *Int J Clin Exp Pathol* **13**, 2513 (2020).
219. Sousa, L. M. *et al.* Tumor and Peripheral Immune Status in Soft Tissue Sarcoma: Implications for Immunotherapy. *Cancers (Basel)* **13**, (2021).
220. Smolle, M. A. *et al.* T-regulatory cells predict clinical outcome in soft tissue sarcoma patients: a clinico-pathological study. *British Journal of Cancer 2021 125:5* **125**, 717–724 (2021).
221. D'Angelo, S. P. *et al.* Prevalence of tumor-infiltrating lymphocytes and PD-L1 expression in the soft tissue sarcoma microenvironment. *Hum Pathol* **46**, 357–365 (2015).
222. Boxberg, M. *et al.* PD-L1 and PD-1 and characterization of tumor-infiltrating lymphocytes in high grade sarcomas of soft tissue–prognostic implications and rationale for immunotherapy. *Oncoimmunology* **7**, (2018).
223. Fujiwara, T. *et al.* Role of Tumor-Associated Macrophages in Sarcomas. *Cancers 2021, Vol. 13, Page 1086* **13**, 1086 (2021).
224. Dancsok, A. R. *et al.* Tumor-associated macrophages and macrophage-related immune checkpoint expression in sarcomas. *Oncoimmunology* **9**, (2020).
225. Hoffmann, M. *et al.* Robust computational reconstitution - A new method for the comparative analysis of gene expression in tissues and isolated cell fractions. *BMC Bioinformatics* **7**, 1–16 (2006).
226. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods 2015 12:5* **12**, 453–457 (2015).
227. Zhong, Y. & Liu, Z. Gene expression deconvolution in linear space. *Nature Methods 2012 9:1* **9**, 8–9 (2011).

228. Li, B. *et al.* Comprehensive analyses of tumor immunity: Implications for cancer immunotherapy. *Genome Biol* **17**, 1–16 (2016).
229. Avila Cobos, F., Vandesompele, J., Mestdagh, P. & de Preter, K. Computational deconvolution of transcriptomics data from mixed cell populations. *Bioinformatics* **34**, 1969–1979 (2018).
230. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & de Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature Communications 2020 11:1* **11**, 1–14 (2020).
231. Petitprez, F. *et al.* B cells are associated with survival and immunotherapy response in sarcoma. *Nature 2020 577:7791* **577**, 556–560 (2020).
232. Sautès-Fridman, C., Petitprez, F., Calderaro, J. & Fridman, W. H. Tertiary lymphoid structures in the era of cancer immunotherapy. *Nature Reviews Cancer 2019 19:6* **19**, 307–325 (2019).
233. Yamaguchi, K. *et al.* Helper T cell-dominant tertiary lymphoid structures are associated with disease relapse of advanced colorectal cancer. *Oncoimmunology* **9**, (2020).
234. SEER Cancer Statistics Review, 1975-2010. *National Cancer Institute* https://seer.cancer.gov/archive/csr/1975_2010/.
235. Ducimetière, F. *et al.* Incidence of Sarcoma Histotypes and Molecular Subtypes in a Prospective Epidemiological Study with Central Pathology Review and Molecular Testing. *PLoS One* **6**, e20294 (2011).
236. George, S., Serrano, C., Hensley, M. L. & Ray-Coquard, I. Soft Tissue and Uterine Leiomyosarcoma. *Journal of Clinical Oncology* **36**, 144 (2018).
237. Watanabe, K., Tajino, T., Sekiguchi, M. & Suzuki, T. h-Caldesmon as a specific marker for smooth muscle tumors. Comparison with other smooth muscle markers in bone tumors. *Am J Clin Pathol* **113**, 663–668 (2000).
238. Pathology Outlines - Leiomyosarcoma. https://www.pathologyoutlines.com/topic/uteruslms.html.
239. Pathology Outlines - Leiomyosarcoma-general. https://www.pathologyoutlines.com/topic/softtissueleiomyosarcoma.html.
240. Carvalho, J. C., Thomas, D. G. & Lucas, D. R. Cluster analysis of immunohistochemical markers in leiomyosarcoma delineates specific anatomic and gender subgroups. *Cancer* **115**, 4186–4195 (2009).
241. Fersini, F., Maselli, V., Miani, E., Palma, A. de & D'Errico, A. Misdiagnosis of leiomyosarcomas: case report and medico-legal issues. *Gynecol Pelvic Med* **4**, 41–41 (2021).

242. Varon, S., Parvataneni, R., Waetjen, E., Dunn, K. & Jacoby, V. L. Misdiagnosis of Leiomyosarcoma after Radiofrequency Ablation of Uterine Myomas. *J Minim Invasive Gynecol* **26**, 564–566 (2019).
243. Guled, M. *et al.* Differentiating soft tissue leiomyosarcoma and undifferentiated pleomorphic sarcoma: A miRNA analysis. *Genes Chromosomes Cancer* **53**, 693–702 (2014).
244. Carneiro, A. *et al.* Indistinguishable genomic profiles and shared prognostic markers in undifferentiated pleomorphic sarcoma and leiomyosarcoma: different sides of a single coin? *Laboratory Investigation 2009 89:6* **89**, 668–675 (2009).
245. Gladdy, R. A. *et al.* Predictors of Survival and Recurrence in Primary Leiomyosarcoma. *Ann Surg Oncol* **20**, 1851 (2013).
246. Schaefer, I. M. *et al.* Relationships Between Highly Recurrent Tumor Suppressor Alterations in 489 Leiomyosarcomas. *Cancer* **127**, 2666 (2021).
247. Kleinerman, R. A. *et al.* Risk of soft tissue sarcomas by individual subtype in survivors of hereditary retinoblastoma. *J Natl Cancer Inst* **99**, 24–31 (2007).
248. Venkatraman, L. *et al.* Soft tissue, pelvic, and urinary bladder leiomyosarcoma as second neoplasm following hereditary retinoblastoma. *J Clin Pathol* **56**, 233 (2003).
249. Ognjanovic, S., Olivier, M., Bergemann, T. L. & Hainaut, P. Sarcomas in TP53 germline mutation carriers: a review of the IARC TP53 database. *Cancer* **118**, 1387–1396 (2012).
250. Movva, S. *et al.* Multi-platform profiling of over 2000 sarcomas: Identification of biomarkers and novel therapeutic targets. *Oncotarget* **6**, 12234 (2015).
251. Cuppens, T. *et al.* Potential Targets' Analysis Reveals Dual PI3K/mTOR Pathway Inhibition as a Promising Therapeutic Strategy for Uterine Leiomyosarcomas-an ENITEC Group Initiative. *Clin Cancer Res* **23**, 1274–1285 (2017).
252. Gibault, L. *et al.* From PTEN loss of expression to RICTOR role in smooth muscle differentiation: complex involvement of the mTOR pathway in leiomyosarcomas and pleomorphic sarcomas. *Modern Pathology 2012 25:2* **25**, 197–211 (2011).
253. Hu, J. *et al.* Loss of DNA copy number of 10q is associated with aggressive behavior of leiomyosarcomas: a comparative genomic hybridization study. *Cancer Genet Cytogenet* **161**, 20–27 (2005).
254. Cote, G. M., He, J. & Choy, E. Next-Generation Sequencing for Patients with Sarcoma: A Single Center Experience. *Oncologist* **23**, 234–242 (2018).
255. Huang, J. & Manning, B. D. A complex interplay between Akt, TSC2, and the two mTOR complexes. *Biochem Soc Trans* **37**, 217 (2009).

256. Vanhaesebroeck, B., Stephens, L. & Hawkins, P. PI3K signalling: the path to discovery and understanding. *Nature Reviews Molecular Cell Biology 2012 13:3* **13**, 195–203 (2012).
257. Hoxhaj, G. & Manning, B. D. The PI3K–AKT network at the interface of oncogenic signalling and cancer metabolism. *Nature Reviews Cancer 2019 20:2* **20**, 74–88 (2019).
258. Tan, M. H. *et al.* Lifetime cancer risks in individuals with germline PTEN mutations. *Clin Cancer Res* **18**, 400–407 (2012).
259. Hühns, M. *et al.* PTEN mutation, loss of heterozygosity, promoter methylation and expression in colorectal carcinoma: two hits on the gene? *Oncol Rep* **31**, 2236–2244 (2014).
260. Li, Y. L., Tian, Z., Wu, D. Y., Fu, B. Y. & Xin, Y. Loss of heterozygosity on 10q23.3 and mutation of tumor suppressor gene PTEN in gastric cancer and precancerous lesions. *World Journal of Gastroenterology : WJG* **11**, 285 (2005).
261. Kwabi-Addo, B. *et al.* Haploinsufficiency of the Pten tumor suppressor gene promotes prostate cancer progression. *Proc Natl Acad Sci U S A* **98**, 11563–11568 (2001).
262. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences* **68**, 820–823 (1971).
263. Tomita, Y. *et al.* Prognostic significance of activated AKT expression in soft-tissue sarcoma. *Clin Cancer Res* **12**, 3070–3077 (2006).
264. Hernando, E. *et al.* The AKT-mTOR pathway plays a critical role in the development of leiomyosarcomas. *Nat Med* **13**, 748–753 (2007).
265. Italiano, A. *et al.* Temsirolimus in advanced leiomyosarcomas: patterns of response and correlation with the activation of the mammalian target of rapamycin pathway. *Anticancer Drugs* **22**, 463–467 (2011).
266. Schwartz, G. K. *et al.* Cixutumumab and temsirolimus for patients with bone and soft-tissue sarcoma: a multicentre, open-label, phase 2 trial. *Lancet Oncol* **14**, 371 (2013).
267. Wu, X. *et al.* Recent Advances in Dual PI3K/mTOR Inhibitors for Tumour Treatment. *Front Pharmacol* **13**, 1597 (2022).
268. Yoo, C. *et al.* Multicenter phase II study of everolimus in patients with metastatic or recurrent bone and soft-tissue sarcomas after failure of anthracycline and ifosfamide. *Invest New Drugs* **31**, 1602–1608 (2013).
269. Lee, Y. R. *et al.* Up-regulation of PI3K/Akt signaling by 17β-estradiol through activation of estrogen receptor-α, but not estrogen receptor-β, and stimulates cell

growth in breast cancer cells. *Biochem Biophys Res Commun* **336**, 1221–1226 (2005).

270. Martin, M. B. *et al.* A Role for Akt in Mediating the Estrogenic Functions of Epidermal Growth Factor and Insulin-Like Growth Factor I. *Endocrinology* **141**, 4503–4511 (2000).
271. Leitao, M. M. *et al.* Immunohistochemical expression of estrogen and progesterone receptors and outcomes in patients with newly diagnosed uterine leiomyosarcoma. *Gynecol Oncol* **124**, 558–562 (2012).
272. Bodner, K., Bodner-Adler, B., Kimberger, O., Czerwenka, K. & Mayerhofer, K. Estrogen and progesterone receptor expression in patients with uterine smooth muscle tumors. *Fertil Steril* **81**, 1062–1066 (2004).
273. Valkov, A. *et al.* Estrogen receptor and progesterone receptor are prognostic factors in soft tissue sarcomas. *Int J Oncol* **38**, 1031–1040 (2011).
274. Anderson, N. D. *et al.* Lineage-defined leiomyosarcoma subtypes emerge years before diagnosis and determine patient survival. *Nature Communications 2021 12:1* **12**, 1–14 (2021).
275. Lord, C. J. & Ashworth, A. BRCAness revisited. *Nature Reviews Cancer 2016 16:2* **16**, 110–120 (2016).
276. Seligson, N. D. *et al.* BRCA1/2 Functional Loss Defines a Targetable Subset in Leiomyosarcoma . *Oncologist* **24**, 973–979 (2019).
277. Rosenbaum, E. *et al.* Clinical Outcome of Leiomyosarcomas With Somatic Alteration in Homologous Recombination Pathway Genes. *https://doi.org/10.1200/PO.20.00122* 1350–1360 (2020) doi:10.1200/PO.20.00122.
278. Javle, M. *et al.* Olaparib Monotherapy for Previously Treated Pancreatic Cancer With DNA Damage Repair Genetic Alterations Other Than Germline BRCA Variants: Findings From 2 Phase 2 Nonrandomized Clinical Trials. *JAMA Oncol* **7**, 693 (2021).
279. Golan, T. *et al.* Phase II study of olaparib for BRCAness phenotype in pancreatic cancer. *https://doi.org/10.1200/JCO.2018.36.4_suppl.297* **36**, 297–297 (2018).
280. FDA. FDA approves olaparib for adjuvant treatment of high-risk early breast cancer. https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-olaparib-adjuvant-treatment-high-risk-early-breast-cancer.
281. Beck, A. H. *et al.* Discovery of molecular subtypes in leiomyosarcoma through integrative molecular profiling. *Oncogene* **29**, 845–854 (2010).
282. Guo, X. *et al.* Clinically Relevant Molecular Subtypes in Leiomyosarcoma. *Clin Cancer Res* **21**, 3501–11 (2015).

283. Hemming, M. L. *et al.* Oncogenic Gene-Expression Programs in Leiomyosarcoma and Characterization of Conventional, Inflammatory, and Uterogenic Subtypes. *Mol Cancer Res* **18**, 1302–1314 (2020).
284. Coosemans, A. Wilms' Tumour gene 1 (WT1) as an immunotherapeutic target. *Facts Views Vis Obgyn* **3**, 89 (2011).
285. Coosemans, A. *et al.* Upregulation of Wilms' tumour gene 1 (WT1) in uterine sarcomas. *Eur J Cancer* **43**, 1630–1637 (2007).
286. Sotobori, T. *et al.* Prognostic significance of Wilms tumor gene (WT1) mRNA expression in soft tissue sarcoma. *Cancer* **106**, 2233–2240 (2006).
287. Campbell, C. E. *et al.* CONSTITUTIVE EXPRESSION OF THE WILMS TUMOR SUPPRESSOR GENE (WT1) IN RENAL CELL CARCINOMA. *J. Cancer* **78**, 182–188 (1998).
288. Inoue, K. *et al.* Aberrant Overexpression of the Wilms Tumor Gene (WT1) in Human Leukemia. *Blood* **89**, 1405–1412 (1997).
289. Oji, Y. *et al.* Overexpression of the Wilms' tumor gene WT1 in de novo lung cancers. *Int J Cancer* **100**, 297–303 (2002).
290. Bruening, W. *et al.* Analysis of the 11p13 Wilms' Tumor Suppressor Gene (WTI) in Ovarian Tumors. *http://dx.doi.org/10.3109/07357909309018871* **11**, 393–399 (2009).
291. Wang, Z., Wang, D. Z., Pipes, G. C. T. & Olson, E. N. Myocardin is a master regulator of smooth muscle gene expression. *Proc Natl Acad Sci U S A* **100**, 7129–7134 (2003).
292. Kirik, U. *et al.* Chromatin, Gene, and RNA Regulation Discovery-Based Protein Expression Profiling Identifies Distinct Subgroups and Pathways in Leiomyosarcomas. (2014) doi:10.1158/1541-7786.MCR-14-0072.
293. Gaeta, R. *et al.* Dedifferentiated soft tissue leiomyosarcoma with heterologous osteosarcoma component: case report and review of the literature. *Clinical Sarcoma Research 2020 10:1* **10**, 1–6 (2020).
294. Chen, E., O'Connell, F. & Fletcher, C. D. M. Dedifferentiated leiomyosarcoma: clinicopathological analysis of 18 cases. *Histopathology* **59**, 1135–1143 (2011).
295. Nicolas, M. M., Tamboli, P., Gomez, J. A. & Czerniak, B. A. Pleomorphic and dedifferentiated leiomyosarcoma: clinicopathologic and immunohistochemical study of 41 cases. *Hum Pathol* **41**, 663–671 (2010).
296. Nosaka, K. A Case of Dedifferentiated Leiomyosarcoma of the Uterus. *International Journal of Pathology and Clinical Research* **2**, (2016).
297. Friedmann-Morvinski, D. & Verma, I. M. Dedifferentiation and reprogramming: origins of cancer stem cells. *EMBO Rep* **15**, 244 (2014).

298. Yuan, S., Norgard, R. J. & Stanger, B. Z. Cellular Plasticity in Cancer. *Cancer Discov* **9**, 837 (2019).
299. Boumahdi, S. & de Sauvage, F. J. The great escape: tumour cell plasticity in resistance to targeted therapy. *Nat Rev Drug Discov* **19**, 39–56 (2020).
300. Miranda, A. *et al.* Cancer stemness, intratumoral heterogeneity, and immune response across cancers. *Proc Natl Acad Sci U S A* **116**, 9020–9029 (2019).
301. Malta, T. M. *et al.* Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **173**, 338-354.e15 (2018).
302. Jung, H. *et al.* DNA methylation loss promotes immune evasion of tumours with high mutation and copy number load. *Nature Communications 2019 10:1* **10**, 1–12 (2019).
303. Toro, J. R. *et al.* Incidence patterns of soft tissue sarcomas, regardless of primary site, in the surveillance, epidemiology and end results program, 1978-2001: An analysis of 26,758 cases. *Int J Cancer* **119**, 2922–2930 (2006).
304. Thway, K. & Fisher, C. Undifferentiated and dedifferentiated soft tissue neoplasms: Immunohistochemical surrogates for differential diagnosis. *Semin Diagn Pathol* **38**, 170–186 (2021).
305. Vodanovich, D. A., Spelman, T., May, D., Slavin, J. & Choong, P. F. M. Predicting the prognosis of undifferentiated pleomorphic soft tissue sarcoma: a 20-year experience of 266 cases. *ANZ J Surg* **89**, 1045–1050 (2019).
306. Chen, S. *et al.* Undifferentiated Pleomorphic Sarcoma: Long-Term Follow-Up from a Large Institution. *Cancer Manag Res* **11**, 10001 (2019).
307. Goldblum, J. R. An approach to pleomorphic sarcomas: can we subclassify, and does it matter? *Modern Pathology 2014 27:1* **27**, S39–S46 (2014).
308. Hofvander, J. *et al.* Recurrent PRDM10 gene fusions in undifferentiated pleomorphic sarcoma. *Clinical Cancer Research* **21**, 864–869 (2015).
309. Zheng, B. *et al.* Identification of Novel Fusion Transcripts in Undifferentiated Pleomorphic Sarcomas by Transcriptome Sequencing. *Cancer Genomics Proteomics* **16**, 399 (2019).
310. Duhoux, F. P. *et al.* PRDM16 (1p36) translocations define a distinct entity of myeloid malignancies with poor prognosis but may also occur in lymphoid malignancies. *Br J Haematol* **156**, 76–88 (2012).
311. Yang, X. H. & Huang, S. PFM1 (PRDM4), a new member of the PR-domain family, maps to a tumor suppressor locus on human chromosome 12q23-q24.1. *Genomics* **61**, 319–325 (1999).
312. Casamassimi, A. *et al.* Multifaceted Role of PRDM Proteins in Human Cancer. *Int J Mol Sci* **21**, (2020).

313. Yan, Z. *et al.* Identification of recurrence-related genes by integrating microRNA and gene expression profiling of gastric cancer. *Int J Oncol* **41**, 2166–2174 (2012).
314. Tam, W. *et al.* Mutational analysis of PRDM1 indicates a tumor-suppressor role in diffuse large B-cell lymphomas. *Blood* **107**, 4090–4100 (2006).
315. Pasqualucci, L. *et al.* Inactivation of the PRDM1/BLIMP1 gene in diffuse large B cell lymphoma. *J Exp Med* **203**, 311–317 (2006).
316. di Tullio, F., Schwarz, M., Zorgati, H., Mzoughi, S. & Guccione, E. The duality of PRDM proteins: epigenetic and structural perspectives. *FEBS J* **289**, 1256–1275 (2022).
317. Carlsten, J. O. P., Zhu, X. & Gustafsson, C. M. The multitalented Mediator complex. *Trends Biochem Sci* **38**, 531–537 (2013).
318. Szilagyi, Z. & Gustafsson, C. M. Emerging roles of Cdk8 in cell cycle control. *Biochim Biophys Acta* **1829**, 916–920 (2013).
319. Wang, H., Shen, Q., Ye, L. hua & Ye, J. MED12 mutations in human diseases. *Protein Cell* **4**, 643–646 (2013).
320. Soutourina, J. Transcription regulation by the Mediator complex. *Nature Reviews Molecular Cell Biology 2017 19:4* **19**, 262–274 (2017).
321. Kämpjärvi, K. *et al.* Somatic MED12 mutations in uterine leiomyosarcoma and colorectal cancer. *Br J Cancer* **107**, 1761 (2012).
322. Chen, C. mun *et al.* Functional Significance of SRJ Domain Mutations in CITED2. *PLoS One* **7**, e46256 (2012).
323. Lau, W. M., Doucet, M., Huang, D., Weber, K. L. & Kominsky, S. L. CITED2 Modulates Estrogen Receptor Transcriptional Activity in Breast Cancer Cells. *Biochem Biophys Res Commun* **437**, 261 (2013).
324. Hofvander, J. *et al.* Undifferentiated pleomorphic sarcomas with PRDM10 fusions have a distinct gene expression profile. *J Pathol* **249**, 425–434 (2019).
325. Yoshimoto, M. *et al.* Comparative Study of Myxofibrosarcoma With Undifferentiated Pleomorphic Sarcoma: Histopathologic and Clinicopathologic Review. *Am J Surg Pathol* **44**, 87–97 (2020).
326. Silveira, S. M. *et al.* Genomic Signatures Predict Poor Outcome in Undifferentiated Pleomorphic Sarcomas and Leiomyosarcomas. *PLoS One* **8**, e67643 (2013).
327. Hélias-Rodzewicz, Z. *et al.* YAP1 and VGLL3, encoding two cofactors of TEAD transcription factors, are amplified and overexpressed in a subset of soft tissue sarcomas. *Genes Chromosomes Cancer* **49**, 1161–1171 (2010).
328. Hori, N. *et al.* Vestigial-like family member 3 (VGLL3), a cofactor for TEAD transcription factors, promotes cancer cell proliferation by activating the Hippo pathway. *Journal of Biological Chemistry* **295**, 8798–8807 (2020).

329. Halperin, D. S., Pan, C., Lusis, A. J. & Tontonoz, P. Vestigial-like 3 is an inhibitor of adipocyte differentiation. *J Lipid Res* **54**, 473–481 (2013).
330. Figeac, N. *et al.* VGLL3 operates via TEAD1, TEAD3 and TEAD4 to influence myogenesis in skeletal muscle. *J Cell Sci* **132**, (2019).
331. Zhang, Z., Lei, B., Wu, H., Zhang, X. & Zheng, N. Tumor suppressive role of miR-194-5p in glioblastoma multiforme. *Mol Med Rep* **16**, 9317–9322 (2017).
332. Yen, Y. T., Yang, J. C., Chang, J. B. & Tsai, S. C. Down-Regulation of miR-194-5p for Predicting Metastasis in Breast Cancer Cells. *Int J Mol Sci* **23**, (2021).
333. Guo, J., Zhang, J., Yang, T., Zhang, W. & Liu, M. MiR-22 suppresses the growth and metastasis of bladder cancer cells by targeting E2F3. *Int J Clin Exp Pathol* **13**, 587 (2020).
334. Wongjampa, W. *et al.* Suppression of miR-22, a tumor suppressor in cervical cancer, by human papillomavirus 16 E6 via a p53/miR-22/HDAC6 pathway. *PLoS One* **13**, (2018).
335. Jiang, X. *et al.* miR-22 has a potent anti-tumour role with therapeutic potential in acute myeloid leukaemia. *Nature Communications 2016 7:1* **7**, 1–15 (2016).
336. Wang, G., Shen, N., Cheng, L., Lin, J. & Li, K. Downregulation of miR-22 acts as an unfavorable prognostic biomarker in osteosarcoma. *Tumor Biology 2015 36:10* **36**, 7891–7895 (2015).
337. Bar, N. & Dikstein, R. miR-22 Forms a Regulatory Loop in PTEN/AKT Pathway and Modulates Signaling Kinetics. *PLoS One* **5**, e10859 (2010).
338. Toulmonde, M. *et al.* Use of PD-1 Targeting, Macrophage Infiltration, and IDO Pathway Activation in Sarcomas: A Phase 2 Clinical Trial. *JAMA Oncol* **4**, 93–97 (2018).
339. Toulmonde, M. *et al.* High throughput profiling of undifferentiated pleomorphic sarcomas identifies two main subgroups with distinct immune profile, clinical outcome and sensitivity to targeted therapies. *EBioMedicine* **62**, (2020).
340. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, (2017).
341. Gronchi, A. *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. *Ann Surg* **263**, 1002–1009 (2016).
342. Toulmonde, M. *et al.* Retroperitoneal sarcomas: patterns of care at diagnosis, prognostic factors and focus on main histological subtypes: a multicenter analysis of the French Sarcoma Group. *Ann Oncol* **25**, 735–742 (2014).

343. Singer, S., Antonescu, C. R., Riedel, E., Brennan, M. F. & Pollock, R. E. Histologic subtype and margin of resection predict pattern of recurrence and survival for retroperitoneal liposarcoma. *Ann Surg* **238**, 358–371 (2003).
344. Pathology Outlines - Dedifferentiated liposarcoma. https://www.pathologyoutlines.com/topic/softtissuedediffl ipo.html.
345. Pathology Outlines - Atypical lipomatous tumor / well differentiated liposarcoma. https://www.pathologyoutlines.com/topic/softtissuewdliposarcoma.html.
346. Barretina, J. *et al.* Subtype-specific genomic alterations define new targets for soft tissue sarcoma therapy. *Nat Genet* **42**, 715 (2010).
347. Kanojia, D. *et al.* Genomic landscape of liposarcoma. *Oncotarget* **6**, 42429 (2015).
348. Jones, R. L., Lee, A. T. J., Thway, K. & Huang, P. H. Clinical and Molecular Spectrum of Liposarcoma. *Journal of Clinical Oncology* **36**, 151 (2018).
349. Beird, H. C. *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* **4**, a002386 (2018).
350. Tap, W. D. *et al.* Evaluation of well-differentiated/de-differentiated liposarcomas by high-resolution oligonucleotide array-based comparative genomic hybridization. *Genes Chromosomes Cancer* **50**, 95–112 (2011).
351. Bill, K. L. J. *et al.* Degree of MDM2 Amplification Affects Clinical Outcomes in Dedifferentiated Liposarcoma . *Oncologist* **24**, 989–996 (2019).
352. Lee, D. S. *et al.* c-Jun regulates adipocyte differentiation via the KLF15-mediated mode. *Biochem Biophys Res Commun* **469**, 552–558 (2016).
353. Hirata, M. *et al.* Integrated exome and RNA sequencing of dedifferentiated liposarcoma. *Nature Communications 2019 10:1* **10**, 1–12 (2019).
354. Mariani, O. *et al.* JUN Oncogene Amplification and Overexpression Block Adipocytic Differentiation in Highly Aggressive Sarcomas. *Cancer Cell* **11**, 361–374 (2007).
355. Chibon, F. *et al.* ASK1 (MAP3K5) as a potential therapeutic target in malignant fibrous histiocytomas with 12q14–q15 and 6q23 amplifications. *Genes Chromosomes Cancer* **40**, 32–37 (2004).
356. Yamashita, K. *et al.* Prognostic significance of the MDM2/HMGA2 ratio and histological tumor grade in dedifferentiated liposarcoma. *Genes Chromosomes Cancer* **60**, 26–37 (2021).
357. Takahira, T. *et al.* Alterations of the RB1 gene in dedifferentiated liposarcoma. *Modern Pathology 2005 18:11* **18**, 1461–1470 (2005).

358. Lee, J. C. *et al.* Alternative lengthening of telomeres and loss of ATRX are frequent events in pleomorphic and dedifferentiated liposarcomas. *Mod Pathol* **28**, 1064–1073 (2015).
359. Amorim, J. P., Santos, G., Vinagre, J. & Soares, P. The Role of ATRX in the Alternative Lengthening of Telomeres (ALT) Phenotype. *Genes (Basel)* **7**, (2016).
360. Venturini, L., Motta, R., Gronchi, A., Daidone, M. G. & Zaffaroni, N. Prognostic relevance of ALT-associated markers in liposarcoma: a comparative analysis. *BMC Cancer* **10**, (2010).
361. Udroiu, I. & Sgura, A. Alternative Lengthening of Telomeres and Chromatin Status. *Genes (Basel)* **11**, (2020).
362. Italiano, A. *et al.* Clinical effect of molecular methods in sarcoma diagnosis (GENSARC): a prospective, multicentre, observational study. *Lancet Oncol* **17**, 532–538 (2016).
363. Gounder, M. M. *et al.* Clinical genomic profiling in the management of patients with soft tissue and bone sarcoma. *Nature Communications 2022 13:1* **13**, 1–15 (2022).
364. Perrier, L. *et al.* The cost-saving effect of centralized histological reviews with soft tissue and visceral sarcomas, GIST, and desmoid tumors: The experiences of the pathologists of the French Sarcoma Group. *PLoS One* **13**, e0193330 (2018).
365. Improving outcomes for people with sarcoma. *NICE guidelines* https://www.nice.org.uk/guidance/csg9.
366. Thway, K., Wang, J., Mubako, T. & Fisher, C. Histopathological Diagnostic Discrepancies in Soft Tissue Tumours Referred to a Specialist Centre: Reassessment in the Era of Ancillary Molecular Diagnosis. *Sarcoma* **2014**, (2014).
367. Koelsche, C. *et al.* Sarcoma classification by DNA methylation profiling. *Nature Communications 2021 12:1* **12**, 1–10 (2021).
368. Chibon, F. *et al.* Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. *Nat Med* **16**, 781–787 (2010).
369. Lesluyes, T. & Chibon, F. A global and integrated analysis of CINSARC-associated genetic defects. *Cancer Res* **80**, 5282–5290 (2020).
370. Pasquali, S. *et al.* The prognostic value of CINSARC in a randomised trial comparing histotype-tailored neoadjuvant chemotherapy versus standard chemotherapy in patients with high-risk soft-tissue sarcomas (ISG-STS 1001). *https://doi.org/10.1200/JCO.2020.38.15_suppl.e23531* **38**, e23531–e23531 (2020).

371. Frezza, A. M. *et al.* CINSARC in high-risk soft tissue sarcoma patients treated with neoadjuvant chemotherapy: Results from the ISG-STS 1001 study. *Cancer Med* **00**, 1–8 (2022).
372. Filleron, T. *et al.* Value of peri-operative chemotherapy in patients with CINSARC high-risk localized grade 1 or 2 soft tissue sarcoma: Study protocol of the target selection phase III CHIC-STS trial. *BMC Cancer* **20**, 1–8 (2020).
373. Italiano, A. *et al.* Benefit of intensified perioperative chemotherapy within high-risk CINSARC patients with resectable soft tissue sarcomas (CIRSARC). *https://doi.org/10.1200/JCO.2019.37.15_suppl.TPS11078* **37**, TPS11078–TPS11078 (2019).
374. le Guellec, S. *et al.* Validation of the Complexity INdex in SARComas prognostic signature on formalin-fixed, paraffin-embedded, soft-tissue sarcomas. *Ann Oncol* **29**, 1828–1835 (2018).
375. Lesluyes, T. *et al.* RNA sequencing validation of the Complexity INdex in SARComas prognostic signature. *Eur J Cancer* **57**, 104–111 (2016).
376. Lesluyes, T., Delespaul, L., Coindre, J. M. & Chibon, F. The CINSARC signature as a prognostic marker for clinical outcome in multiple neoplasms. *Sci Rep* **7**, (2017).
377. Lagarde, P. *et al.* Chromosome instability accounts for reverse metastatic outcomes of pediatric and adult synovial sarcomas. *J Clin Oncol* **31**, 608–615 (2013).
378. Crombé, A. *et al.* Gene expression profiling improves prognostication by nomogram in patients with soft-tissue sarcomas. *Cancer Commun* **42**, 563–566 (2022).
379. Sotiriou, C. *et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* **98**, 262–272 (2006).
380. Bertucci, F. *et al.* The Genomic Grade Index predicts postoperative clinical outcome in patients with soft-tissue sarcoma. *Annals of Oncology* **29**, 459–465 (2018).
381. Hendrickx, W. *et al.* Identification of genetic determinants of breast cancer immune phenotypes by integrative genome-scale analysis. *https://doi.org/10.1080/2162402X.2016.1253654* **6**, (2017).
382. Bertucci, F. *et al.* Immunologic constant of rejection signature is prognostic in soft-tissue sarcoma and refines the CINSARC signature. *J Immunother Cancer* **10**, e003687 (2022).

383. Merry, E., Thway, K., Jones, R. L. & Huang, P. H. Predictive and prognostic transcriptomic biomarkers in soft tissue sarcomas. *NPJ Precis Oncol* **5**, (2021).
384. Forker, L. *et al.* The hypoxia marker CAIX is prognostic in the UK phase III VorteX-Biobank cohort: an important resource for translational research in soft tissue sarcoma. *Br J Cancer* **118**, 698–704 (2018).
385. Nordsmark, M. *et al.* Hypoxia in human soft tissue sarcomas: Adverse impact on survival and no association with p53 mutations. *British Journal of Cancer 2001 84:8* **84**, 1070–1075 (2001).
386. Toustrup, K. *et al.* Gene expression classifier predicts for hypoxic modification of radiotherapy with nimorazole in squamous cell carcinomas of the head and neck. *Radiotherapy and Oncology* **102**, 122–129 (2012).
387. Aggerholm-Pedersen, N. *et al.* A prognostic profile of hypoxia-induced genes for localised high-grade soft tissue sarcoma. *British Journal of Cancer 2016 115:9* **115**, 1096–1104 (2016).
388. Yang, L. *et al.* Validation of a hypoxia related gene signature in multiple soft tissue sarcoma cohorts. *Oncotarget* **9**, 3946 (2018).
389. Schwaederle, M. *et al.* Association of Biomarker-Based Treatment Strategies With Response Rates and Progression-Free Survival in Refractory Malignant Neoplasms: A Meta-analysis. *JAMA Oncol* **2**, 1452–1459 (2016).
390. Schwaederle, M. *et al.* Impact of Precision Medicine in Diverse Cancers: A Meta-Analysis of Phase II Clinical Trials. *Journal of Clinical Oncology* **33**, 3817 (2015).
391. Ray-Coquard, I. *et al.* Effect of the MDM2 antagonist RG7112 on the P53 pathway in patients with MDM2-amplified, well-differentiated or dedifferentiated liposarcoma: an exploratory proof-of-mechanism study. *Lancet Oncol* **13**, 1133–1140 (2012).
392. Konopleva, M. *et al.* MDM2 inhibition: an important step forward in cancer therapy. *Leukemia 2020 34:11* **34**, 2858–2874 (2020).
393. Italiano, A., Bellera, C. & D'Angelo, S. PD1/PD-L1 targeting in advanced soft-tissue sarcomas: A pooled analysis of phase II trials. *J Hematol Oncol* **13**, 1–4 (2020).
394. Keung, E. Z. *et al.* Correlative analyses of the SARC028 trial reveal an association between sarcoma-associated immune infiltrate and response to pembrolizumab. *Clinical Cancer Research* **26**, 1258–1266 (2020).
395. Zhou, J. *et al.* Soluble PD-L1 as a biomarker in malignant melanoma treated with checkpoint blockade. *Cancer Immunol Res* **5**, 480–492 (2017).

396. Abu Hejleh, T., Furqan, M., Ballas, Z. & Clamon, G. The clinical significance of soluble PD-1 and PD-L1 in lung cancer. *Crit Rev Oncol Hematol* **143**, 148–152 (2019).
397. O'Malley, D. M. *et al.* LBA34 Single-agent anti-PD-1 balstilimab or in combination with anti-CTLA-4 zalifrelimab for recurrent/metastatic (R/M) cervical cancer (CC): Preliminary results of two independent phase II trials. *Annals of Oncology* **31**, S1164–S1165 (2020).
398. Chung, H. C. *et al.* Pembrolizumab treatment of advanced cervical cancer: Updated results from the phase 2 KEYNOTE-158 study. *https://doi.org/10.1200/JCO.2018.36.15_suppl.5522* **36**, 5522–5522 (2018).
399. Mehra, R. *et al.* Efficacy and safety of pembrolizumab in recurrent/metastatic head and neck squamous cell carcinoma: pooled analyses after long-term follow-up in KEYNOTE-012. *British Journal of Cancer 2018 119:2* **119**, 153–159 (2018).
400. Kefford, R. *et al.* Clinical efficacy and correlation with tumor PD-L1 expression in patients (pts) with melanoma (MEL) treated with the anti-PD-1 monoclonal antibody MK-3475. *https://doi.org/10.1200/jco.2014.32.15_suppl.3005* **32**, 3005–3005 (2014).
401. Wendel Naumann, R. *et al.* Safety and Efficacy of Nivolumab Monotherapy in Recurrent or Metastatic Cervical, Vaginal, or Vulvar Carcinoma: Results From the Phase I/II CheckMate 358 Trial. *Journal of Clinical Oncology* **37**, 2825 (2019).
402. Incorvaia, L. *et al.* Programmed Death Ligand 1 (PD-L1) as a Predictive Biomarker for Pembrolizumab Therapy in Patients with Advanced Non-Small-Cell Lung Cancer (NSCLC). *Adv Ther* **36**, 2600 (2019).
403. D'Angelo, S. P. *et al.* Prevalence of tumor-infiltrating lymphocytes and PD-L1 expression in the soft tissue sarcoma microenvironment. *Hum Pathol* **46**, 357–365 (2015).
404. Que, Y. *et al.* PD-L1 Expression Is Associated with FOXP3+ Regulatory T-Cell Infiltration of Soft Tissue Sarcoma and Poor Patient Prognosis. *J Cancer* **8**, 2018 (2017).
405. Bertucci, F. *et al.* PDL1 expression is a poor-prognosis factor in soft-tissue sarcomas. *Oncoimmunology* **6**, (2017).
406. Italiano, A. *et al.* Pembrolizumab in soft-tissue sarcomas with tertiary lymphoid structures: a phase 2 PEMBROSARC trial cohort. *Nature Medicine 2022 28:6* **28**, 1199–1206 (2022).
407. Bintrafusp Alfa and Doxorubicin Hydrochloride in Treating Patients With Advanced Sarcoma. *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT04874311.

408. Neoadjuvant Chemotherapy and Retifanlimab in Patients With Selected Retroperitoneal Sarcomas (TORNADO). *ClinicalTrials.gov* https://clinicaltrials.gov/ct2/show/NCT04968106.
409. Goodman, A. M. *et al.* Tumor Mutational Burden as an Independent Predictor of Response to Immunotherapy in Diverse Cancers. *Mol Cancer Ther* **16**, 2598 (2017).
410. Johnson, D. B. *et al.* Targeted next generation sequencing identifies markers of response to PD-1 blockade. *Cancer Immunol Res* **4**, 959 (2016).
411. Snyder, A. *et al.* Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N Engl J Med* **371**, 2189 (2014).
412. Dudley, J. C., Lin, M. T., Le, D. T. & Eshleman, J. R. Microsatellite Instability as a Biomarker for PD-1 Blockade. *Clin Cancer Res* **22**, 813–820 (2016).
413. Zuo, W. & Zhao, L. Recent advances and application of PD-1 blockade in sarcoma. *Onco Targets Ther* **12**, 6887 (2019).
414. Eso, Y., Shimizu, T., Takeda, H., Takai, A. & Marusawa, H. Microsatellite instability and immune checkpoint inhibitors: toward precision medicine against gastrointestinal and hepatobiliary cancers. *J Gastroenterol* **55**, 15–26 (2020).
415. Zou, X. L. *et al.* Prognostic Value of Neoantigen Load in Immune Checkpoint Inhibitor Therapy for Cancer. *Front Immunol* **12**, 5319 (2021).
416. Mardis, E. R. Neoantigens and genome instability: Impact on immunogenomic phenotypes and immunotherapy response. *Genome Med* **11**, 1–12 (2019).
417. Campanella, N. C. *et al.* Absence of Microsatellite Instability In Soft Tissue Sarcomas. *Pathobiology* **82**, 36–42 (2015).
418. Lam, S. W. *et al.* Mismatch repair deficiency is rare in bone and soft tissue tumors. *Histopathology* **79**, 509–520 (2021).
419. Saito, T. *et al.* Possible association between tumor-suppressor gene mutations and hMSH2/hMLH1 inactivation in alveolar soft part sarcoma. *Hum Pathol* **34**, 841–849 (2003).
420. Engel, C. & Fischer, C. Breast Cancer Risks and Risk Prediction Models. *Breast Care* **10**, 7 (2015).
421. Bell, D. *et al.* Integrated Genomic Analyses of Ovarian Carcinoma. *Nature* **474**, 609 (2011).
422. Armstrong, N., Ryder, S., Forbes, C., Ross, J. & Quek, R. G. W. A systematic review of the international prevalence of BRCA mutation in breast cancer. *Clin Epidemiol* **11**, 543 (2019).
423. Neff, R. T., Senter, L. & Salani, R. BRCA mutation in ovarian cancer: testing, implications and treatment considerations. *Ther Adv Med Oncol* **9**, 519 (2017).

424. Li, H. *et al.* Molecular signatures of BRCAness analysis identifies PARP inhibitor Niraparib as a novel targeted therapeutic strategy for soft tissue Sarcomas. *Theranostics* **10**, 9477 (2020).
425. Kovac, M. *et al.* Exome sequencing of osteosarcoma reveals mutation signatures reminiscent of BRCA deficiency. *Nature Communications 2015 6:1* **6**, 1–9 (2015).
426. Holme, H. *et al.* Chemosensitivity profiling of osteosarcoma tumour cell lines identifies a model of BRCAness. *Scientific Reports 2018 8:1* **8**, 1–9 (2018).
427. Engert, F., Kovac, M., Baumhoer, D., Nathrath, M. & Fulda, S. Osteosarcoma cells with genetic signatures of BRCAness are susceptible to the PARP inhibitor talazoparib alone or in combination with chemotherapeutics. *Oncotarget* **8**, 48794–48806 (2017).
428. Zoumpoulidou, G. *et al.* Therapeutic vulnerability to PARP1,2 inhibition in RB1-mutant osteosarcoma. *Nature Communications 2021 12:1* **12**, 1–16 (2021).
429. Zalenski, A. A. & Venere, M. Capitalizing on ATRX loss in glioma via PARP inhibition: Comment on “Loss of ATRX confers DNA repair defects and PARP inhibitor sensitivity” by Garbarino et al. *Transl Oncol* **14**, (2021).
430. Garbarino, J., Eckroate, J., Sundaram, R. K., Jensen, R. B. & Bindra, R. S. Loss of ATRX confers DNA repair defects and PARP inhibitor sensitivity. *Transl Oncol* **14**, (2021).
431. Reichert, Z. R., Daignault, S., Teply, B. A., Devitt, M. E. & Heath, E. I. Targeting resistant prostate cancer with ATR and PARP inhibition (TRAP trial): A phase II study. *https://doi.org/10.1200/JCO.2020.38.6_suppl.TPS254* **38**, TPS254–TPS254 (2020).
432. Forrest, S. J. *et al.* Phase II trial of olaparib in combination with ceralasertib in patients with recurrent osteosarcoma. *https://doi.org/10.1200/JCO.2021.39.15_suppl.TPS11575* **39**, TPS11575–TPS11575 (2021).
433. ClinicalTrials.gov. Combination ATR and PARP Inhibitor (CAPRI) Trial With AZD6738 and Olaparib in Recurrent Ovarian Cancer. https://clinicaltrials.gov/ct2/show/NCT03462342.
434. ClinicalTrials.gov. Olaparib With Ceralasertib in Recurrent Osteosarcoma. https://clinicaltrials.gov/ct2/show/NCT04417062.
435. ClinicalTrials.gov. Targeted Therapy Directed by Genetic Testing in Treating Patients With Advanced Refractory Solid Tumors, Lymphomas, or Multiple Myeloma (The MATCH Screening Trial) . https://clinicaltrials.gov/ct2/show/NCT02465060.

436. ClinicalTrials.gov. TAPUR: Testing the Use of Food and Drug Administration (FDA) Approved Drugs That Target a Specific Abnormality in a Tumor Gene in People With Advanced Stage Cancer. https://clinicaltrials.gov/ct2/show/NCT02693535.
437. NCI. NCI-MATCH Precision Medicine Clinical Trial. https://www.cancer.gov/about-cancer/treatment/clinical-trials/nci-supported/nci-match.
438. ECOG-ACRIN Cancer Research Group. NCI-ComboMATCH Background and Introduction. https://ecog-acrin.org/nci-combomatch/.
439. Flaherty, K. T. *et al.* Molecular Landscape and Actionable Alterations in a Genomically Guided Cancer Clinical Trial: National Cancer Institute Molecular Analysis for Therapy Choice (NCI-MATCH). *Journal of Clinical Oncology* **38**, 3883 (2020).
440. Chae, Y. K. *et al.* Phase II Study of AZD4547 in Patients With Tumors Harboring Aberrations in the FGFR Pathway: Results From the NCI-MATCH Trial (EAY131) Subprotocol W. *Journal of Clinical Oncology* **38**, 2407 (2020).
441. Krop, I. E. *et al.* Phase II Study of Taselisib in PIK3CA-Mutated Solid Tumors Other Than Breast and Squamous Lung Cancer: Results From the NCI-MATCH ECOG-ACRIN Trial (EAY131) Subprotocol I. *JCO Precis Oncol* **6**, (2022).
442. Damodaran, S. *et al.* Phase II Study of Copanlisib in Patients with Tumors with PIK3CA Mutations: Results from the NCI-MATCH ECOG-ACRIN Trial (EAY131) Subprotocol Z1F. *Journal of Clinical Oncology* **40**, 1552–1561 (2022).
443. Jour, G. *et al.* Molecular profiling of soft tissue sarcomas using next-generation sequencing: a pilot study toward precision therapeutics. *Hum Pathol* **45**, 1563–1571 (2014).
444. Groisberg, R. *et al.* Clinical genomic profiling to identify actionable alterations for investigational therapies in patients with diverse sarcomas. *Oncotarget* **8**, 39254–39267 (2017).
445. Gounder, M. M. *et al.* Impact of next-generation sequencing (NGS) on diagnostic and therapeutic options in soft-tissue and bone sarcoma. *https://doi.org/10.1200/JCO.2017.35.15_suppl.11001* **35**, 11001–11001 (2017).
446. Italiano, A. *et al.* Molecular profiling of advanced soft-tissue sarcomas: the MULTISARC randomized trial. *BMC Cancer* **21**, (2021).
447. Italiano, A. *et al.* Molecular profiling of advanced soft-tissue sarcomas: the MULTISARC randomized trial. *BMC Cancer* **21**, (2021).
448. Gry, M. *et al.* Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics* **10**, 1–14 (2009).

449. Payne, S. H. The utility of protein and mRNA correlation. *Trends Biochem Sci* **40**, 1 (2015).
450. Alberts, B. *et al. Molecular Biology of the Cell*. (Garland Science, 2002).
451. Ramazi, S. & Zahiri, J. Post-translational modifications in proteins: resources, tools and prediction methods. *Database* **2021**, (2021).
452. Burns, J., Wilding, C. P., L Jones, R. & H Huang, P. Proteomic research in sarcomas – current status and future opportunities. *Seminars in Cancer Biology* vol. 61 56–70 Preprint at https://doi.org/10.1016/j.semcancer.2019.11.003 (2020).
453. de Matos, L. L., Trufelli, D. C., de Matos, M. G. L. & Pinhal, M. A. da S. Immunohistochemistry as an Important Tool in Biomarkers Detection and Clinical Practice. *Biomark Insights* **5**, 9 (2010).
454. Kalebi, A. Y. & Dada, M. A. Application of immunohistochemistry in clinical practice: a review. *East Afr Med J* **84**, 389–397 (2007).
455. Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery 2016 16:1* **16**, 19–34 (2016).
456. Omenn, G. S. *et al.* Progress Identifying and Analyzing the Human Proteome: 2021 Metrics from the HUPO Human Proteome Project. *J Proteome Res* **20**, 5227–5240 (2021).
457. Cho, W. C. S. Proteomics technologies and challenges. *Genomics Proteomics Bioinformatics* **5**, 77–85 (2007).
458. Steen, H. & Mann, M. The abc's (and xyz's) of peptide sequencing. *Nat Rev Mol Cell Biol* **5**, 699–711 (2004).
459. Chen, Z., Dodig-Crnković, T., Schwenk, J. M. & Tao, S. C. Current applications of antibody microarrays. *Clinical Proteomics 2018 15:1* **15**, 1–15 (2018).
460. MacBeath, G. & Schreiber, S. L. Printing proteins as microarrays for high-throughput function determination. *Science* **289**, 1760–3 (2000).
461. Bordeaux, J. *et al.* Antibody validation. *Biotechniques* **48**, 197–209 (2010).
462. Vidova, V. & Spacil, Z. A review on mass spectrometry-based quantitative proteomics: Targeted and data independent acquisition. *Anal Chim Acta* **964**, 7–23 (2017).
463. Lai, X., Wang, L. & Witzmann, F. A. Issues and Applications in Label-Free Quantitative Mass Spectrometry. *Int J Proteomics* **2013**, 1–13 (2013).
464. O'Connell, J. D., Paulo, J. A., O'Brien, J. J. & Gygi, S. P. Proteome-Wide Evaluation of Two Common Protein Quantification Methods. *J Proteome Res* **17**, 1934–1942 (2018).
465. Li, J. *et al.* TMTpro-18plex: The Expanded and Complete Set of TMTpro Reagents for Sample Multiplexing. *J Proteome Res* **20**, 2964–2972 (2021).

466. Brenes, A., Hukelmann, J., Bensaddek, D. & Lamond, A. I. Multibatch TMT reveals false positives, batch effects and missing values. *Molecular and Cellular Proteomics* **18**, 1967–1980 (2019).
467. Wei, R. *et al.* Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports 2018 8:1* **8**, 1–10 (2018).
468. Jin, L. *et al.* A comparative study of evaluating missing value imputation methods in label-free proteomics. *Scientific Reports 2021 11:1* **11**, 1–11 (2021).
469. Bø, T. H., Dysvik, B. & Jonassen, I. LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res* **32**, e34 (2004).
470. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
471. Gillet, L. C. *et al.* Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: A new concept for consistent and accurate proteome analysis. *Molecular and Cellular Proteomics* **11**, (2012).
472. Krasny, L. & Huang, P. H. Data-independent acquisition mass spectrometry (DIA-MS) for proteomic applications in oncology. *Mol Omics* **17**, 29–42 (2021).
473. TMT10plex Mass Tag Labeling Kits and Reagents Instructions. Preprint at https://tools.thermofisher.com/content/sfs/manuals/MAN0016969_2162457_TMT10plex_UG.pdf.
474. Chen, F., Chandrashekar, D. S., Varambally, S. & Creighton, C. J. Pan-cancer molecular subtypes revealed by mass-spectrometry-based proteomic characterization of more than 500 human cancers. *Nature Communications 2019 10:1* **10**, 1–15 (2019).
475. Krug, K. *et al.* Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. *Cell* **183**, 1436-1456.e31 (2020).
476. Cao, L. *et al.* Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell* **184**, 5031-5052.e26 (2021).
477. Gao, Y. *et al.* Quantitative proteomics by SWATH-MS reveals sophisticated metabolic reprogramming in hepatocellular carcinoma tissues. *Scientific Reports 2017 7:1* **7**, 1–12 (2017).
478. Bouchal, P., Schubert, O. T., Budinska, E., Nenutil, R. & Aebersold, R. Breast Cancer Classification Based on Proteotypes Obtained by SWATH Mass Spectrometry. *Cell Rep* **28**, 832–843 (2019).
479. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nature Methods 2019 17:1* **17**, 41–44 (2019).

480. Metz, B. *et al.* Identification of Formaldehyde-induced Modifications in Proteins: REACTIONS WITH MODEL PEPTIDES *. *Journal of Biological Chemistry* **279**, 6235–6243 (2004).
481. Balls, A. K. *et al.* The action of formaldehyde on the cystine disulphide linkages of wool2. The conversion of subfraction A of the combined cystine into combined lanthionine and djenkolic acid and subfraction B into combined thiazolidine-4-carboxylic acid. *Biochemical Journal* **41**, 218–223 (1947).
482. Fraenkel-Conrat, H. & Olcott, H. S. The Reaction of Formaldehyde with Proteins. V. Cross-linking between Amino and Primary Amide or Guanidyl Groups. *J Am Chem Soc* **70**, 2673–2684 (1948).
483. Milighetti, M. *et al.* Proteomic profiling of soft tissue sarcomas with SWATH mass spectrometry. *J Proteomics* **241**, (2021).
484. Stewart, E. *et al.* Identification of Therapeutic Targets in Rhabdomyosarcoma Through Integrated Genomic, Epigenomic, and Proteomic Analyses. *Cancer Cell* **34**, 411 (2018).
485. Meng, X., Gao, J. Z., Gomendoza, S. M. T., Li, J. W. & Yang, S. Recent Advances of WEE1 Inhibitors and Statins in Cancers With p53 Mutations. *Front Med (Lausanne)* **8**, 1703 (2021).
486. Do, K., Doroshow, J. H. & Kummar, S. Wee1 kinase as a target for cancer therapy. *Cell Cycle* **12**, 3159 (2013).
487. Ghelli Luserna Di Rorà, A., Cerchione, C., Martinelli, G. & Simonetti, G. A WEE1 family business: regulation of mitosis, cancer progression, and therapeutic target. *Journal of Hematology & Oncology 2020 13:1* **13**, 1–17 (2020).
488. Liu, Y. *et al.* Proteomic Maps of Human Gastrointestinal Stromal Tumor Subgroups. *Mol Cell Proteomics* **18**, 923 (2019).
489. Lessard, L., Stuible, M. & Tremblay, M. L. The two faces of PTP1B in cancer. *Biochim Biophys Acta* **1804**, 613–619 (2010).
490. Nakamura, Y. *et al.* Role of protein tyrosine phosphatase 1B in vascular endothelial growth factor signaling and cell-cell adhesions in endothelial cells. *Circ Res* **102**, 1182–1191 (2008).
491. Wang, B. *et al.* Quantitative proteomic analysis of aberrant expressed lysine acetylation in gastrointestinal stromal tumors. *Clin Proteomics* **18**, 1–15 (2021).
492. Xia, C., Tao, Y., Li, M., Che, T. & Qu, J. Protein acetylation and deacetylation: An important regulatory modification in gene transcription (Review). *Exp Ther Med* **20**, 2923–2940 (2020).

493. Drazic, A., Myklebust, L. M., Ree, R. & Arnesen, T. The world of protein acetylation. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1864**, 1372–1401 (2016).
494. Liang, Y. M., Li, X. H., Li, W. M. & Lu, Y. Y. Prognostic significance of PTEN, Ki-67 and CD44s expression patterns in gastrointestinal stromal tumors. *World Journal of Gastroenterology : WJG* **18**, 1664 (2012).
495. Belev, B. *et al.* Role of Ki-67 as a prognostic factor in gastrointestinal stromal tumors. *World Journal of Gastroenterology : WJG* **19**, 523 (2013).
496. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat Methods* **6**, 359–362 (2009).
497. R Core Team. R: a language and environment for statistical computing. *R Foundation for Statistical Computing* http://www.r-project.org/ (2018).
498. Hastie, T., Tibshirani, R., Narasimhan, B. & Chu, G. impute: Imputation for microarray data. Preprint at (2020).
499. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116–5121 (2001).
500. Bhattacharya, S. *et al.* ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Sci Data* **5**, (2018).
501. Naba, A. *et al.* The matrisome: in silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Mol Cell Proteomics* **11**, M111.014647 (2012).
502. Winograd-Katz, S. E., Fässler, R., Geiger, B. & Legate, K. R. The integrin adhesome: From genes and proteins to human disease. *Nat Rev Mol Cell Biol* **15**, 273–288 (2014).
503. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* vol. 298 1912–1934 Preprint at https://doi.org/10.1126/science.1075762 (2002).
504. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545–15550 (2005).
505. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
506. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* vol. 25 25–29 Preprint at https://doi.org/10.1038/75556 (2000).
507. The Gene Ontology resource: enriching a GOld mine. doi:10.1093/nar/gkaa1113.

508. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* **1**, 417–425 (2015).
509. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature 2009 462:7269* **462**, 108–112 (2009).
510. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
511. Yoo, M. *et al.* DSigDB: drug signatures database for gene set analysis. *Bioinformatics* **31**, 3069 (2015).
512. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
513. Hinton, G. & Roweis, S. Stochastic Neighbor Embedding.
514. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw* **3**, 861 (2018).
515. Jolliffe, I. T. *Principal Component Analysis*. *S* (Springer-Verlag, 2002). doi:10.1007/B98835.
516. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).
517. Wilkerson, D., M., Hayes & Neil, D. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010).
518. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* **20**, 53–65 (1987).
519. Xu, T. *et al.* CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation, and visualization. *Bioinformatics* (2017).
520. Liu, Y., Hayes, D. N., Nobel, A. & Marron, J. S. Statistical Significance of Clustering for High-Dimension, Low–Sample Size Data. *https://doi.org/10.1198/016214508000000454* **103**, 1281–1293 (2012).
521. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 559 (2008).
522. Complement and coagulation cascades (Homo sapiens) - WikiPathways. https://www.wikipathways.org/index.php/Pathway:WP558.
523. Huang, C. *et al.* Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. *Cancer Cell* **39**, 361-379.e16 (2021).
524. Satpathy, S. *et al.* A proteogenomic portrait of lung squamous cell carcinoma. *Cell* **184**, 4348-4371.e40 (2021).

525. Gillette, M. A. *et al.* Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell* **182**, 200-225.e35 (2020).
526. Wang, L. B. *et al.* Proteogenomic and metabolomic characterization of human glioblastoma. *Cancer Cell* **39**, 509-528.e20 (2021).
527. Dou, Y. *et al.* Proteogenomic Characterization of Endometrial Carcinoma. *Cell* **180**, 729-748.e26 (2020).
528. Clark, D. J. *et al.* Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell* **179**, 964-983.e31 (2019).
529. Vasaikar, S. *et al.* Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell* **177**, 1035-1049.e19 (2019).
530. Stratton, K. G. *et al.* PmartR: Quality Control and Statistics for Mass Spectrometry-Based Biological Data. *J Proteome Res* **18**, 1418–1425 (2019).
531. Pfister, R., Schwarz, K. A., Janczyk, M., Dale, R. & Freeman, J. B. Good things peak in pairs: A note on the bimodality coefficient. *Front Psychol* **4**, 700 (2013).
532. Hartigan, P. M. Algorithm AS 217: Computation of the Dip Statistic to Test for Unimodality. *Appl Stat* **34**, 320 (1985).
533. SAS Institute Inc. *SAS/STAT User's Guide, Version 6, 4th Edn*. (SAS Institute Inc, 1990).
534. Watanabe, M. *et al.* Estimation of age-related DNA degradation from formalin-fixed and paraffin-embedded tissue according to the extraction methods. *Exp Ther Med* **14**, 2683 (2017).
535. Yi, Q. Q. *et al.* Effect of preservation time of formalin-fixed paraffin-embedded tissues on extractable DNA and RNA quantity. *Journal of International Medical Research* **48**, 1–10 (2020).
536. Rossouw, S. C. *et al.* Evaluation of Protein Purification Techniques and Effects of Storage Duration on LC-MS/MS Analysis of Archived FFPE Human CRC Tissues. *Pathology and Oncology Research* **27**, 622855 (2021).
537. Kokkat, T. J., Patel, M. S., McGarvey, D., Livolsi, V. A. & Baloch, Z. W. Archived Formalin-Fixed Paraffin-Embedded (FFPE) Blocks: A Valuable Underexploited Resource for Extraction of DNA, RNA, and Protein. *Biopreserv Biobank* **11**, 101 (2013).
538. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
539. Kuhn, M. & Johnson, K. *Applied Predictive Modeling*. (2013).
540. Munoz, A. C., Jain, N. K. & Gupta, M. Albumin Colloid. *StatPearls* (2022).

541. Kainov, Y. *et al.* CRABP1 provides high malignancy of transformed mesenchymal cells and contributes to the pathogenesis of mesenchymal and neuroendocrine tumors. *Cell Cycle* **13**, 1530 (2014).
542. Demicco, E. G. *et al.* Progressive loss of myogenic differentiation in leiomyosarcoma has prognostic value. *Histopathology* **66**, 627–638 (2015).
543. Tabb, D. L. *et al.* Reproducibility of Differential Proteomic Technologies in CPTAC Fractionated Xenografts. *J Proteome Res* **15**, 691–706 (2016).
544. Clinical Proteomic Tumor Analysis Consortium (CPTAC) | NCI Genomic Data Commons. https://gdc.cancer.gov/about-gdc/contributed-genomic-data-cancer-research/clinical-proteomic-tumor-analysis-consortium-cptac.
545. Zaidel-Bar, R., Itzkovitz, S., Ma'ayan, A., Iyengar, R. & Geiger, B. Functional atlas of the integrin adhesome. *Nat Cell Biol* **9**, 858–867 (2007).
546. Oken, M. M. *et al.* Toxicity and response criteria of the Eastern Cooperative Oncology Group - PubMed. *Am J Clin Oncol.*
547. Lahat, G. *et al.* Outcome of Locally Recurrent and Metastatic Angiosarcoma. *Annals of Surgical Oncology 2009 16:9* **16**, 2502–2509 (2009).
548. Buehler, D. *et al.* Angiosarcoma Outcomes and Prognostic Factors: A 25-Year Single Institution Experience. *Am J Clin Oncol* **37**, 473 (2014).
549. Ikoma, N. *et al.* Recurrence Patterns of Retroperitoneal Leiomyosarcoma and Impact of Salvage Surgery. *J Surg Oncol* **116**, 313 (2017).
550. Wang, Z. *et al.* Survival of patients with metastatic leiomyosarcoma: the MD Anderson Clinical Center for targeted therapy experience. *Cancer Med* **5**, 3437 (2016).
551. Tirumani, S. H. *et al.* Metastatic pattern of uterine leiomyosarcoma: retrospective analysis of the predictors and outcome in 113 patients. *J Gynecol Oncol* **25**, 306 (2014).
552. Penel, N. *et al.* Performance status is the most powerful risk factor for early death among patients with advanced soft tissue sarcoma The European Organisation for Research and Treatment of Cancer – Soft Tissue and Bone Sarcoma Group (STBSG) and French Sarcoma Group (FSG) study. *Br J Cancer* **104**, 1544 (2011).
553. Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187–202 (1972).
554. Lugano, R. *et al.* CD93 promotes β1 integrin activation and fibronectin fibrillogenesis during tumor angiogenesis. *J Clin Invest* **128**, 3280 (2018).
555. Cao, G. *et al.* Involvement of human PECAM-1 in angiogenesis and in vitro endothelial cell migration. *Am J Physiol Cell Physiol* **282**, (2002).

556. Bao, L. *et al.* Elevated expression of CD93 promotes angiogenesis and tumor growth in nasopharyngeal carcinoma. *Biochem Biophys Res Commun* **476**, 467–474 (2016).
557. Gounder, M. M., Thomas, D. M. & Tap, W. D. Locally Aggressive Connective Tissue Tumors. *J Clin Oncol* **36**, 202–209 (2018).
558. George, S., Serrano, C., Hensley, M. L. & Ray-Coquard, I. Soft Tissue and Uterine Leiomyosarcoma. *J Clin Oncol* **36**, 144–150 (2018).
559. Yamasaki, H. *et al.* Synovial sarcoma cell lines showed reduced DNA repair activity and sensitivity to a PARP inhibitor. *Genes Cells* **21**, 852–860 (2016).
560. Jones, S. E. *et al.* ATR Is a Therapeutic Target in Synovial Sarcoma. *Cancer Res* **77**, 7014–7026 (2017).
561. Conus, S. & Simon, H. U. Cathepsins and their involvement in immune responses. *Swiss Med Wkly* **140**, (2010).
562. Sarma, J. V. & Ward, P. A. The complement system. *Cell Tissue Res* **343**, 227–235 (2011).
563. Pankova, V., Thway, K., Jones, R. L. & Huang, P. H. The Extracellular Matrix in Soft Tissue Sarcomas: Pathobiology and Cellular Signalling. *Front Cell Dev Biol* **9**, (2021).
564. Pietilä, E. A. *et al.* Co-evolution of matrisome and adaptive adhesion dynamics drives ovarian cancer chemoresistance. *Nature Communications 2021 12:1* **12**, 1–19 (2021).
565. Yuzhalin, A. E., Urbonas, T., Silva, M. A., Muschel, R. J. & Gordon-Weeks, A. N. A core matrisome gene signature predicts cancer outcome. *Br J Cancer* **118**, 435 (2018).
566. bin Lim, S. *et al.* Pan-cancer analysis connects tumor matrisome to immune response. *npj Precision Oncology 2019 3:1* **3**, 1–9 (2019).
567. Rafaeva, M. & Erler, J. T. Framing cancer progression: influence of the organ- and tumour-specific matrisome. *FEBS J* **287**, 1454–1477 (2020).
568. Kalluri, R. Basement membranes: structure, assembly and role in tumour angiogenesis. *Nature Reviews Cancer 2003 3:6* **3**, 422–433 (2003).
569. Engbring, J. A. & Kleinman, H. K. The basement membrane matrix in malignancy. *J Pathol* **200**, 465–470 (2003).
570. Yurchenco, P. D. Basement Membranes: Cell Scaffoldings and Signaling Platforms. *Cold Spring Harb Perspect Biol* **3**, 1–27 (2011).
571. Bella, J. & Hulmes, D. J. S. Fibrillar Collagens. *Subcell Biochem* **82**, 457–490 (2017).

572. Krasny, L. & Huang, P. H. Advances in the proteomic profiling of the matrisome and adhesome. *https://doi.org/10.1080/14789450.2021.1984885* **18**, 781–794 (2021).
573. Humphries, J. D., Byron, A. & Humphries, M. J. Integrin ligands at a glance. *J Cell Sci* **119**, 3901–3903 (2006).
574. Villacis, R. A. R. *et al.* Gene Expression Profiling in Leiomyosarcomas and Undifferentiated Pleomorphic Sarcomas: SRC as a New Diagnostic Marker. *PLoS One* **9**, e102281 (2014).
575. Chan, J. Y. *et al.* Multiomic analysis and immunoprofiling reveal distinct subtypes of human angiosarcoma. *J Clin Invest* **130**, 5833–5846 (2020).
576. Oshi, M. *et al.* G2M checkpoint pathway alone is associated with drug response and survival among cell proliferation-related pathways in pancreatic cancer. *Am J Cancer Res* **11**, 3070 (2021).
577. Deming, S. L., Nass, S. J., Dickson, R. B. & Trock, B. J. C-myc amplification in breast cancer: a meta-analysis of its occurrence and prognostic relevance. *Br J Cancer* **83**, 1688–1695 (2000).
578. de Cássia S. Alves, R., Meurer, R. T. & Roehe, A. V. MYC amplification is associated with poor survival in small cell lung cancer: a chromogenic in situ hybridization study. *J Cancer Res Clin Oncol* **140**, 2021–2025 (2014).
579. Paclitaxel: Uses, Interactions, Mechanism of Action | DrugBank Online. https://go.drugbank.com/drugs/DB01229.
580. Vinblastine: Uses, Interactions, Mechanism of Action | DrugBank Online. https://go.drugbank.com/drugs/DB00570.
581. Vincristine: Uses, Interactions, Mechanism of Action | DrugBank Online. https://go.drugbank.com/drugs/DB00541.
582. Schlemmer, M. *et al.* Paclitaxel in patients with advanced angiosarcomas of soft tissue: a retrospective study of the EORTC soft tissue and bone sarcoma group. *Eur J Cancer* **44**, 2433–2436 (2008).
583. Penel, N. *et al.* Phase II trial of weekly paclitaxel for unresectable angiosarcoma: The ANGIOTAX study. *Journal of Clinical Oncology* **26**, 5269–5274 (2008).
584. Subbiah, V. *et al.* Multimodality Treatment of Desmoplastic small round cell tumor: Chemotherapy and Complete Cytoreductive Surgery Improve Patient Survival. *Clin Cancer Res* **24**, 4865 (2018).
585. Hayes-Jordan, A., LaQuaglia, M. P. & Modak, S. Management of Desmoplastic Small Round Cell Tumor. *Semin Pediatr Surg* **25**, 299 (2016).

586. Ferrari, A. *et al.* Soft Tissue Sarcoma Across the Age Spectrum: A Population-Based Study from the Surveillance Epidemiology and End Results Database. *Pediatr Blood Cancer* **57**, 943 (2011).
587. Francis, I. R., Cohan, R. H., Varma, D. G. K. & Sondak, V. K. Retroperitoneal sarcomas. *Cancer Imaging* **5**, 89 (2005).
588. Painter, C. A. *et al.* The Angiosarcoma Project: enabling genomic and clinical discoveries in a rare cancer through patient-partnered research. *Nature Medicine* vol. 26 181–187 Preprint at https://doi.org/10.1038/s41591-019-0749-z (2020).
589. Wagner, M. J. *et al.* Original research: Multicenter phase II trial (SWOG S1609, cohort 51) of ipilimumab and nivolumab in metastatic or unresectable angiosarcoma: a substudy of dual anti-CTLA-4 and anti-PD-1 blockade in rare tumors (DART). *J Immunother Cancer* **9**, 2990 (2021).
590. Momen, S. *et al.* Dramatic response of metastatic cutaneous angiosarcoma to an immune checkpoint inhibitor in a patient with xeroderma pigmentosum: whole-genome sequencing aids treatment decision in end-stage disease. *Cold Spring Harb Mol Case Stud* **5**, (2019).
591. Sindhu, S., Gimber, L. H., Cranmer, L., McBride, A. & Kraft, A. S. Angiosarcoma treated successfully with anti-PD-1 therapy - a case report. *J Immunother Cancer* **5**, (2017).
592. Hamacher, R. *et al.* Dramatic Response of a PD-L1-Positive Advanced Angiosarcoma of the Scalp to Pembrolizumab. *JCO Precis Oncol* **2**, 1–7 (2018).
593. Florou, V. *et al.* Angiosarcoma patients treated with immune checkpoint inhibitors: A case series of seven patients from a single institution. *J Immunother Cancer* **7**, 1–8 (2019).
594. Complement and coagulation cascades (Homo sapiens) - WikiPathways. https://www.wikipathways.org/index.php/Pathway:WP558.
595. Foley, J. H. *et al.* Interplay between fibrinolysis and complement: plasmin cleavage of iC3b modulates immune responses. *Journal of Thrombosis and Haemostasis* **13**, 610–618 (2015).
596. Barthel, D., Schindler, S. & Zipfel, P. F. Plasminogen Is a Complement Inhibitor. *J Biol Chem* **287**, 18831 (2012).
597. LeBleu, V. S., MacDonald, B. & Kalluri, R. Structure and Function of Basement Membranes. *Exp Biol Med* **232**, 1121–1129 (2007).
598. Reuten, R. *et al.* Basement membrane stiffness determines metastases formation. *Nature Materials 2021 20:6* **20**, 892–903 (2021).
599. Vatcheva, K., Lee, M., McCormick, J. & Rahbar, M. The Effect of Ignoring Statistical Interactions in Regression Analyses Conducted in Epidemiologic

Studies: An Example with Survival Analysis Using Cox Proportional Hazards Regression Model. *Epidemiology (Sunnyvale)* **6**, (2015).

600. Conger, A. J. A Revised Definition for Suppressor Variables: a Guide To Their Identification and Interpretation. *http://dx.doi.org/10.1177/001316447403400105* **34**, 35–46 (2016).
601. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach Learn* **52**, 91–118 (2003).
602. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607–D613 (2019).
603. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res* **45**, D362–D368 (2017).
604. Lei, M. The MCM complex: its role in DNA replication and implications for cancer therapy. *Curr Cancer Drug Targets* **5**, 365–380 (2005).
605. Zou, Y., Liu, Y., Wu, X. & Shell, S. M. Functions of Human Replication Protein A (RPA): From DNA Replication to DNA Damage and Stress Responses. *J Cell Physiol* **208**, 267 (2006).
606. Lee, A. T. J. *et al.* The adequacy of tissue microarrays in the assessment of inter- and intra-tumoural heterogeneity of infiltrating lymphocyte burden in leiomyosarcoma. *Scientific Reports 2019 9:1* **9**, 1–12 (2019).
607. Tirumani, S. H. *et al.* Metastatic pattern of uterine leiomyosarcoma: retrospective analysis of the predictors and outcome in 113 patients. *J Gynecol Oncol* **25**, 306 (2014).
608. van Cann, T. *et al.* Retrospective Analysis of Outcome of Patients with Metastatic Leiomyosarcoma in a Tertiary Referral Center. *Oncol Res Treat* **41**, 206–213 (2018).
609. Perl, K. *et al.* Reduced changes in protein compared to mRNA levels across non-proliferating tissues. *BMC Genomics* **18**, 1–14 (2017).
610. Edfors, F. *et al.* Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol* **12**, 883 (2016).
611. Nota, S. P. F. T. *et al.* High TIL, HLA, and Immune Checkpoint Expression in Conventional High-Grade and Dedifferentiated Chondrosarcoma and Poor Clinical Course of the Disease. *Front Oncol* **11**, 944 (2021).

612. Kitsou, M., Ayiomamitis, G. D. & Zaravinos, A. High expression of immune checkpoints is associated with the TIL load, mutation rate and patient survival in colorectal cancer. *Int J Oncol* **57**, 237 (2020).
613. Aricò, E., Castiello, L., Capone, I., Gabriele, L. & Belardelli, F. Type I Interferons and Cancer: An Evolving Story Demanding Novel Clinical Applications. *Cancers (Basel)* **11**, (2019).
614. Castro, F., Cardoso, A. P., Gonçalves, R. M., Serre, K. & Oliveira, M. J. Interferon-gamma at the crossroads of tumor immune surveillance or evasion. *Front Immunol* **9**, 847 (2018).
615. Lu, C. *et al.* Type i interferon suppresses tumor growth through activating the STAT3-granzyme B pathway in tumor-infiltrating cytotoxic T lymphocytes. *J Immunother Cancer* **7**, 1–11 (2019).
616. Jorgovanovic, D., Song, M., Wang, L. & Zhang, Y. Roles of IFN-γ in tumor progression and regression: a review. *Biomarker Research 2020 8:1* **8**, 1–16 (2020).
617. Charles A Janeway, J., Travers, P., Walport, M. & Shlomchik, M. J. The Humoral Immune Response. (2001).
618. Dzik, S. Complement and Coagulation: Cross Talk Through Time. *Transfus Med Rev* **33**, 199–206 (2019).
619. Zafar, R. & Wheeler, Y. Liposarcoma. *StatPearls* (2022).
620. Mangla, A. & Yadav, U. Leiomyosarcoma. *StatPearls* (2022).
621. Robles-Tenorio, A. & Solis-Ledesma, G. Undifferentiated Pleomorphic Sarcoma. *StatPearls* (2022).
622. Mahmud, S. A., Manlove, L. S. & Farrar, M. A. Interleukin-2 and STAT5 in regulatory T cell development and function. *JAKSTAT* **2**, e23154 (2013).
623. Burchill, M. A., Yang, J., Vogtenhuber, C., Blazar, B. R. & Farrar, M. A. IL-2 Receptor β-Dependent STAT5 Activation Is Required for the Development of Foxp3+ Regulatory T Cells. *The Journal of Immunology* **178**, 280–290 (2007).
624. Jones, D. M., Read, K. A. & Oestreich, K. J. Dynamic Roles for IL-2–STAT5 Signaling in Effector and Regulatory CD4+ T Cell Populations. *The Journal of Immunology* **205**, 1721–1730 (2020).
625. Mori, D. N., Kreisel, D., Fullerton, J. N., Gilroy, D. W. & Goldstein, D. R. Inflammatory triggers of acute rejection of organ allografts. *Immunol Rev* **258**, 132 (2014).
626. Ravindranath, M. H., el Hilali, F. & Filippone, E. J. The Impact of Inflammation on the Immune Responses to Transplantation: Tolerance or Rejection? *Front Immunol* **12**, 4510 (2021).

627. Ricklin, D. & Lambris, J. D. Complement in immune and inflammatory disorders: pathophysiological mechanisms. *J Immunol* **190**, 3831 (2013).
628. Novy, P., Quigley, M., Huang, X. & Yang, Y. CD4 T Cells Are Required for CD8 T Cell Survival during Both Primary and Memory Recall Responses. *The Journal of Immunology* **179**, 8243–8251 (2007).
629. Laidlaw, B. J., Craft, J. E. & Kaech, S. M. The multifaceted role of CD4+ T cells in the regulation of CD8+ T cell memory maturation. *Nat Rev Immunol* **16**, 102 (2016).
630. Baumjohann, D. & Brossart, P. T follicular helper cells: linking cancer immunotherapy and immune-related adverse events. *J Immunother Cancer* **9**, 2588 (2021).
631. Nagasaki, J. *et al.* The critical role of CD4+ T cells in PD-1 blockade against MHC-II–expressing tumors such as classic Hodgkin lymphoma. *Blood Adv* **4**, 4069–4082 (2020).
632. Zuazo, M. *et al.* Systemic CD4 Immunity as a Key Contributor to PD-L1/PD-1 Blockade Immunotherapy Efficacy. *Front Immunol* **11**, (2020).
633. Ben-Ami, E. *et al.* Immunotherapy with single agent nivolumab for advanced leiomyosarcoma of the uterus: Results of a phase 2 study. *Cancer* **123**, 3285–3290 (2017).
634. Bursać, S., Prodan, Y., Pullen, N., Bartek, J. & Volarević, S. Dysregulated Ribosome Biogenesis Reveals Therapeutic Liabilities in Cancer. *Trends Cancer* **7**, 57–76 (2021).
635. Bruggeman, J. W., Koster, J., Lodder, P., Repping, S. & Hamer, G. Massive expression of germ cell-specific genes is a hallmark of cancer and a potential target for novel treatment development. *Oncogene 2018 37:42* **37**, 5694–5700 (2018).
636. Bruggeman, J. W. *et al.* Tumors Widely Express Hundreds of Embryonic Germline Genes. *Cancers (Basel)* **12**, 1–16 (2020).
637. Penel, N. *et al.* Performance status is the most powerful risk factor for early death among patients with advanced soft tissue sarcoma The European Organisation for Research and Treatment of Cancer – Soft Tissue and Bone Sarcoma Group (STBSG) and French Sarcoma Group (FSG) study. *Br J Cancer* **104**, 1544 (2011).
638. Charles A Janeway, J., Travers, P., Walport, M. & Shlomchik, M. J. T Cell-Mediated Immunity. (2001).
639. Leung, L. L. & Morser, J. Plasmin as a complement C5 convertase. *EBioMedicine* **5**, 20 (2016).

640. Xie, C. B., Jane-Wit, D. & Pober, J. S. Complement Membrane Attack Complex: New Roles, Mechanisms of Action, and Therapeutic Targets. *Am J Pathol* **190**, 1138–1150 (2020).
641. Markiewski, M. M., Dunstone, M. A., Honeychurch, J., Fishelson, Z. & Kirschfink, M. Complement C5b-9 and Cancer: Mechanisms of Cell Damage, Cancer Counteractions, and Approaches for Intervention. *Frontiers in Immunology | www.frontiersin.org* **1**, 752 (2019).
642. Towner, L. D., Wheat, R. A., Hughes, T. R. & PaulMorgan, B. Complement Membrane Attack and Tumorigenesis. *J Biol Chem* **291**, 14927 (2016).
643. Nabizadeh, J. A. *et al.* The Complement C3a Receptor Contributes to Melanoma Tumorigenesis by Inhibiting Neutrophil and CD4+ T Cell Responses. *J Immunol* **196**, 4783–4792 (2016).
644. Markiewski, M. M. *et al.* Modulation of the antitumor immune response by complement. *Nat Immunol* **9**, 1225–1235 (2008).
645. Kwak, J. W. *et al.* Complement Activation via a C3a Receptor Pathway Alters CD4 + T Lymphocytes and Mediates Lung Cancer Progression. *Cancer Res* **78**, 143–156 (2018).
646. Risitano, A. M. *et al.* Anti-complement Treatment for Paroxysmal Nocturnal Hemoglobinuria: Time for proximal complement inhibition? A position paper from the SAAWP of the EBMT. *Front Immunol* **10**, 1157 (2019).
647. Smith, P. K. *et al.* Effects of C5 complement inhibitor pexelizumab on outcome in high-risk coronary artery bypass grafting: combined results from the PRIMO-CABG I and II trials. *J Thorac Cardiovasc Surg* **142**, 89–98 (2011).
648. Okroj, M., Heinegård, D., Holmdahl, R. & Blom, A. M. Rheumatoid arthritis and the complement system. *Ann Med* **39**, 517–530 (2007).
649. Zha, H. *et al.* Blocking C5aR signaling promotes the anti-tumor efficacy of PD-1/PD-L1 blockade. *Oncoimmunology* **6**, (2017).
650. Ajona, D. *et al.* A Combined PD-1/C5a Blockade Synergistically Protects against Lung Cancer Growth and Metastasis. *Cancer Discov* **7**, 694–703 (2017).
651. Budczies, J. *et al.* Cutoff Finder: A Comprehensive and Straightforward Web Application Enabling Rapid Biomarker Cutoff Optimization. *PLoS One* **7**, 51862 (2012).
652. Camp, R. L., Dolled-Filhart, M. & Rimm, D. L. X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. *Clin Cancer Res* **10**, 7252–7259 (2004).
653. Williams, B. A. *et al.* Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes. (2006).

654. Ogłuszka, M., Orzechowska, M., Jędroszka, D., Witas, P. & Bednarek, A. K. Evaluate Cutpoints: Adaptable continuous data distribution system for determining survival in Kaplan-Meier estimator. *Comput Methods Programs Biomed* **177**, 133–139 (2019).
655. Yip, A. M. & Horvath, S. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics* **8**, 1–14 (2007).
656. Albert, R. Scale-free networks in cell biology. *J Cell Sci* **118**, 4947–4957 (2005).
657. Barabási, A. L. & Albert, R. Emergence of Scaling in Random Networks. *Science (1979)* **286**, 509–512 (1999).
658. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature 1998 393:6684* **393**, 440–442 (1998).
659. Erdos, P. & Renyi, A. On random graphs I.
660. Morgenstern, D. A., Gibson, S., Brown, T., Sebire, N. J. & Anderson, J. Clinical and pathological features of paediatric malignant rhabdoid tumours. *Pediatr Blood Cancer* **54**, 29–34 (2010).
661. Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
662. Sabidussi, G. The centrality index of a graph. *Psychometrika 1966 31:4* **31**, 581–603 (1966).
663. Tsutakawa, S. E. *et al.* Human Flap Endonuclease Structures, DNA Double-Base Flipping, and a Unified Understanding of the FEN1 Superfamily. *Cell* **145**, 198–211 (2011).
664. Chapados, B. R. *et al.* Structural Basis for FEN-1 Substrate Specificity and PCNA-Mediated Activation in DNA Replication and Repair. *Cell* **116**, 39–50 (2004).
665. Popoff, V., Adolf, F., Brügge, B. & Wieland, F. COPI Budding within the Golgi Stack. *Cold Spring Harb Perspect Biol* **3**, (2011).
666. Shibata, H., Huynh, D. P. & Pulst, S. M. A novel protein with RNA-binding motifs interacts with ataxin-2. *Hum Mol Genet* **9**, 1303–1313 (2000).
667. Jerby-Arnon, L. *et al.* Opposing immune and genetic mechanisms shape oncogenic programs in synovial sarcoma. *Nature Medicine 2021 27:2* **27**, 289–300 (2021).
668. Przybyl, J. *et al.* Recurrent and novel SS18-SSX fusion transcripts in synovial sarcoma: description of three new cases. *Tumour Biol* **33**, 2245–2253 (2012).
669. Miallot, R., Galland, F., Millet, V., Blay, J. Y. & Naquet, P. Metabolic landscapes in sarcomas. *Journal of Hematology & Oncology 2021 14:1* **14**, 1–23 (2021).

670. Castellano, M. D. M., Boniotti, M. B., Caro, E., Schnittger, A. & Gutierrez, C. DNA Replication Licensing Affects Cell Proliferation or Endoreplication in a Cell Type–Specific Manner. *Plant Cell* **16**, 2380 (2004).
671. Stoeber, K. *et al.* DNA replication licensing and human cell proliferation. *J Cell Sci* **114**, 2027–2041 (2001).
672. Cooper, G. M. The Mechanism of Vesicular Transport. (2000).
673. Krishnan, P. D. G., Golden, E., Woodward, E. A., Pavlos, N. J. & Blancafort, P. Rab GTPases: Emerging Oncogenes and Tumor Suppressive Regulators for the Editing of Survival Pathways in Cancer. *Cancers (Basel)* **12**, (2020).
674. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).
675. Minakshi, P. *et al.* Single-Cell Proteomics: Technology and Applications. *Single-Cell Omics: Volume 1: Technological Advances and Applications* 283–318 (2019) doi:10.1016/B978-0-12-814919-5.00014-2.
676. Perkel, J. M. Single-cell proteomics takes centre stage. *Nature* **597**, 580–582 (2021).
677. Lee, P. Y. *et al.* Molecular tissue profiling by MALDI imaging: recent progress and applications in cancer research. *https://doi.org/10.1080/10408363.2021.1942781* **58**, 513–529 (2021).
678. Aichler, M. & Walch, A. MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Laboratory Investigation 2015 95:4* **95**, 422–431 (2015).