**IMPORTANT NOTE:** The following is a thorough annotation of a subset of clusters we identified within our transcriptional heterogeneity analysis project. Currently we are considering publishing this series of documents independently of the main paper, as an online resource for users to reference when diving into our pan-cancer atlas. These will be then updated with time as new and better information emerges. Alternatively, a stripped-down version could be published as supplemental information.

**T002 Mesodermal tumours with high immune infiltration and T003 Mesodermal tumours with high stemness**

**T002 MESODERM IMM high**

We find that most sarcomas and other mesenchymal cancers, together with a few other sarcomatous and non-sarcomatous solid tumours, are divided into two major families. Samples in **T002 MESODM IMMhigh** (n = 565) have a higher immune infiltration, while **T003 MESODM STEMhigh** (n = 483) have higher stemness (Fig. 4b) .

**Leiomyosarcoma**

**T002** splits into two clusters at the first level: **T067 LMS** (n = 78), which contains most leiomyosarcomas and **T066 MESODM IMMhigh A** (n = 487), carrying the remaining malignancies (Fig. S23a, b). **T067** shows, as expected, overexpression of genes relating to smooth muscle development, including *ACTA2* (logFC = -4.29, FDR = 6.853e-133)*,* *ACTG2* (logFC = -7.92, FDR = 2.126e-168), *DES* (logFC = -7.62, FDR = 4.876e-121), and *MYOC* (logFC = -8.66, FDR = 5.518e-154), and has significantly better survival than samples in **T066** (lrt p-val = 2.92e-03 at 5840 days).

It is then divided in three subclasses, roughly defined by the tumour location. **T087 ULMS** (n = 18) is composed of uterine LMS (n = 14, χ2 p-val = 8.96e-11), **T088 STLMS ABD** (n = 23) is largely composed of abdominal and retroperitoneal soft tissue LMS (n= 16, χ2 p-val = 7.85e-05), while **T089 STLMS EXT** (n = 29) contains a significant portion of LMS of the extremities (n = 9, χ2 p-val =3.56e-04). Though there is no significant difference in overall survival between the groups (lrt p-val = 1.32e-01 at 3765 days), **T089** has a higher incidence of relapsed tumours (χ2 p-val = 2.93e-02).

**Other Sarcomas**

**T066** splits into osteosarcomas, **T068 OSARC** (n = 131), a class of sarcomas with mixed diagnoses **T069 SARC IMMhigh** (n = 275), and mesotheliomas **T070 MPM** (n = 81) (Fig. S23a, b). These clusters differ significantly in age (KW adj. p-val = 1.67e-33) and in number of paediatric samples (χ2 p-val ≤ 4.49e-56); **T068** is the cluster with the youngest patients (median age of 15 y.o.) and is almost exclusively paediatric (96.18%). At variance, mesotheliomas in **T070** are almost exclusively adult tumours, with a patients’ median age of 63 y.o, and significantly worse overall survival than the other two classes (lrt p-val = 3.90e-11 at 5840 days). Finally, **T069,** the mixed sarcoma class, is in between, with a patients’ median age of 60 y.o. and has 23.63% pediatric samples.

Of note, within **T069 SARC IMM high** we observe the presence of a number of samples (such as osteosarcoma and leiomyosarcoma) for which a matching type-specific cluster is available elsewhere. These likely reflect clinically and/or developmentally distinct groups within these specific tumour types and may exhibit variable levels of immune activity, idiosyncratic genomic lesions and distinct disease progression, as recently seen in literature.

**Osteosarcoma**

The osteosarcomas in our cohort divide into four distinct subtypes (Fig. S22b, i). All samples for which we have clinical data are osteosarcomas of the long bones or pelvis (cite TARGET Discovery cohort).

**T071 OSARC OSSIF** (n = 32) has patients with a median age of 15.65 y.o. and predominantly male, (75.00%). It exhibits overexpression (FDR < 0.05) of cancer testis antigen (CTA) genes, most notably the *SSX* (8/9 genes), *MAGEA* (10/12), *MAGEB* (6/10), *CSAG* (2/2) and *XAGE* (4/5) families, several of which are known to be upregulated in osteosarcoma47. Though CTA expression has been associated with poor prognosis in osteosarcoma1, this cluster exhibits favourable prognosis when compared to its siblings (lrt p-val = 5.56e-05 at 5840 days, median OS not reached) (Fig. S22j). As this cluster is also associated with direct ossification (medNES ≥ 1.01, KW adj. p-val = 6.69e-10) and positive regulation of osteoblast differentiation (medNES ≥ 1.05, KW adj. p-val = 4.18e-02), it may represent a subtype of osteoblastic or non-specific osteosarcoma with good prognosis.

**T072 OSARC CHOND** (n = 38) has patients with a median age of 15 y.o. and predominantly male (57.89%). It is enriched for chondrocyte marker genes, such as *COL9A1* (median logFC= 7.73, FDR ≤ 7.08e-08), *SOX9* (median logFC= 2.20, FDR ≤ 3.34e-05), and *OGN* (median logFC= 3.98, FDR ≤ 1.16e-03), as well as genesets for collagen synthesis (mednes ≥ 1.97, KW adj. p-val = 1.20e-12, Dunn adj. p-val-val < 1.00e-03), chondrocyte differentiation (medNES ≥ 1.13, KW adj. p-val = 1.77e-09, , Dunn adj. p-val-val < 1.00e-03), and cartilage development involved in endochondral morphogenesis (medNES = 1.13, KW adj. p-val = 4.21e-09)2,3(Fig. S23d) suggesting these tumours have significant chondroid components and may represent chondroblastic osteosarcoma (Fig. S22k). Furthermore, **T072** contains all osteosarcomas of the pelvis, including the ilium and sacrum, in our dataset (0/22 vs. 4/16 vs. 0/17 vs. 0/3, χ2 p-val = 1.03e-02), a location associated with chondroblastic osteosarcomas4,5. **T072** also overexpresses *MYC* (median logFC= 1.2, FDR ≤ 4.67e-04), and has the lowest expression of *RB1* (median logFC= -1.11, FDR ≤ 1.63e-03). Samples in this cluster exhibit poor overall survival, reaching median OS at 1906 days post diagnosis (Fig. S22j).

**T073 OSARC OSBLA** (n = 37) has the youngest group of patients with a median age of 13.66 y.o. and is composed predominantly of female patients (57.89%). It significantly overexpresses the master bone regulator *SP7* (median logFC= 0.939, FDR ≤ 1.712e-02) (Fig. S22m), and osteoblast markers *SOST* (median logFC= 5.66, FDR ≤ 1.437e-04) (Fig. S23d) and *SATB2* (median logFC= 1.52, FDR ≤ 1.142e-03)6. Furthermore, it is enriched for genesets for bone mineralization (medNES ≥ 1.02, KW adj. p-val = 4.26e-05)2,3, and replacement ossification of existing non-cartilagenous tissues (medNES ≥ 1.07, KW adj. p-val = 2.23e-03, Dunn adj. p-val < 5.00e-02) (Fig. S23d). It also displays enrichment of mTORC1 signalling (medNES ≥ 1.03, KW adj. p-val = 1.41e-06, Dunn adj. p-val < 1.00e-02), associated with poor prognosis in osteosarcoma7, and cell cycle progression (medNES ≥ 1.01, KW adj. p-val = 5.76e-05, Dunn adj. p-val < 5.00e-02). Samples within this cluster exhibit the worst overall survival of all osteosarcoma clusters, reaching median OS at 679 days post diagnosis (Fig. S22j). As this cluster is composed of ossifying tumours with very poor prognosis, it may represent a highly aggressive subtype of osteoblastic osteosarcoma.

Finally, **T074 OSARC OSCL** (n = 11) is the smallest cluster, with the oldest median age of 22.57 y.o., and has the greatest female composition (75.00%). It also exhibits the best overall survival among all its siblings, with no deaths recorded in our dataset at 5840 days.

Gene sets analysis revealed significant enrichment of sets related to osteoclast differentiation (medNES≥ 1.16, KW adj. p-val = 6.80e-11, Dunn adj. p-val < 1.00e-02), bone remodelling (medNES ≥ 1.14, KW adj. p-val = 3.14e-06, Dunn adj. p-val < 5.00e-02), and fibrinolysis (medNES ≥ 9.43, KW adj. p-val = 8.83e-06, Dunn adj. p-val < 1.00e-04) (Fig. S23d). This profile suggests this cluster contains osteoclast-rich and highly lytic/unstable tumours and may represent telangiectatic osteosarcoma, though we lack the clinical annotation to confirm this.

**Mesothelioma**

Similarly, mesotheliomas in **T083 MPM** also follow a simple path in their subtyping hierarchy (Fig. S23b). They first split into **T096 MPM BAP1 LOH** (n = 59), a mixed biphasic and epithelial class, and **T097 MPM EPITH** (n = 23) which is composed almost exclusively of epithelial tumours (17/21). **T096** shows higher *BAP1* loss of heterozygosity (p-val = 3.96e-2)8,9, has significantly worse prognosis than **T097** (lrt p-val = 1.20e-3 at 2800 days post diagnosis), and shows higher EMT scores (MWU p-val = 4.24e-05)10 due to its biphasic component, as well as lower ploidy (MWU p-val = 1.55e-3).

**T096** further splits by histology, with **T098 MPM BAP1 LOH BIPHAS** (n = 23) containing both biphasic and epithelial samples and **T099 MPM BAP1 LOH EPITH** (n = 23) being almost exclusively composed of epithelial tumours. As before, **T098,** with a majority component of biphasic samples exhibit a significantly higher EMT score (MWU p-val = 1.73e-2).

**Mixed sarcomas**

The hierarchy of **T069 SARC IMMhigh** is deeper and more complex (Fig. S23b). At the first level, we see the separation of **T075 SARC IMM high A** (n = 218) and **T076 STS CIN** (n = 57) (Fig. S23c). Both are mixed clusters, but **T076** contains, for the major part, soft tissue sarcomas, including dedifferentiated liposarcomas (DDLPS) undifferentiated pleomorphic sarcomas (UPS) and myxofibrosarcomas (MFS). The former class is significantly younger (58.00 vs. 65.00, MWU p-val = 9.16e-04) but we observed no difference in survival between the two classes (lrt p-val = 5.10e-01 at 5204 days).

**T075** overexpresses (FDR < 0.05 & median logFC > 0) cancer testis antigen (CTA) family genes, which show considerable promise for immunotherapeutics11. These include *GAGE* (9/13), *PAGE* (4/6), *MAGEA* (11/12), *MAGEC* (3/3) and *XAGE* (3/5)12.We then investigated immune checkpoint ligands and receptors expression, revealing overexpression of *PD1* (median logFC= -3.37, FDR ≤ 2.044e-03), *PDL1* (median logFC= -0.87, FDR ≤ 1.078e-02), and *CTLA4* in **T075** (median logFC= -2.75, FDR ≤ 8.047e-06). Consistent with this, **T075** displays the lowest leukocyte fraction of its siblings (KW adj. p-val = 5.11e-10, Dunn adj. p-val ≤ 3.69e-04)13.

More interestingly, samples in **T076** show significantly higher chromosomal instability (CIN, MWU p-val = 1.15e-05) (Fig. S23f) without a corresponding difference in mutation load (MWU p-val = 4.75e-01). This holds true for both DDLPS (MWU p-val = 3.70e-03) and UPS (MWU p-val = 1.43e-02) subpopulations when taken independently.

**T076** class further splits by diagnosis into **T094 UPS/MFS CIN** (n = 33), containing UPS and MFS, and **T095 DDLPS CIN** (n = 24) (Fig. S23b), which is composed by the majority of DDLPS. This is reflected in the higher amplification of chr12q15, common to DDLPSs,14 in **T095** (median amp. 23.00 vs 2.00, MWU p-val = 8.335e-07). Furthermore, **T095** has both significantly higher genomic *MDM2* amplification (2.40e-2 vs. 3.66, MWU p-val = 4.37e-06) and expression (logFC = -4.55, FDR = 5.39e-18), as well as *CDK4* amplification (0.00 vs. 3.66 MWU p-val = 7.48e-06) and expression (logFC = -4.27, FDR = 5.39e-18).

Finally, **T075 SARC IMMhigh A** separates in four terminal classes, with varying composition, immunogenicity and age (Fig. S23b). Nevertheless, there are no significant differences in survival between these clusters.

**T077 SARC HYPOX** (n = 58) contains a high variability of diagnoses but is mostly composed of osteosarcomas, UPSs, and LMSs. Only 31.03% of these samples are paediatric, the median age is 60 y.o. Sarcomas within this cluster display the highest mitotic rate compared to those in sibling clusters (KW p-val = 4.75e-05), as also reflected in gene set enrichment analysis (logFC = 0.871, adj. p-val = 7.76e-01). This cluster also has the lowest expression of *TP53* (logFC= -1.66, FDR ≤ 1.78e-11). Tumours in **T077** display the lowest leukocyte fraction (KW p-val = 5.11e-10, Dunn adj. p-val ≤ 1.70e-07) among this family, and also exhibit the lowest expression of the immune checkpoint genes *PD1* (median logFC= -3.37, FDR ≤ 2.044e-03), *PDL1* (median logFC= -0.87, FDR ≤ 1.078e-02), and *CTLA4* (median logFC= -2.75, FDR ≤ 8.047e-06). It is enriched for genes associated with hypoxia in soft tissue sarcomas (medNES ≥ 1.08, KW adj. p-val ≤ 1.44e-05, Dunn adj. p-val < 0.05)15 (Fig. S23d).

**T078 SARC EPITH/KIT** (n = 77) is the largest cluster, and is mostly composed of osteosarcomas, DDLPSs, and also contains five gastrointestinal stromal tumours (GIST)s. It is the cluster with the youngest median age (24 y.o. KW p-val = 2.50e-06, 57.14% paediatric, χ2 p-val = 4.66e-11). **T078** has the highest expression of the *c-KIT* proto-oncogene (median logFC= 1.33, FDR ≤ 3.88e-02) (Fig. S23d). Mutations in *c-KIT* are a major driver of GISTs16 and may explain their affinity to this class. Nevertheless, *c-KIT* mutations are not exclusive of this tumour type17, and indeed, the significance in overexpression is maintained after the removal of GISTs (median logFC = 1.52, FDR ≤ 4.11e-02). We confirmed enrichment *of c-KIT* downstream genes with gene sets analysis (medNES ≥ 1.02, adj. p-val = 7.82e-08)18 (Fig. S23d). **T078** also displays the highest chr12q13~15 amplification (KW p-val = 6.77e-04), likely a consequence of its high population of DDLPS. Furthermore, **T078** has the highest expression of epithelial markers *EPCAM* (median logFC= 1.41, FDR ≤ 1.280e-02), CLDN1 (median logFC= 1.86, FDR ≤ 2.941e-04), and *CDH1* (median logFC= 2.02, FDR ≤ 7.437e-03) among its siblings and shows enrichment of epithelial development gene sets (medNES ≥ 1.12, KW adj. p-val = 4.43e-16, Dunn adj. p-val < 1.00e-04)2,3 (Fig. S23d). It is also enriched for gene sets involving angiogenesis (medNES ≥ 1.10, KW adj. p-val = 3.12e-08, Dunn adj. p-val < 0.05)19, which has been implicated in the pathogenesis of sarcomas with epithelial features. We hypothesize that this class comprises sarcomas with epithelial differentiation and related tumours, possibly including epitheloid subtypes of DDLPS, Osteosarcoma, and others20–22.

The majority of samples present in **T079 SARC CARCN** (n = 41) are not labelled by their source institutions as malignancies of mesenchymal origin but rather as carcinomas or related ecto- or endodermal tumours (Fig. S23c) (23/41 are carcinomas or skin cutaneous melanoma compared to 15/41 being sarcomas); however, sarcomatoid components were noted in many of these samples’ clinical data when available (TCGA)14. **T079** is enriched for *E2F* targets (medNES = 1.06, KW adj. p-val = 6.17e-28) , *MYC* targets (medNES ≥ 1.02, KW adj. p-val ≤ 3.59e-25, Dunn adj. p-val < 1.00e-03), and DNA synthesis (medNES ≥ 1.04, KW adj. p-val = 5.04e-24, Dunn adj. p-val < 5.00e-02) and G2M checkpoint (medNES ≥ 1.04, KW adj. p-val = 1.16e-28) pathways2,3, suggesting its constituents share a pool of mutations whose pathways converge upon increased translation, protein processing, and cell cycle progression. It is also highly enriched for gene sets involving translation (medNES ≤ 1.02, KW adj. p-val = 1.07e-19, Dunn adj. p-val < 5.00e-02) and protein processing (medNES ≤ 1.10, KW adj. p-val = 2.31e-18, Dunn adj. p-val < 1.00e-04)2,3. Sarcomas within this class have a significantly higher mutation load than those in its sibling clusters (median 96.00, KW p-val = 1.60e-03) possibly indicating hypermutation. This cluster also shows a high leukocyte fraction, with the highest lymphocyte content of its siblings (KW p-val = 3.325e-4), specifically CD8+ T cells (KW p-val = 3.14e-06)23.

**T080 STS DIFFlow IMMhigh** (n = 30) is the smallest cluster with patients of oldest median age of 62 y.o. (KW p-val = 2.50e-06) and no paediatric samples. It is mainly composed of DDLPSs and UPSs, similar to **T089 DDLPS/SARC CIN**, but significantly lower in chromosomal instability. It is possible a similar subdivision by diagnosis would have been observed with a more sizeable cohort. It has the highest leukocyte fraction of all classes (KW adj. p-val = 5.11e-10) and is significantly enriched (medNES ≥ 1.06, KW adj. p-val ≤ 8.90e-11) for a myriad of gene sets relating to the immune response, proinflammatory signalling, and complement activation19. We hypothesize that **T080** represents a group of poorly differentiated soft tissue sarcomas with high immune infiltration.

**T004 MESODM STEMhigh**

Following the hierarchy along the high stemness sarcomas branch, **T004,** we first observe a separation by diagnosis (Fig. 5em Fig. S24a, Fig. 25c). **T090 MYOGEN** (n = 152) is composed of myogenic sarcomas with a median age of 7.00 y. o. **T091 MESODM STEM high A** (n = 212) is the largest and most diverse cluster; it is composed of Testicular Germ Cell Tumours (TGCT) synovial sarcomas (SYSARC), and uterine carcinomas (UCS), among other tumours. It is the cluster with the oldest patients, with a median age of 33.00 y. o. Finally, we observe a homogeneous Wilms tumours class, **T092 WILMS** (n = 119). It is the cluster with the youngest median age of 4.38 y.o.

**Myogenic tumours**

Myogenic tumours in **T090** further split into **T093 MYOGEN FUS-** (n = 108) with a majority of rhabdomyosarcomas of the embryonal subtype and other myogenic malignancies and **T094 RMSARC ALV FUS+** (n = 47), which contains instead a majority of alveolar rhabdomyosarcomas (Fig. S24a, c). Indeed **T093** shows significantly higher expression of the *FOXO1*-*PAX3/7* fusion-negative markers *HMGA2* (logFC = 4.76, FDR = 3.82e-17), *EGFR* (logFC = 2.73, FDR = 2.72e-19), and *FBN2* (logFC = 5.35, FDR = 9.920e-35), while **T094** shows overexpression (logFC ≤ -3.17, FDR ≤ 9.73e-22) of fusion-positive markers *TFAP2B* (logFC = -9.14, FDR = 5.078e-45) and *CDH3* (logFC = -3.17, FDR = 9.728e-22)24,25, and significant enrichment (medNES ≥ 1.12, MWU adj. p-val ≤ 3.79e-18) of *FOXO1*-*PAX3/7* fusion-associated pathways25,26.(Fig. S24e). Though **T093** contains a handful of samples labelled as alveolar rhabdomyosarcoma, it is sensible to speculate these may be fusion-negative; this occurrence is common, the fusion is not a necessary feature of this histotype.

At the next level in **T093,** we see the separation of a small group of samples labelled as Wilms tumours, **T096 WILMS MYO** (n = 12) from the rest of fusion-negative myogenic tumours in **T095 MYOGEN FUS- A** (n = 95) (Fig. S24c).

When compared to **T092 WILMS** (see below for details), the major Wilms tumour class, **T096** shows high expression of striated muscle genes such as *MYL1* (logFC = 11.9, FDR = 4.61e-51), *MYOG* (logFC = 9.45, FDR = 3.93e-55), and *MYOD1* (logFC = 8.93, FDR = 1.12e-51). Furthermore, **T096** is enriched for gene sets related to skeletal muscle development (MWU adj. p-val = 3.56e-09, medNES = 1.85)2,3 (Fig. S24e), suggesting this specific subtype of Wilms tumours to be myoblast-rich.

As an alternative hypothesis, we also advance the possibility this may be a class of misdiagnosed rhabdomyosarcoma of the kidney,27,28 or a striated-muscle-like Wilms tumour phenotype. **T096** is enriched for gene sets of rhabdomyosarcomas both with and without *FOXO1*-*PAX3/7* fusions (medNES ≥ 1.04, MWU adj. p-val = 7.17e-04)25, . Furthermore, **T096** has significantly higher expression of NOGGIN (*NOG*) (logFC = 1.77, FDR = 4.31e-07), when compared to other Wilms tumours29. A common classification of Wilms tumours separates them between those with favourable histology (FHWT) and those with diffuse anaplasia (DAWT); **T096** is the only Wilms tumour class within our cohort with a majority of diffuse anaplasia samples.

The remaining myogenic tumours found in **T095** separate into three subclasses (Fig. S24c). Two have a majority component of embryonal rhabdomyosarcomas, **T097 RMSARC EMB MYO** (n = 30) which also contains a few fusion-negative alveolar rhabdomyosarcomas, and **T098 RMSARC EMB MYOD1mut** (n = 35) which also includes two spindle cell/sclerotizing rhabdomyosarcomas. The other is **T099 UCS MYO** (n = 19) a small class of uterine carcinosarcomas. These classes split by age, as expected the former two have a considerable paediatric component (median age 5 y.o. for both), while the carcinosarcomas are mostly adult (median age 63 y.o., KW adj. p-val = 4.30e-05 with 21.06% of paediatric samples, χ2 p-val = 3.76e-13).

Comparing the two embryonal RMSARC classes, we observed **T097** having an significantly elevated expression of skeletal muscle developmental gene sets (medNES = 1.15, KW adj. p-val = 2.61e-07) and high expression of muscle genes (see below) (Fig. S24e), suggesting a well-differentiated subtype.25

**T098** is characterized instead by high expression of gene sets related to immune activation2,3, low expression of skeletal muscle genes, including *MYH8* (logFC= - 4.54, FDR = 1.13e-06), *ACTA1* (logFC = -4.9, FDR = 9.23e-09), and *MYOG* (logFC = -1.74, FDR = 2.31e-04), and enrichment of PI3K signalling (medNES = 1.11, MWU adj. p-val = 4.58e-05)18, a gene expression pattern characteristic of rhabdomyosarcomas with *MYOD1* L122R mutations30 . Further gene set enrichment analysis of targets downregulated by *MYOD1* LI22R compared to wild type *MYOD1* shows significant underexpression in T098 compared to both T097 and T099 (medNES ≤ 4.04e-02, KW adj. p-val = 2.44e-11, Dunn adj. p-val < 1.00e-04) (Fig. S24e); however, we lack genomic information to confirm this match.

Finally, when compared to other uterine carcinosarcomas (**T111**, see below for details), **T099** is significantly higher in sarcomatous components (median 100.00% vs. 70.00, MWU p-val = 8.462e-04) and heterologous rhabdomyosarcomatous components (mean 23.5% vs. 0.00%, MWU p-val = 6.13e-04), while **T111** exhibits a higher carcinomatous component (median 1.00% vs. 30.00%, MWU p-val = 1.03e-02)

**T099** also exhibits significantly higher expression of skeletal muscle genes, *MYOD1* (logFC = 6.88, FDR = 7.27e-16), and *MYOG* (logFC = 9.55, FDR = 9.30e-17), conforming with a recently described myogenic subtype of UCS (subtype II)31.

**Mixed sarcomas II**

Following along the mixed diagnoses branch, **T091**, we find 6 different subclasses with a wide variety of diagnoses (Fig. S24b, c). Some, like **T100 SYSARC** (n = 37) and **T02 CPC** (n =6) are clearly defined by a single tumour type, in this case, synovial sarcomas and choroid plexus carcinomas, respectively.

Others, like **T104 SARC CICr** (n = 21), are composed of samples of disparate origins brought together by specific lesions. This is an exemplary case and is similar to that of *BCOR* altered samples within the CNS branch, here both CNS malignancies and sarcomas carrying *CIC*-*DUX4* fusions (Fig. 5e, f, and h). Gene set enrichment analysis of this cluster revealed both significant enrichment of upregulated targets (medNES ≤ 1.07, KW adj. p-val = 2.36e-18) and significant paucity of downregulated targets (medNES ≥ 8.34e-01, KW adj. p-val = 7.29e-11) in *CIC-DUX4* fusion-positive round cell tumours32,33. **T104** also exhibits overexpression of *MYC* (median logFC= 2.47, FDR ≤ 1.39e-06) (Fig. 5g), frequently amplified in *CIC* rearranged tumours34, as well as its canonical downstream effector *CDKN1A* (median logFC= 2.69, FDR ≤ 5.70e-09). The class includes a few samples labelled as Ewing sarcomas, which are likely misdiagnosed.

**T103 SARC NF1low** (n = 47) contains a majority of uterine carcinosarcomas (UCS), mixed with retroperitoneal DDLPS, MPNST, two ovarian serous cystadenocarcinomas, and a few uterine corpus endometrial carcinomas, among others. UCS samples are clustered into T111 UCS (n = 37) at the next level (Fig. S24c), separating them from all other malignancies, which are found in **T110** **SARC** **NF1mut** (n=25). When compared to the myogenic UCS in **T099,** we observe higher expression of cell adhesion and apoptotic genes, *SIPA1L1* (logFC = 1.33, FDR = 7.290e-08))*, STAT6* (logFC = 0.846, FDR = 4.461e-02)*, CASP6* (logFC = 1.01, FDR = 2.585e-05), *CASP8* (logFC = 0.702, FDR = 4.084e-02) in **T111**, associated with a recently described UCS group (subtype I)31.

**T110** contains a majority of MPNST and DDLPS (n = 4 for each), which seem to be characterized by a loss of *NF1*. This explains the marked separation of this group from few other samples of the same type, foundin **T069 SARC IMMhigh**, along the immune high branch instead.

We observe highly significant enrichment of genes upregulated in *NF1* mutants and impoverishment of genes downregulated by the same lesions between all diagnoses included within **T110** and their counterparts in all other clusters (medNES ≥ 1.26, MWU adj. p-val ≤ 6.53e-07) (Fig. S24e), and more specifically between MPNST and DDLPS in **T110** vs. **T069** (medNES ≥ 1.33, MWU adj. p-val ≤ 6.25e-04)35. As **T110** contains samples from different lineages, including two GBM and three melanomas, it is likely this class contains *NF1* mutant tumours regardless of their histotype, similarly to what observed for *BCOR* altered samples in CNS and *CIC*-fusion samples36–38.

We observe no significant difference in these downstream *NF1* gene sets, between **T110 SARC NF1mut** and **T111** **UCS** (p-val ≥ 3.80e-01), suggesting this expression pattern is characteristic of their whole parent class **T103 SARC NF1**. Among the **T111** cluster only one sample is reported as *NF1* mutant, the only case in the whole TCGA UCS cohort, possibly suggesting a role of *NF1* in UCS regardless of mutation status. Interestingly, when comparing **T111** with **T099** **UCS MYO**, myogenic uterine carcinosarcomas, and its parent **T093 MYO FUS-,** fusion-negative myogenic tumours, only the downregulated targets of *NF1* mutations are significantly lower (KW adj. p-val = 2.40e-05 and 6.30e-01 vs. **T099**, 6.22e-03 and 6.97e-01 vs. **T093**). This is possibly due to the reported role of *NF1* in myogenesis39, and suggests that only the loss of expression in downstream target may be the specific marker of *NF1* alterations in these malignancies.

**Testicular tumours**

Within the child classes of **T091** we also find two separate groups of testicular germ cell tumours non-seminomas (TGCT NON-SEM) (Fig. S24c, d). **T101 TGCT nonSEM MAT/YOLK** (n = 45) is composed of both mature teratoma and yolk sac tumours, as evident both from clinical annotation (χ2 p-val ≤ 3.67e-02) and tissue type percentage information (MWU adj. p-val ≤ 8.35e-03)40. Conversely, **T104 TGCT nonSEM EMB** (n = 45) contains embryonal carcinoma-rich tumours, gleaned from both from clinical annotation (0 vs. 27/39, χ2 p-val = 1.21e-07) and embryonal carcinoma data (MWU adj. p-val =1.48e-8). **T114** exhibits elevated *AFP* expression (logFC = 2.23, FDR = 8.421e-03), while **T118** overexpresses Lactate dehydrogenase genes (4/6, FDR < 1.00e-6) and CGB (*β-HCG*) genes (4/5, FDR < 1.00e-04). **T104** is highly enriched for an embryonal carcinoma gene set (medNES = 216.80, MWU adj. p-val = 3.27e-06), while **T101** is more enriched for a yolk sac (medNES = 2.06, MWU adj. p-val = 3.27e-06) gene sets41.

**T101** further divides into four separate subtypes, falling on a spectrum of differentiation from yolk sac to mature tumours (KW p-val < 1.00e-04) (Fig. S24d). Indeed, we observe **T108** **TGCT nonSEM YOLK** H (n = 13) carrying yolk sac and yolk sac dominant samples with the highest percentage of yolk sac tissue (median 95.00%) and lowest of mature tissue (0.00%), **T107 TGCT nonSEM YOLK I** (n = 9), being just below (yolk sac 42.50%, mature 25.00%), **T109 TGCT nonSEM MAT** I (n = 10) containing mature teratoma dominant samples with low yolk sac content 10.0%) and a considerably higher mature tissue component (65.00%), and finally **T106 TGCT nonSEM MAT H** (n = 13) with the samples showing lowest yolk sac (2.00%) and highest maturation (95.00%) (Fig. S24d).

This separation is further confirmed in the case of **T108** by gene sets, where we see an enrichment of yolk sac tumours genes (medNES = 2.26, KW adj. p-val = 6.21e-05)42.

**T105**, containing samples of the embryonal subtype, splits into two subclasses**. T112 TGCT nonSEM EMB I** (n = 20) contains samples labelled with a variety of subtypes, while **T113 TGCT nonSEM EMB H** (n = 25) is composed almost entirely of samples marked as embryonal. The embryonal carcinoma percentages (median 40.00% vs. 100.00%, MWU p-val = 5.11e-05) further the idea of a continuous spectrum of tissue type, analogous to what we observed in the subtypes of **T101**. Here, **T112** contains samples almost exclusively composed of embryonal tissue, while **T113** contains those with a more intermediate component. This is confirmed by gene set enrichment, where **T112** is enriched for yolk sac and teratoma gene sets (medNES ≥ 1.24 MWU adj. p-val ≤ 1.00e-5) while **T113** is enriched for an established embryonal carcinoma gene set (medNES = 1.04, MWU adj. p-val = 1.15e-03)41.

**Wilms tumours**

When compared to the myogenic subtype in **T099 WILMS MYO**, we observe significantly higher expression of metanephrogenic genes *PAX2* (logFC = 1.81, FDR = 1.03e-07)*, OSR1* (logFC = 1.77, FDR = 9.07e-04)*, EYA1* (logFC = 1.44, FDR = 1.17e-06)*, MEOX1* (logFC = 1.13, FDR = 2.269e-03), and *SALL2* (logFC = 0.96, FDR = 3.962e-04) in **T092 WILMS**.43 This class then divides into 5 different subtypes with characteristic transcriptional profiles (Fig. S24c), in line with FHWT transcriptional clusters recently described by a joint COG-TARGET initiative44.

Importantly, we observe a mixture of both FHWT and DAWT categories across all classes; however, all our bona fide Wilms subtypes (children of **T092**) have significantly higher proportion of FHWT apart from **T117** - which is evenly divided. **T096 WILMS MYO**, the straited-muscle-like components group of Wilms tumours, is the only cluster to have a higher DAWT component (see section on **T096**). **T096** is also the only Wilms tumour cluster to be composed exclusively of histologically mixed tumours.

**T114 WILMS PI3K/mTOR** (n=11) is the smallest cluster and is exclusively composed of COG-TARGET FHWT expression cluster 2 samples (χ2 p-val = 5.587e-07) and is defined by significant enrichment of gene sets related to PI3K-mTOR signalling (medNES ≥ 1.01, KW adj. p-val = 4.14e-04) and the interferon response (medNES ≥ 1.06, KW adj p-val = 1.32e-04) (Fig. S24e). It also exclusively contains SIX1/2 mutants (5 and 4). Furthermore, it has the greatest proportion of blastemal samples (6/7, χ2 p-val = 5.74e-04).

**T115 WILMS OXYPHO** (n = 27), the largest cluster, is defined by enrichment of gene sets related to oxidative phosphorylation (medNES ≥ 1.06, KW adj. p-val = 1.13e-08, Dunn adj. p-val < 1.00e-02) and low expression of mitotic spindle related sets (medNES ≥ 0.90, KW adj p-val = 6.03e-10, Dunn adj. p-val < 1.00e-03)19(Fig. S24e), which is similar to COG-TARGET cluster 5. That **T115** is the only cluster to contain expression class 5 FHWT samples (χ2 p-val = 2.03e-3) confirms this identity, though it also contains an equal number of expression class 1 and class 2 samples. Like **T114**, **T115** is also composed of a majority of blastemal samples (11/17).

**T116 WILMS EMT** (n = 26) is defined by enrichment of gene sets related to the epithelial mesenchymal transition (medNES ≥ 1.07, KW adj. p-val = 2.41e-08, Dunn adj. p-val < 5.00e-02) and angiogenesis (medNES ≥ 1.07, KW adj. p-val = 2.30e-06)19 (Fig. S24e) It also exhibits the lowest expression of *WT1* amongst its siblings (median logFC= -1.24, FDR ≤ 1.30e-02). It should be noted that this expression profile also corresponds to **T096 Wilms MYO**, with these two classes corresponding to the profile of COG-TARGET cluster 4. However, while **T116** is composed of a majority of expression class 3 samples (13/23, χ2 p val = 6.806e-05), it and **T096** contain the largest expression class 4 components (n = 3 each). Histologically, **T116** is composed mainly of mixed tumours (15/23), and also contains the majority of FHWT samples marked as having WT1 loss (χ2 p val = 3.04e-02).

**T117** **WILMS KDEV** (n = 23) is defined by enrichment of the estrogen (medNES ≥ 1.07, KW adj. p-val ≤ 1.67e-06) and androgen responses (medNES ≥ 1.05, KW adj. p-val = 7.38e-04, Dunn adj. p-val < 5.00e-02) and notch signalling (medNES ≥ 1.04, KW adj. p-val = 3.97e-08)19; consequently, gene sets of kidney development relating to the ureteric metanephric mesenchyme (medNES ≥ 1.54, KW adj. p-val = 4.68e-04, Dunn adj. p-val < 1.00e-02) and loop of Henle (medNES ≥ 8.11, KW adj. p-val = 1.25e-08)2,3 are also upregulated (Fig. S24e). It is the only cluster to contain COG-TARGET expression class 6 samples (n = 3, χ2 p val = 2.61e-04) but contains a higher amount of class 3 samples (n = 6). It contains the highest number of *TP53* mutants (4 in **T145**, 6 in **T113**, 10 in **T115**, 7 in **T118**), and is composed entirely of mixed and epithelial tumours (5/9 and 4/9, respectively).

Finally, **T118 WILMS E2F** (n = 26) is defined by enrichment cell proliferation sets, including the G2M checkpoint (medNES = 1.01, KW adj. p-val = 6.39e-03) and mitotic spindle (medNES = 1.03 KW adj. p-val = 6.03e-10)19, as well as genesets for *E2F* (E2F6 KW adj. p-val = 7.86e-03, and E2F1 KW adj. p-val = 5.82e-06) 45 activity, and histone modifications (Fig. S24e). Given its enrichment for *E2F* signalling and proliferative gene sets, **T118** corresponds to COG-TARGET cluster 1. Indeed, it is the only cluster to be composed of a majority of expression class 1 samples (14/17, χ2 p val = 1.927e-05). Its samples exhibit variable histology, with a majority of samples being blastemal (8/17), with small mixed (5/7) and epithelial (4/17) components. Although no differences in survival reached significance between any of the clusters (lrt p-val = 7.40e-02 at 4795 days), this group exhibits the worst overall survival and is the only cluster to reach median OS (1229 days post-diagnosis).

Our clusters of Wilms tumours seem to represent, like the TGCTs non seminomas in the same supergroup, a spectrum of differentiation from blastemal through mixed to epithelial tumours.

**Bibliography**

1. Zou, C. *et al.* Cancer-testis antigens expressed in osteosarcoma identified by gene microarray correlate with a poor patient prognosis. *Cancer* **118,** 1845–1855 (2012).

2. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25,** 25–29 (2000).

3. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47,** D330–D338 (2019).

4. Saab, R. *et al.* Osteosarcoma of the pelvis in children and young adults: the St. Jude Children’s Research Hospital experience. *Cancer* **103,** 1468–1474 (2005).

5. Kawai, A., Huvos, A. G., Meyers, P. A. & Healey, J. H. Osteosarcoma of the pelvis. Oncologic results of 40 patients. *Clin. Orthop. Relat. Res.* 196–207 (1998).

6. Conner, J. R. & Hornick, J. L. SATB2 is a novel marker of osteoblastic differentiation in bone and soft tissue tumours. *Histopathology* **63,** 36–49 (2013).

7. Hu, K., Dai, H.-B. & Qiu, Z.-L. mTOR signaling in osteosarcoma: Oncogenesis and therapeutic aspects (Review). *Oncol. Rep.* **36,** 1219–1225 (2016).

8. Pulford, E., Huilgol, K., Moffat, D., Henderson, D. W. & Klebe, S. Malignant mesothelioma, BAP1 immunohistochemistry, and VEGFA: does BAP1 have potential for early diagnosis and assessment of prognosis? *Dis. Markers* **2017,** 1310478 (2017).

9. Alakus, H. *et al.* BAP1 mutation is a frequent somatic event in peritoneal malignant mesothelioma. *J. Transl. Med.* **13,** 122 (2015).

10. Hmeljak, J. *et al.* Integrative molecular characterization of malignant pleural mesothelioma. *Cancer Discov.* **8,** 1548–1565 (2018).

11. Gjerstorff, M. F., Andersen, M. H. & Ditzel, H. J. Oncogenic cancer/testis antigens: prime candidates for immunotherapy. *Oncotarget* **6,** 15772–15787 (2015).

12. Carregaro, F., Stefanini, A. C. B., Henrique, T. & Tajara, E. H. Study of small proline-rich proteins (SPRRs) in health and disease: a review of the literature. *Arch Dermatol Res* **305,** 857–866 (2013).

13. Thorsson, V. *et al.* The immune landscape of cancer. *Immunity* **48,** 812–830.e14 (2018).

14. The Cancer Genome Atlas Research Network. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell* **171,** 950–965.e28 (2017).

15. Yang, L. *et al.* Validation of a hypoxia related gene signature in multiple soft tissue sarcoma cohorts. *Oncotarget* **9,** 3946–3955 (2018).

16. Hirota, S. *et al.* Gain-of-function mutations of c-kit in human gastrointestinal stromal tumors. *Science* **279,** 577–580 (1998).

17. Smithey, B. E., Pappo, A. S. & Hill, D. A. C-kit expression in pediatric solid tumors: a comparative immunohistochemical study. *Am. J. Surg. Pathol.* **26,** 486–492 (2002).

18. Schaefer, C. F. *et al.* PID: the pathway interaction database. *Nucleic Acids Res.* **37,** D674-9 (2009).

19. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1,** 417–425 (2015).

20. Thway, K., Jones, R. L., Noujaim, J. & Fisher, C. Epithelioid sarcoma: diagnostic features and genetics. *Adv Anat Pathol* **23,** 41–49 (2016).

21. Deyrup, A. T. & Montag, A. G. Epithelioid and Epithelial Neoplasms of Bone. *Arch. Pathol. Lab. Med.* (2007).

22. Makise, N. *et al.* Dedifferentiated liposarcoma with epithelioid/epithelial features. *Am. J. Surg. Pathol.* **41,** 1523–1531 (2017).

23. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12,** 453–457 (2015).

24. Parham, D. M. & Barr, F. G. Classification of rhabdomyosarcoma and its molecular basis. *Adv Anat Pathol* **20,** 387–397 (2013).

25. Davicioni, E. *et al.* Molecular classification of rhabdomyosarcoma--genotypic and phenotypic determinants of diagnosis: a report from the Children’s Oncology Group. *Am. J. Pathol.* **174,** 550–564 (2009).

26. Gryder, B. E. *et al.* PAX3-FOXO1 Establishes Myogenic Super Enhancers and Confers BET Bromodomain Vulnerability. *Cancer Discov.* **7,** 884–899 (2017).

27. Samkari, A. & Al-Maghrabi, H. Rhabdomyosarcoma of the kidney. *Journal of Pediatric Surgery Case Reports* **32,** 62–67 (2018).

28. Mehrain, R. & Nabahati, M. A case of rhabdomyosarcoma of kidney mimicking nephroblastoma. *Caspian J Intern Med* **4,** 621–623 (2013).

29. Gerhart, J. *et al.* Rhabdomyosarcoma and Wilms tumors contain a subpopulation of noggin producing, myogenic cells immunoreactive for lens beaded filament proteins. *PLoS One* **14,** e0214758 (2019).

30. Kohsaka, S. *et al.* A recurrent neomorphic mutation in MYOD1 defines a clinically aggressive subset of embryonal rhabdomyosarcoma associated with PI3K-AKT pathway mutations. *Nat. Genet.* **46,** 595–600 (2014).

31. An, Y. *et al.* Identification of distinct molecular subtypes of uterine carcinosarcoma. *Oncotarget* **8,** 15878–15886 (2017).

32. Specht, K. *et al.* Distinct transcriptional signature and immunoprofile of CIC-DUX4 fusion-positive round cell tumors compared to EWSR1-rearranged Ewing sarcomas: further evidence toward distinct pathologic entities. *Genes. Chromosomes Cancer* **53,** 622–633 (2014).

33. Yoshimoto, T. *et al.* CIC-DUX4 Induces Small Round Cell Sarcomas Distinct from Ewing Sarcoma. *Cancer Res.* **77,** 2927–2937 (2017).

34. Smith, S. C. *et al.* CIC-DUX sarcomas demonstrate frequent MYC amplification and ETS-family transcription factor expression. *Mod. Pathol.* **28,** 57–68 (2015).

35. Pemov, A., Li, H., Presley, W., Wallace, M. R. & Miller, D. T. Genetics of human malignant peripheral nerve sheath tumors. *Neurooncol Adv* **2,** i50–i61 (2020).

36. Kiuru, M. & Busam, K. J. The NF1 gene in tumor syndromes and melanoma. *Lab. Invest.* **97,** 146–157 (2017).

37. Costa, A. D. A. & Gutmann, D. H. Brain tumors in Neurofibromatosis type 1. *Neurooncol Adv* **1,** vdz040 (2019).

38. Kim, Y.-S., Shin, S., Jung, S.-H. & Chung, Y.-J. Pathogenic NF1 truncating mutation and copy number alterations in a dedifferentiated liposarcoma with multiple lung metastasis: a case report. *BMC Med. Genet.* **21,** 200 (2020).

39. Kossler, N. *et al.* Neurofibromin (Nf1) is required for skeletal muscle development. *Hum. Mol. Genet.* **20,** 2697–2709 (2011).

40. Shen, H. *et al.* Integrated molecular characterization of testicular germ cell tumors. *Cell Rep.* **23,** 3392–3406 (2018).

41. Korkola, J. E. *et al.* Gene expression-based classification of nonseminomatous male germ cell tumors. *Oncogene* **24,** 5101–5107 (2005).

42. Korkola, J. E. *et al.* Down-regulation of stem cell genes, including those in a 200-kb gene cluster at 12p13.31, is associated with in vivo differentiation of human male germ cell tumors. *Cancer Res.* **66,** 820–827 (2006).

43. Li, C.-M. *et al.* Gene expression in Wilms’ tumor mimics the earliest committed stage in the metanephric mesenchymal-epithelial transition. *Am. J. Pathol.* **160,** 2181–2190 (2002).

44. Gadd, S. *et al.* A Children’s Oncology Group and TARGET initiative exploring the genetic landscape of Wilms tumor. *Nat. Genet.* **49,** 1487–1494 (2017).

45. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).