


# Genomic, transcriptomic, and immunogenomic landscape of over 1300 sarcomas of diverse histology subtypes

Received: 7 May 2024

Accepted: 24 March 2025

Published online: 06 May 2025

 Check for updates

Alex Soupir<sup>1,2,3</sup>, Oscar E. Ospina<sup>1,2,3</sup>, Oliver Hampton<sup>2</sup>, Michelle Churchman<sup>2</sup>, Michael Radmacher<sup>2</sup>, Dale Hedges<sup>2</sup>, David McKean<sup>2</sup>, Phaedra Agius<sup>2</sup>, Saman Zeeshan<sup>3</sup>, Nathan D. Seligson<sup>4</sup>, Raphael Pollock<sup>5</sup>, David Liebner<sup>6</sup>, James L. Chen<sup>6</sup>, Gabriel Tinoco<sup>6</sup>, Bodour Salhia<sup>7</sup>, Martin McCarter<sup>8</sup>, Breelyn A. Wilky<sup>8</sup>, Benjamin J. Miller<sup>9</sup>, Michael J. Cavnar<sup>10</sup>, John S. Groundland<sup>11</sup>, Bryan P. Schneider<sup>12</sup>, Gregory Riedlinger<sup>13</sup>, Stephen B. Edge<sup>14</sup>, Christopher A. Moskaluk<sup>15</sup>, Kenneth Cardona<sup>16</sup>, Abdul Rafeh Naqash<sup>17</sup>, Ricardo J. Gonzalez<sup>18</sup>, John E. Mullinax<sup>18</sup>, David M. Joyce<sup>18</sup>, Odion Binitie<sup>18</sup>, G. Douglas Letson<sup>18</sup>, Arash O. Naghavi<sup>19</sup>, Mihaela Druta<sup>18</sup>, Damon R. Reed<sup>20</sup>, Erin M. Siegel<sup>21</sup>, Jamie K. Teer<sup>1</sup>, Brooke L. Fridley<sup>1,22</sup> & Andrew S. Brohl<sup>18</sup> 

Given their rarity and diversity, a fundamental understanding of the genomic underpinnings for many sarcoma subtypes is still lacking. To better define the molecular landscape of this group of diseases, we perform matched whole exome sequencing and RNA sequencing on a cohort of 1340 sarcoma tumor specimens. We identify recurrent somatic mutations and observe an increased mutational burden in metastatic vs. primary samples ( $p < 0.001$ ). We observe frequent copy number alterations including whole genome doubling, with this feature being more common in metastatic tumors ( $p = 0.026$ ). Estimation of immune cell abundances followed by hierarchical clustering identifies five immune subtypes ranging from low to high and we observe inferior overall survival in immune deplete clusters compared to immune enriched ( $p < 0.01$ ). Interestingly, GIST predominantly form a distinct “immune intermediate” cluster that is marked by a specific enrichment for NK cells (FDR  $< 0.01$ ).

The term sarcoma encompasses a diverse group of rare malignancies that arise from mesenchymal tissues. Though only accounting for approximately 1% of cancer incidence, there are more than 100 distinct sarcoma subtypes<sup>1</sup>. These features of rarity and diversity pose significant challenges to both research efforts and the clinical management of this group of diseases. Historically, sarcomas have been grouped into broad categories (e.g. “soft tissue sarcoma”) despite highly divergent disease biology and clinical behavior amongst these diseases. Unfortunately, for most patients with advanced-stage sarcoma, treatment options are limited and non-curative. Cytotoxic

anthracycline-based chemotherapy remains the standard first line systemic treatment for most sarcoma varieties<sup>2</sup>.

In recent years, there has been an increasing emphasis on improving the precision of sarcoma diagnosis, treatment, and clinical management using molecular profiling. Clinical molecular profiling efforts have consistently demonstrated the ability to enhance or refine sarcoma diagnosis in a clinically meaningful percentage of patients and, in some cases, uncover opportunities for molecularly guided therapeutic strategies<sup>3–6</sup>. For several of the more common sarcoma types, immunogenomic characterization might also help to select

patients for checkpoint inhibitor therapy<sup>7</sup>, the subject of ongoing clinical study. For several of the rarest subtypes, there have been new diagnostic categorizations based largely on disease-defining genomic features and the resultant recognition of distinct clinical behaviors, for example round cell sarcomas with alternate non-EWS-ETS fusions<sup>8,9</sup>.

Despite the widespread use of next generation sequencing (NGS) in both the lab and clinic, our fundamental molecular understanding of many sarcoma subtypes remains limited. Most initial comprehensive NGS efforts in adult sarcomas, such as The Cancer Genome Atlas (TCGA) project and cohort studies, have focused only on a handful of the more common sarcoma subtypes, and largely have characterized only primary tumor samples<sup>10,11</sup>. Clinical molecular profiling efforts, while including broader diagnostic categories and tumor stages, are often limited to panel-based sequencing designed to capture molecular features relevant in common cancer types and therefore potentially less relevant to sarcomas<sup>3,6</sup>. To further our understanding of the sarcoma molecular landscape across a broad group of sarcomas, we performed comprehensive profiling including whole exome sequencing and transcriptome sequencing on a cohort of >1,300 sarcomas, including both primary and metastatic tumor samples collected as part of the Oncology Research Information Exchange Network (ORIEN) collaborative. In addition to describing foundational genomic features and a landscape of gene expression patterns, we assessed the immunogenomic features of these tumors and the associated micro-environments captured by bulk sequencing and evaluated these immunogenomic features for prognostic relevance.

Results

A total of 1340 tumor samples from 1232 patients representing 42 different sarcoma subtypes were included in this analysis (Table 1). Of these, 1162 samples underwent tumor and germline whole exome sequencing (WES, Supplementary Table 1) and 974 samples underwent RNA sequencing (Supplementary Table 2), with 796 having data available from both. Patients with samples from the primary tumor ( $n = 971$ ), a metastatic site ( $n = 225$ ), or both ( $n = 36$ ) were included in our cohort. Prior to disease-based clustering and analysis, we reclassified the diagnosis in 14 (1.0%) samples based on the presence of a pathognomonic driver fusion (Supplementary fig. 1; Supplementary Table 3).

Somatic mutational analysis and tumor mutational burden

Somatic mutations were identified for all samples and were assessed for enrichment at the gene level per histology subtype. To provide an overview of the most recurrent and relevant findings, the top 10 most significantly mutated genes from each histology were intersected with the COSMIC Tier 1 Cancer Gene Census and used to create an Oncoplot (Fig. 1A, Supplementary Data 1). As expected, *TP53* showed the highest mutation frequency across the entire cohort (18.2%). Additional known disease-specific mutational patterns were recapitulated, such as recurrent *KIT* and *PDGFRA* mutations in GIST.

Interestingly, the frequency of somatic mutations in these cancer-associated genes was significantly higher in metastatic tumors compared to primary tumors (median mutations in COSMIC Tier 1 genes 3 vs. 2, Wilcoxon Rank Sum  $p < 0.001$ ). Notably, the higher mutational rate in known oncologic drivers in metastatic vs. primary samples was largely accounted for by an increase in frequency of the most common tumor suppressors *TP53* (26% vs. 16%, Chi-squared  $p = 0.0007$ ) and *ATRX* (8.9% vs 5.2%, Chi-squared  $p = 0.04$ ) and not by an increase in frequency of driver kinase mutations such as *KIT* (11% vs. 12%, Chi-squared  $p = 0.88$ ).

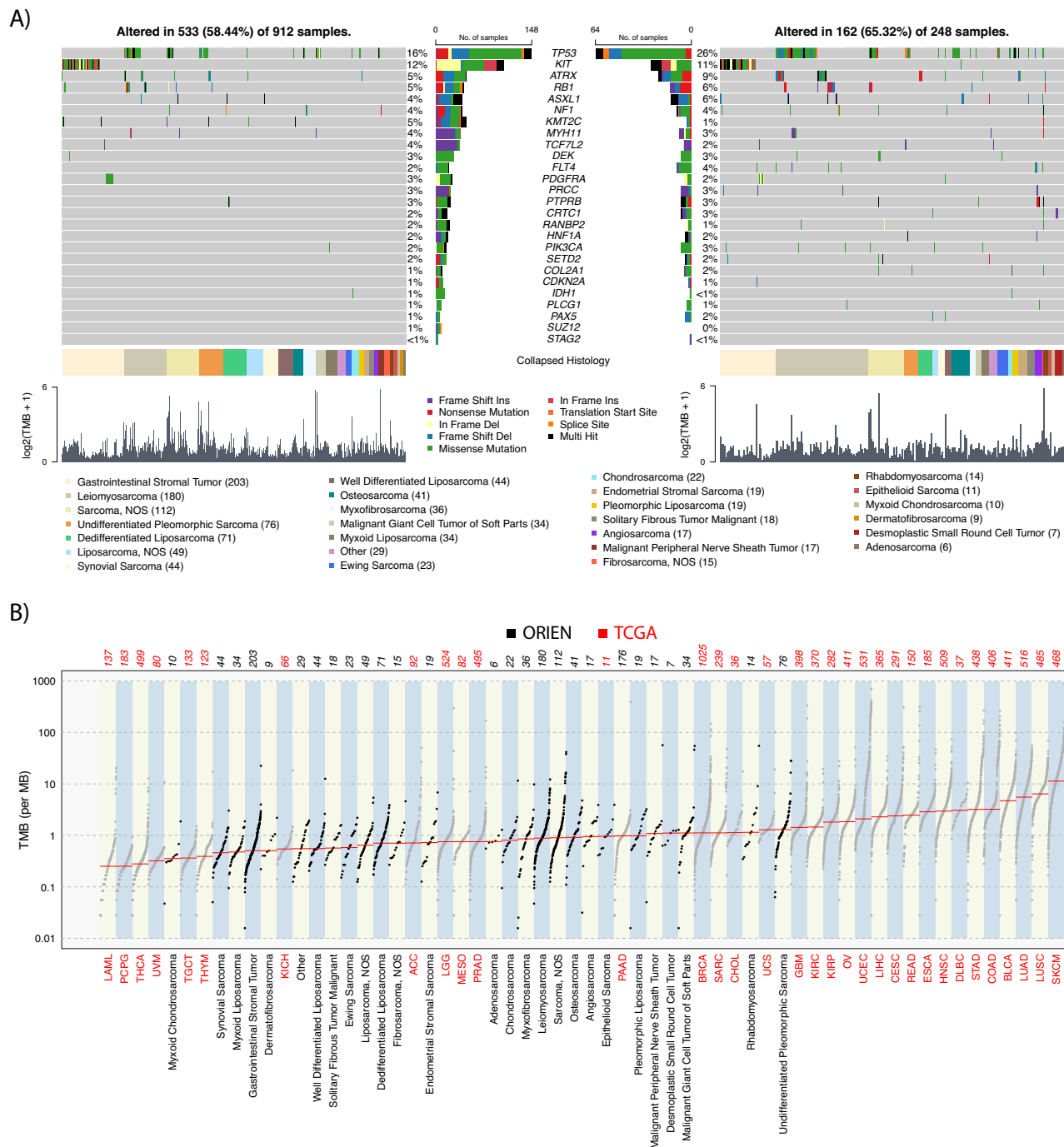
We observed a wide range of tumor mutational burden (TMB), both intra- and inter-histology, in our cohort (Fig. 1B). Undifferentiated pleomorphic sarcoma had the highest median mutations per MB (mut/MB; median = 1.3 mut/MB, range 0.06 to 28.2 mut/MB) amongst histologic types evaluated. Overall, we found that 98.3% of ORIEN samples

Table 1 | Cohort description

	Number	Percent
ORIEN Avatars	1232	
Total Tumors	1340	
Tumor Site		
Primary	1058	79.0
Metastatic	282	21.0
Preservation Method		
Formalin-Fixed Paraffin Embedded	658	49.1
Fresh Frozen	671	50.1
Other	11	0.8
Molecular Data		
Both	796	59.4
RNA	178	13.3
WES	366	27.3
Histology Subtype		
Adenosarcoma	6	0.4
Angiosarcoma	24	1.8
Chondrosarcoma	28	2.1
Dedifferentiated Liposarcoma	80	6.0
Dermatofibrosarcoma	10	0.7
Desmoplastic Small Round Cell Tumor	9	0.7
Endometrial Stromal Sarcoma	22	1.6
Epithelioid Sarcoma	11	0.8
Ewing Sarcoma	23	1.7
Fibrosarcoma, NOS	20	1.5
Gastrointestinal Stromal Tumor	216	16.1
Leiomyosarcoma	204	15.2
Liposarcoma, NOS	54	4.0
Malignant Giant Cell Tumor of Soft Parts	35	2.6
Malignant Peripheral Nerve Sheath Tumor	22	1.6
Myxofibrosarcoma	41	3.1
Myxoid Chondrosarcoma	12	0.9
Myxoid Liposarcoma	45	3.4
Osteosarcoma	50	3.7
Other <sup>a</sup>	37	2.8
Pleomorphic Liposarcoma	21	1.6
Rhabdomyosarcoma	18	1.3
Sarcoma, NOS	133	9.9
Solitary Fibrous Tumor Malignant	20	1.5
Synovial Sarcoma	54	4.0
Undifferentiated Pleomorphic Sarcoma	92	6.9
Well Differentiated Liposarcoma	53	4.0

<sup>a</sup>Histological subtypes with  $n < 5$ , including Adamantinoma of Long Bones ( $n = 1$ ), Alveolar Rhabdomyosarcoma ( $n = 3$ ), Alveolar Soft Part Sarcoma ( $n = 4$ ), Chordoma ( $n = 4$ ), Clear Cell Chondrosarcoma ( $n = 3$ ), Clear Cell Sarcoma ( $n = 3$ ), Dedifferentiated Chondrosarcoma ( $n = 2$ ), Epithelioid Hemangiioendothelioma ( $n = 2$ ), Inflammatory Myofibroblastic Tumor ( $n = 2$ ), Intimal Sarcoma ( $n = 1$ ), Kaposi Sarcoma ( $n = 1$ ), Malignant Rhabdoid Tumor ( $n = 1$ ), Mesenchymal Chondrosarcoma ( $n = 1$ ), Myoepithelial Carcinoma ( $n = 1$ ), Peripheral Neuroectodermal Tumor ( $n = 4$ ), Pleomorphic Rhabdomyosarcoma ( $n = 2$ ), Round Cell Sarcoma ( $n = 2$ ).

demonstrated a low ( $< 5$  mut/MB) or intermediate ( $\geq 5$  mut/MB and  $< 10$  mut/MB) TMB (97.1% and 1.2%, respectively) which was not statistically different from the TCGA SARC cohort (94.6% and 2.1%, respectively, Fisher's Exact  $p = 0.14$ ; Fig. 1B). Average tumor mutational burden across the whole ORIEN cohort was 1.46 mut/MB, with a



**Fig. 1 | COSMIC Tier 1 genes significantly mutated in sarcomas.** Top 10 most significantly mutated genes from robust regression in each histology subtype intersected with the COSMIC Tier 1 gene list (A) shows the high mutation frequency of *TP53* across all histology subtypes and histology-specific mutations like *KIT*.

Tumor mutation burden of all non-silent mutations in sarcomas compared to TCGA PanCancer Study (B) show similar tumor mutation burden in ORIENT compared to TCGA SARC.

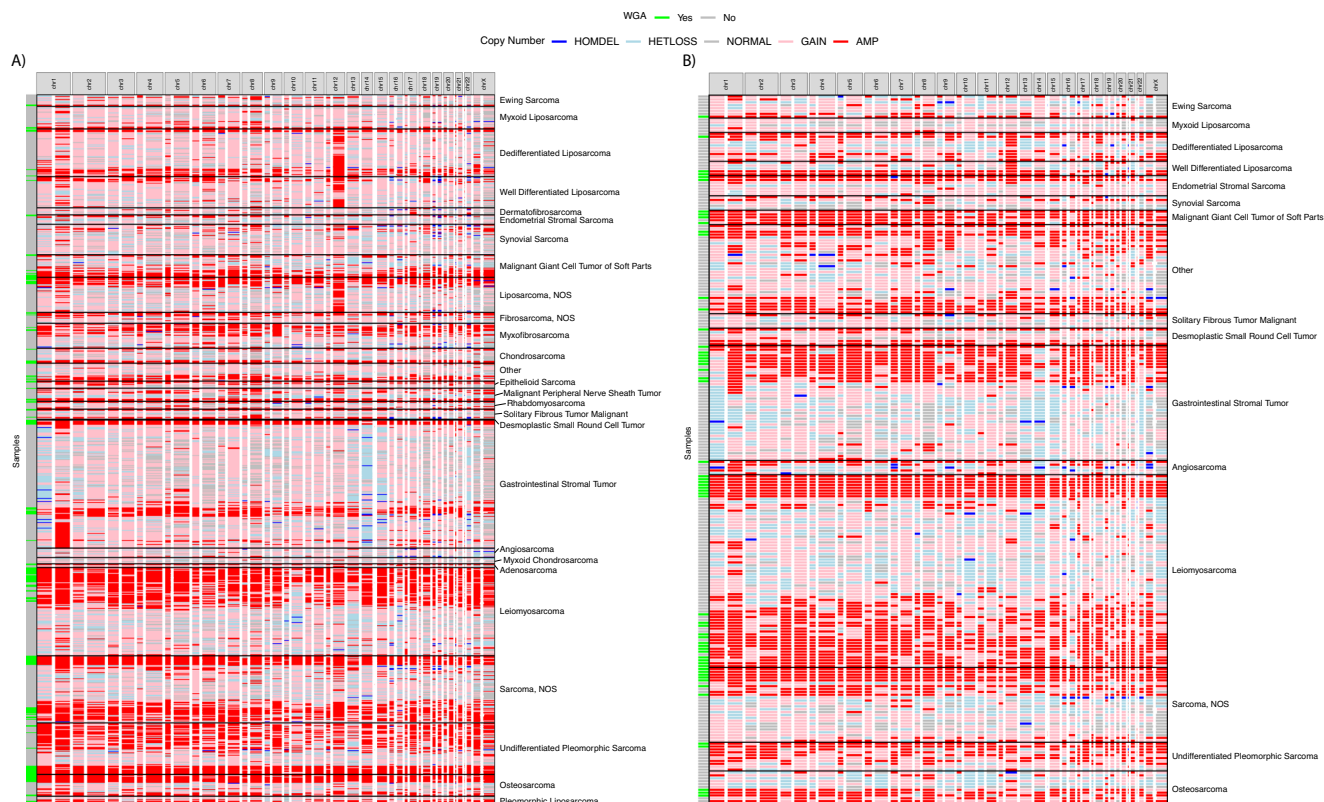
significantly higher burden observed in metastatic samples (1.88 mut/MB) compared to primary samples (1.34 mut/MB; Wilcoxon Rank Sum  $p < 0.001$ ). For histology subtypes with both primary and metastatic samples greater than 20 (GIST, Leiomyosarcoma, and Sarcoma, NOS), we compared TMB between disease sites. Leiomyosarcoma was the only of the three histology subtypes to display a significant increase in mut/MB in metastatic disease (Wilcoxon Rank Sum  $p$  value = 0.022), while GIST (Wilcoxon Rank Sum  $p$  value = 0.065) and Sarcoma, NOS (Wilcoxon Rank Sum  $p$  value = 0.725) did not.

Microsatellite-instability (MSI) was identified in 6 of 1162 WES samples (0.5%), including 4 sarcoma NOS, 1 leiomyosarcoma, and 1

malignant peripheral nerve sheath tumor. MSI tumors demonstrated a higher TMB (median = 11.0 mut/MB, range 1.1 to 56.5 mut/MB) than the overall cohort (median = 0.73 mut/MB, range 0 to 54.9 mut/MB). Non-silent mutations in mismatch repair (MMR) genes were identified in three of six MSI samples tumors, specifically in genes *MLH3*, *MSH3*, *MSH6*, *PMS2*, and *RFC3*.

### Copy Number Variations and Whole-Genome Amplification

In many sarcoma subtypes, copy number variations (CNV) are thought to be the primary driving events, such as recurrent *MDM2/CDK4* amplification in well-differentiated/de-differentiated liposarcoma



**Fig. 2 | Copy number variation landscape of sarcomas.** Arm-level copy number alterations in primary (A) and metastatic (B) disease. Green bar indicates samples that had >50% of segments lengths classified as having GAIN or AMP copy number.

(WD/DDLPs). To provide a landscape of CNV across our diverse cohort, we summarized CNV by chromosome arm with length-weighted average, plotted by sarcoma subtype and primary/metastatic status (Fig. 2). Known histology specific arm changes, including recurrent 12q gain in WD/DDLPs and recurrent loss of 14q in gastrointestinal stromal tumor (GIST), were readily identified. Overall, chromosome/segment level copy number gains were more frequent than losses (segment length with “gain” greater than segment length with “loss” in 860 of 1162 samples; 74%). Whole genome doubling (WGD) was a common feature across many sarcoma subtypes and this feature was more frequent in metastatic samples as compared to primary samples for both the entire study cohort (23.4% vs. 17.0%, Chi-squared  $p = 0.026$ ) and well as for GIST, specifically (22.5% vs. 9.15%, Chi-squared  $p = 0.038$ ). Furthermore, several COSMIC Tier 1 genes showed patterns consistent with recurrent arm losses at the histology subtype level, such as heterozygous losses of *PDGFB* and *FOXA1* in GIST samples (Supplementary fig. 2).

### Transcriptional Landscape and Disease Clustering

Transcriptome sequencing was assessed for global expression patterns across sarcoma subtypes. As expected, assessment of gene expression similarities via dimension reduction (UMAP) showed that tumors assigned to the same histology tended to be closer in the multidimensional embedding, providing further support to the histological classification of the samples (Fig. 3A). We chose initially to focus on several of the more common histologies of our dataset including leiomyosarcoma, GIST, myxoid liposarcoma, and dedifferentiated and well-differentiated liposarcomas (WD/DDLPs) that formed visually obvious distinct expression subgroups (Fig. 3B). We also noted that fusion-driven sarcoma subtypes including Ewing sarcoma, synovial sarcoma, DSCRT and solitary fibrous tumors comprised

Samples were clustered within their respective histology subtype, and the histology subtypes of were clustered for Primary disease.

transcriptomically distinct expression subgroups amongst this study cohort (Fig. 3A, Supplementary Fig. 3).

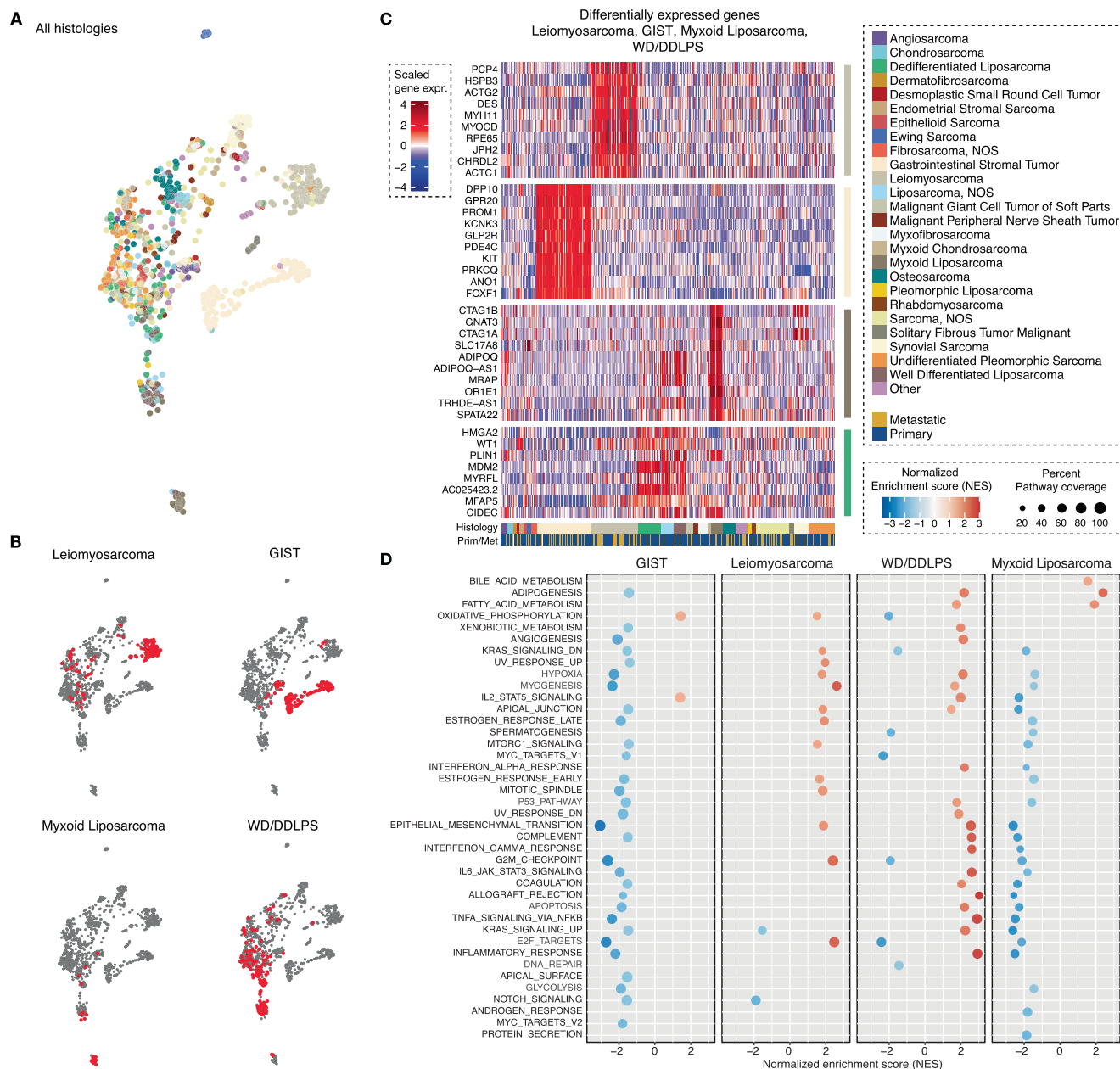
To further probe the expression differences driving these subgroups, we performed differential expression analysis with limma<sup>12</sup> comparing each of these divergent subtypes to the remaining samples (Fig. 3C, Supplementary fig. 4, Supplementary Data 2). As expected, disease subtypes maintained higher expression levels of markers of the cell lineage of origin e.g. genes related to smooth muscle differentiation such as *MYOCD*, *TAGLN*, *JPH2*, and *CSRPI* amongst the top differentially expressed genes in leiomyosarcoma and markers of adipogenesis in WD/DDLPs including *HMG2* and *CIDEA*. Known subtype-specific oncologic drivers were also amongst the top False Discovery Rate (FDR)-corrected differentially expressed genes, including *KIT* in GIST and *MDM2* in WD/DDLPs.

Gene set enrichment analysis was performed using the differentially expressed genes in these histology subtypes (Fig. 3D). In GIST, the top upregulated pathways were *IL2/STAT5* signaling and oxidative phosphorylation. Interestingly, in addition to adipogenesis and fatty acid metabolic pathways, DDLPs was highly enriched for multiple inflammatory pathways, including *JAK/STAT3* signaling and TNF $\alpha$  signaling via NF $\kappa$ B. As expected, the leiomyosarcoma group was enriched for myogenesis but also several cell cycle related pathways as well as mTORC1 signaling. Myxoid liposarcoma was marked by a relative negative enrichment of most oncologic signaling pathways evaluated, highlighting the expected unique oncogenic program of a transcription factor fusion-driven malignancy.

### Immunogenomic Profiling

To characterize the immune microenvironment across our diverse sarcoma cohort, we estimated the immune infiltration based on gene expression profiling of the bulk RNA sequencing followed by





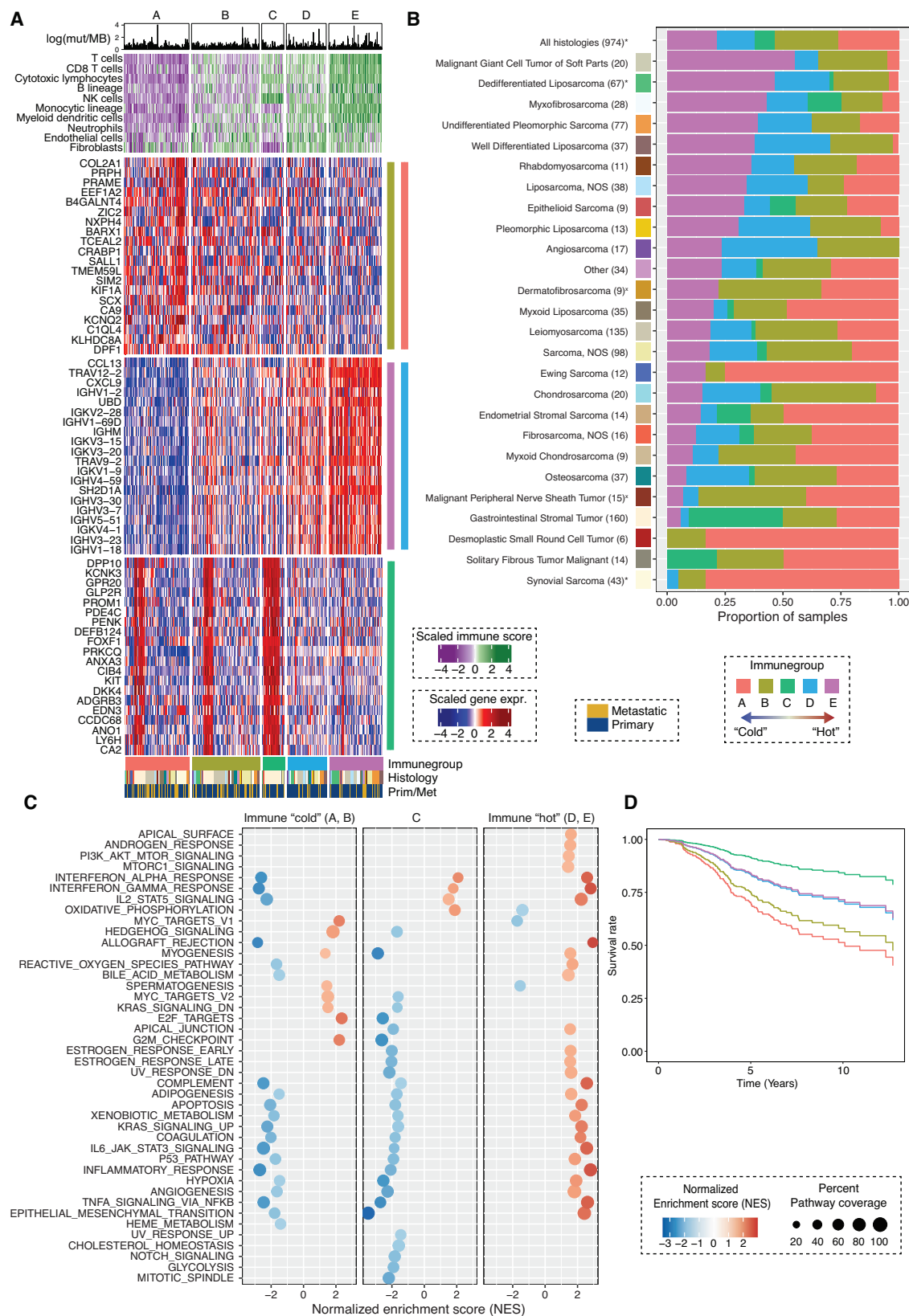
**Fig. 3 | The gene expression landscape of sarcomas.** Sarcomas were grouped according to their gene expression similarities and displayed in an UMAP projection (A). Four histologies (Leiomyosarcomas, GISTs, Myxoid liposarcomas, and dedifferentiated/well-differentiated liposarcomas) showed cohesive clusters (B).

Differential expression analysis indicated that each of the four histologies expressed a unique set of genes (C). Gene set enrichment analysis (GSEA) supported the high-level differences observed in differential expression analysis (D).

clustering analysis using these immune infiltration estimates from MCPCounter<sup>13</sup>. This analysis yielded five distinct immune groups (Fig. 4A) that exhibited a gradient of immune infiltration ranging from low immune cell enrichment or “cold” (Groups A and B) to high immune cell enrichment or “hot” (Groups D and E). Strikingly, the immune “intermediate” group C was almost exclusively composed of GIST (Fig. 4A, B). The enrichment scores for all 10 evaluated cell types were significantly different across the five immune groups (Kruskal-Wallis test FDR < 0.001). For most immune cell subtypes, there was a consistent and progressive increase from groups A to E, including in T (median score -1.18 to 1.21) and B cells (-0.71 to 0.81), dendritic cells (-1.17 to 0.89), and cells of monocyte lineage (-1.22 to 1.10). One notable exception is that Group C showed a specific infiltration enrichment for NK cells (Wilcoxon Rank Sum test FDR < 0.001) compared to all other groups except the most immune “hot” Group E, despite modest or

intermediate enrichment of other immune cell subtypes. Group B was marked by a higher abundance of endothelial cells and fibroblasts compared to groups A and C.

The distribution of tumors amongst immune subgroups varied by histology (Fig. 4B). Notably, angiosarcoma and undifferentiated pleomorphic sarcoma (UPS), diseases that have been reported to be more responsive to checkpoint inhibitor therapy, clustered more predominantly in the immune hot groups D and E. Within individual sarcoma subtypes, metastatic vs. primary status did not significantly associate with a difference in the distribution of immune subgrouping with the exception of dedifferentiated liposarcoma and synovial sarcoma. In DDLPS, metastatic samples were depleted of Group E tumors and in synovial sarcoma, primary samples were predominantly assigned to Group A (Fisher’s test  $p < 0.05$ ) (Fig. 4B, Supplementary fig. 5). In sarcomas with high immune infiltration (Group E) and with



**Fig. 4 | Sarcomas samples classified according to the level of immune infiltration.** Cell type scores (MCPcounter) showed five distinct immune groups of sarcomas, each with a distinct transcriptional profile. Tumor mutation burden (TMB), estimated as log(mut/MB) was slightly higher in the immune group E (A). Each histology in the data set was composed by different proportions of the five immune groups (B). GSEA also highlighted the differences in gene expression among the

immune groups (C). Survival analysis of patients with tumors in the different immune groups showed higher overall survival for those with immune group C-E tumors (D). Asterisks (\*) in (B) indicate  $p$ -value  $< 0.05$  from two-sided Fisher's Exact tests for association between number of samples in each immune group and number primary/metastatic samples. An "x" next to the histology indicated that the Fisher's test could not be completed.

high fibroblast content (Group B), there was a trend towards higher TMB (median  $\log(\text{TMB} + 1)$ : 0.44 and 0.50 respectively), compared to other immune groups ( $\log(\text{TMB} + 1)$ : 0.39 – 0.43), but the difference was not statistically significant (Kruskal-Wallis test  $p = 0.58$ ).

We performed differential expression analysis between immune subgroup A/B (“cold”), C (“intermediate/GIST”) and D/E (“hot”). Tumors in groups D and E (“hot”) showed upregulation of immune relevant genes *PTPRC*, *IL2RG*, *CD53*, *CCL13*, *TRAV12-1*, HLA genes, and multiple immunoglobulins (Fig. 4A and Supplementary Data 3). In contrast, we observed in Group A and B tumors upregulation of genes predominantly involved in tumor proliferation, such as *CIQL4*, *PRAME*, *KLHDC8A*, and *CA9* (Fig. 4A). Gene set enrichment analysis (Fig. 4C) confirmed the inverse relationship between gene sets related to cell cycle/proliferation (consistently enriched in groups A/B) and those related to immune signaling and response (consistently enriched in groups D/E). We further identified enrichment of signaling pathways potentially contributing to the observed immunological clustering, including upregulation of MYC targets and Hedgehog signaling and downregulation of KRAS signaling in the immune cold groups A/B.

Differentially expressed genes in Group C were driven mostly by histology, predominantly resulting in genes also upregulated in the GIST histology. Hence, we tested for differentially expressed genes only among GIST specimens belonging to the various immune subgroups (Supplementary Data 4). Similar to the patterns observed across the entire study cohort, GISTs that cluster in groups A and B showed enriched expression of cell proliferation and apoptosis-related genes such as *TYRO3* and *CSRNP3*, whereas GISTs that cluster in Groups D and E exhibited upregulation of genes involved in immune regulation and/or inflammation (e.g., *CSF3R*, *ANXA1*, *ALOX5*, *ITPR3*, *CXCL1*, *CXCL3*, *ZEB2*). In group C GIST tumors, we detected upregulated expression of immune-related genes associated with NK cells, including *KLRB1*, *KLRC1*, *KLRC3*, *KLRF1*, *KLRK1*, *GZMA*, *GZMK*, *NCRL*, *NKG7*.

Using matched clinical data, we performed survival analysis to evaluate the prognostic impact of immune subgroups. This analysis revealed that there was a significant difference between groups in overall survival, whether considered for the entire cohort (Fig. 4D, Likelihood Ratio Test – LRT  $p < 0.01$ ) or stratified by primary/metastatic status (Supplementary fig. 6, LRT  $p < 0.01$ ). Patients with tumors assigned to Group C had higher overall survival, which was expected given the predominance of GIST tumors in this group as GIST has greater survival expectations compared to most other sarcoma subtypes due to the availability of highly effective tyrosine kinase inhibitor therapy and also excellent expected survival of early-stage tumors included in the cohort treated with curative surgery. Patients with tumors in the immune hot Groups D and E had higher overall survival compared to those with tumors in immune cold Groups A and B. After excluding GISTs from the survival analysis, the same pattern was observed, in spite of non-statistical significance (Supplementary fig. 6).

## Discussion

In this study we analyze the mutational, transcriptomic and immunogenomic landscape of over 1,300 adult sarcomas of diverse histologic representation. The molecular profiling generated from this work provides a rich resource, expanding upon prior foundational work such as TCGA<sup>10</sup> towards a more comprehensive molecular profiling of this set of diseases. We provide a broad scope of findings to further our molecular understanding of previously understudied sarcoma subtypes, perform cross-sarcoma comparisons to highlight the divergent biology inherent to sarcoma, and generate novel insights into the underpinnings of these diseases.

An overarching theme of our analysis is the observation of differences in molecular findings between samples taken from primary versus metastatic tumor samples. We observe a modest increase in mutational burden between these groups that is primarily driven by an increase in inactivating mutations of tumor suppressor genes rather

than an accumulation of additional oncogene drivers. Whole genome doubling (WGD), a known frequent copy number variation pattern in sarcoma, has previously been proposed as a potential negative prognostic feature<sup>6</sup>, though prior analyses have been limited by small sample sizes with matched clinical outcomes, particularly after subgrouping. Concordantly, we find WGD more frequently in metastatic versus primary samples in our cohort. Interestingly, we found this difference most pronounced in GIST, a tumor type primarily driven by activating kinase mutations, suggesting that CNV and WGD specifically might be an important source of secondary mutation leading to tumor progression in this disease.

Transcriptomic profiling of our cohort of sarcomas highlights the diversity of this group of diseases. As expected, sarcoma subtypes cluster predominantly by disease. The sarcomas that form the most distinct transcriptional clusters include those known to be driven by a transcription factor fusion such as Ewing sarcoma and myxoid liposarcoma. Comparing the expressional landscape of one sarcoma subtype to others may help to identify disease specific pathways of interest for further testing and therapeutic development, such as the identification of mTORC1 upregulation in leiomyosarcoma that was identified in this work and is already a target for therapeutic development<sup>14</sup>. To our knowledge this study represents the largest database of whole transcriptome sequencing for adult sarcomas reported in the literature to date and therefore a resource for hypothesis generating comparisons in additional sarcoma types beyond the scope of this initial landscape.

Success of immunotherapy in an increasing number of advanced solid tumors has led to a desire to better understand the tumor-immune microenvironment. In sarcoma, an immune subgroup schema has been proposed based on a combined analysis of RNA sequencing (from TCGA) and several microarray based expressional studies totaling 608 tumors total and encompassing 6 sarcoma histologies<sup>7</sup>. Our work differs from this prior in that it includes a greater diversity of sarcoma subtypes, is larger in sample size, and is restricted only to the analysis of RNAseq. Concordant to this prior work, we find that immune cell enriched sarcoma clusters have a more favorable prognosis than immune cell deplete sarcoma clusters. Intriguingly, in our analysis, we find that most GIST form a unique cluster that has intermediate levels of immune cell enrichment overall but is specifically enriched for NK cell infiltrate. To our knowledge this specific association of GIST with NK cell infiltrate relative to other sarcoma subtypes has not been previously reported or studied, though NK cells have been reported to be relatively abundant in GIST based on prior transcriptomics analyses<sup>15,16</sup> as well as by immunohistochemistry<sup>17</sup>. NK cell activity in GIST has also been shown to be predictive of tyrosine kinase inhibitor efficacy and patient survival<sup>17,18</sup>. Further study is indicated to assess the clinical relevance of our immune subtyping schema, compare and/or combine our data with prior datasets for refinement of sarcoma immune subtyping, and to confirm immune associations identified in this study such as that between NK cells infiltrate and GIST.

We acknowledge several important limitations of our study. The diversity of our cohort, while overall a strength, equates to having small sample sizes ( $\leq 5$  tumors each) for many of the less common disease subtypes. For the purposes of this initial landscape analysis, we bluntly categorized these lower frequency sarcomas into the category of “other,” which will obviously limit disease specific insights for these rarer sarcomas. Additionally, while both primary and metastatic samples were available in this study, very few of these were matched samples from the same patient. Finally, extended clinical data including therapeutic outcomes were still being curated and therefore not available for this analysis, limiting the ability to assess for molecular determinants of response. Despite these limitations, we believe that our study has many strengths and fills an important gap in molecularly profiling efforts in sarcoma.

In conclusion, we report a landscape analysis of whole exome sequencing and RNA sequencing performed on >1300 diverse sarcomas, representing one of the largest molecular profiling efforts in this disease group. We identify the mutational and copy number variation landscape and observe differences between primary and metastatic samples. We highlight expression pathways that are enriched in histologic subtypes that cluster most distinctly from others, providing a subtype-specific roadmap for further translational efforts. Finally, we define immune enriched or depleted sarcoma subgroupings that carry a prognostic impact.

## Methods

### Patient Population and Study Design

The study was approved by the Institutional Review Board (IRB) at Moffitt Cancer Center (Advarra, IRB#00000971, Pro00048786). The Oncology Research Information Exchange Network (ORIEN) is an alliance of 18 U.S. cancer centers established in 2014. All ORIEN alliance members utilize a standard Total Cancer Care® (TCC) protocol that is approved by their local IRB. As part of the TCC study, participants undergo written informed consent to agree to have their clinical data followed over time, to undergo germline and tumor sequencing, and to be contacted in the future by their provider if an appropriate clinical trial or other study becomes available. TCC is a prospective cohort study with a subset of patients enrolled to the ORIEN Avatar program, which includes research use only (RUO) grade whole-exome tumor sequencing, RNA sequencing, germline sequencing, and collection of deep longitudinal clinical data with lifetime follow up. Aster Insights, the commercial and operational partner of ORIEN, harmonizes all abstracted clinical data elements and molecular sequencing files into a standardized, structured format to enable aggregation of de-identified data for sharing across the Network. 1242 ORIEN Avatar patients diagnosed with sarcoma and consented to the TCC protocol from 13 participating members of ORIEN were included in this study. No sex and/or gender analysis was carried out as this was felt to be out of the scope of this report.

### DNA/RNA Preparation and Sequencing

Tumor DNA was purified from frozen tissue and germline DNA was purified from blood using QIASymphony DNA purification kit (Qiagen; Venlo, The Netherlands), followed by ultrasonification to an average size of 213 bp (Covaris; Woburn, MA). Alternatively, tumor DNA was purified from formalin fixed paraffin embedded (FFPE) tissue using truXTRAC FFPE DNA purification kit (Covaris, Woburn, MA), generating an average size of 165 bp. RNA was purified from frozen tissue with RNAeasy plus mini (Qiagen; Venlo, The Netherlands) to generate an average size of 216nt; alternatively, RNA was purified from FFPE tissue using truXTRAC FFPE RNA ultrasonification kit (Covaris; Woburn, MA) generating an average size of 165nt.

DNA libraries were generated, then captured on custom WES arrays (IDT; Coralville, IA (38.7 Mb), or Nimblegen; Madison, WI (63.38 Mb)), with custom probes designed to increase capture of up to 676 cancer-related genes. Libraries were sequenced (100 bp, paired end reads) on Illumina NovaSeq 6000 (Illumina, San Diego, CA) to a target depth of 100x (200x for cancer-related genes) for germline WES, and 300x (600x for cancer-related genes) for tumor WES. Quality checks were performed for gender identity matches. Minimum hybrid selection threshold was set at >80% of bases with >100x coverage for tumor samples and >50x coverage for normal samples. RNA libraries were generated using TruSeq RNA Exome adapters (Illumina; San Diego, CA), captured on exome array, and sequenced (100 bp, paired end reads) at Hudson Alpha (Huntsville, AL) or sequenced (150 bp, paired end reads) at Fulgent (Temple City, CA) to a target depth of 100 M reads (50 M pairs).

### WES Quality Control and Alignment

Adapter sequences were trimmed with BBDuk (version 38.46; [sourceforge.net/projects/bbmap/](https://sourceforge.net/projects/bbmap/)). WES samples were processed using the Sentieon App (sentieon\_release\_201911)<sup>19</sup>. Fastq files were aligned to human reference genome (GRCh38/hg38) using BWA-mem<sup>20</sup>. Resulting cram files underwent extensive quality control checks using Picard (v2.9.0).

### Identification of Somatic Mutations

Somatic single nucleotide variants (SNVs) and insertions/deletions (INDELs) were called with TNseq (Sentieon; San Jose, CA) using matched tumor and germline cram files. Resulting vcf files were annotated with Annotator<sup>21</sup> and Funcotator (GATK v4.1.6.0; <https://gatk.broadinstitute.org/hc/en-us/articles/360037224432-Funcotator>).

### Filtering Somatic Mutations and Calculation of Tumor Mutation Burden

Somatic variants were filtered for population polymorphisms and recurrent sequencing artifacts using Aster Insights Panel of Normals (PoN). Aster Insights PoN is constructed from the ORIEN AVATAR germline variant catalog and is applied by checking every sample's reported somatic mutation against the catalog to add a VCF "panel\_of\_normals" flag. A germline variant is included in the PoN if it is present in > 0.5% of ORIEN AVATAR population of unrelated normal samples. For a variant to pass all filters, it must not be present in the PoN, have a variant allele fraction (VAF) greater than 0.04 (4%), at least 1 ALT read in both the F1R2 and F2R1 read orientations, and F1R2 + F2R1 must be greater than 10 ALT reads.

Protein coding length for all genes in HG38 (bp) were calculated using a custom python script. For each collapsed histology subtype, these lengths were modeled with robust regression against frequency of samples that have mutations in that gene. Robust regression was used due to its ability to better handle outliers<sup>22</sup>. Oncoplots were constructed with *maftools*, ignoring variant classification of 'silent' or 'unknown'<sup>23</sup>. Histology subtypes were collapsed to "Other" if they were composed of less than 5 samples. Top 10 most significant genes from robust regression were selected for each histology subtype and intersected with COSMIC Tier 1 genes to identify cancer-related genes<sup>24</sup>. TMB was calculated with 'tcgaCompare', adjusting for WES kit capture size. Comparisons of mutation frequency between primary and metastatic samples was performed with two-tailed Wilcoxon Rank Sum tests. Tumor mutation burden was classified as low (< 5 mut/MB), intermediate (≥ 5 mut/MB and <10 mut/MB), and high (≥ 10 mut/MB)<sup>25</sup> for ORIEN and TCGA SARC samples then compared with two-tailed Fisher's Exact tests due Chi-squared test producing a warning about an expected value being less than 5.

### Copy Number Calling

Somatic Copy Number Variation (CNV) was calculated by Sequenza (v3.0.0, <https://sequenzatools.bitbucket.io/#/home>)<sup>26</sup>. Files were generated on the alignments of tumor and germline reads by sequenza-utils and processed by CNTools v1.30.0 to generate the final output for copy number at the gene level<sup>27</sup>. Copy number was classified into homologous deletion (HOMDEL, 0), heterozygous loss (HETLOSS, 1), normal (Normal, 2), gain (GAIN, 3), or amplification (AMP, 4 or greater). Samples were determined to have whole genome doubling (WGD) if more than 50% of all segment lengths were classified as amplified (i.e. CN=4 or greater). Differences between the number of primary and metastatic samples with WGD was calculated using two-tailed Chi-squared tests for all samples, as well as the top 5 histology subtypes by sample number for consistency. Arm-level copy number change was calculated as the weighted average of copy number segments on each arm (i.e. longer segments contribute more to arm-level copy number) and then classified as outlined above.



## Microsatellite Instability

Microsatellite instability is scored using MSIsensor 2 (Niu, et al. (2014)). Briefly, the median number of microsatellites assessed per tumor-normal pair was 21,648 (range: 9954–34,873), and MSI scores are reported as the number of unstable microsatellites divided by the total number of microsatellites assessed. For downstream analyses, a threshold of 18% of sites being somatic (somatic microsatellites divided by total number of microsatellites) was used to separate high MSI and stable MSI. TMB was filtered to include non-silent mutations while removing transcriptional start site and splice site mutations before calculating differences between high and stable MSI samples.

## RNA-seq Quality Control and Alignment

Adapter sequences were trimmed with BBDuk (version 38.46; [sourceforge.net/projects/bbmap/](https://sourceforge.net/projects/bbmap/)) and fastq files were aligned to the human reference genome (GRCh38/hg38 with Gencode transcriptome version 32) using STAR (v.2.7.3a, <https://github.com/alexdobin/STAR>)<sup>28</sup>. Resulting RNAseq expression data were then normalized at the TPM level, using a linear scaling method based on the trimmed mean among protein-coding genes (mean of TPMs between the 35th and 95th quantiles) of each sample, log2 transformation of scaled TPMs after adding 0.001, loess smoothing, and ‘ComBat’ normalization (*sva* R package, v3.34.0), adjusting for batch effects induced by sample preservation methods and sequencing capture kit (Supplementary fig. 7)<sup>29</sup>. The resulting isoform-level counts were then collapsed to gene-level by summing the transcripts originating from a given gene. Gene counts mapped to Y RNAs, Metazoa SRP, 5.8S and 5S rRNA, 7SK, Vault, SnoRNAs MIR, RNAs, and RNU were removed.

To ensure that repeated sampling from the same patient would not overly bias results of downstream RNA-seq analysis, we repeated selected analyses with removal of repeated measures (by random selection of 1 tumor per patient when >1 tumor was available for analysis). The results closely mirrored the primary analysis, supporting the inclusion of all available tumors for the analysis (Supplementary fig. 8).

## RNA Gene Fusion Prediction and Identification of Histology Subtype Changing Fusions

STAR-Fusion and Arriba Gene Fusion algorithms are applied to the STAR aligner output files. Gene Fusion predictions from both STAR-Fusion (v1.8.0, <https://github.com/STAR-Fusion/STAR-Fusion/wiki>)<sup>30</sup> and Arriba (v1.1.0, <https://github.com/suhrig/arriba>)<sup>31</sup> were merged into a single output file that removes duplicate putative gene fusion calls, putative gene fusion calls of low confidence – reporting gene fusions with at least one (1) junction read and at least one (1) spanning read, and gene fusion calls occurring within the same gene, within SnoRNAs, within rRNAs, or mitochondrial genes – which are areas considered to be contributing to high false-positive rate and generally uninformative. Fusions in rRNA, mitochondrial genes, and fusions within the same genes were removed. Fusions with poor supporting evidence in the spanning and junction reads (<1 read), as well as gene fusions that have multiple break-points were removed. Merged putative fusion calls were coalesced to the most dominate gene isoform in a set of reported putative gene fusion calls that either share an identical breakpoint or within a set of overlapping putative gene fusion calls.

Fusions associated with histology reassignment were derived from Gounder et al. These were then intersected with our fusion calls, ignoring donor-acceptor gene order to identify candidate samples in need of histology reassignment. List of samples along with clinical attributes were then reviewed to determine whether change in histology assignment was appropriate given fusion confidence and disease location.

## Assessment of Gene Expression and Immune Infiltration

After batch correction and normalization procedures, we applied uniform manifold approximation and projection (UMAP) using the expression from the 10% of genes with the highest standard deviation

to explore the gene expression similarities among sarcoma samples. The umap R package was used with spread=2 and random\_state=12345. Given the separation of Leiomyosarcomas, Gastrointestinal stromal tumors (GIST), Myxoid liposarcomas, and liposarcomas (well- and dedifferentiated), the determination of differentially expressed genes between each of those sarcoma groups and the rest of the samples was completed using linear models as implemented in the limma package<sup>12</sup>. The false discovery rate (FDR) correction was used to adjust for multiple comparisons. Adjusted *p*-values were used in the fgsea package to calculate enrichment scores for the Hallmark gene sets<sup>32</sup>.

Immune infiltration scores were calculated with the deconvolution algorithm MCPcounter as conducted in a previous sarcoma study<sup>7</sup>. MCPcounter provides abundance scores for multiple populations of immune cells, as well as other stromal cell types. In order to detect sarcoma immune groups, the MCPcounter scores were used in hierarchical clustering with Ward’s linkage. Unsupervised detection of clusters was performed using hybrid Dynamic Tree Cuts with deepSplit=1.75<sup>33</sup>. We collapsed the determined immune groups into three categories: Immune “cold” (Groups A and B), immune “hot” (Groups D and E), and Group C. Differential gene expression analysis was conducted among the three collapsed immune groups by comparing the samples from each group against the samples in the two other groups. We also used FDR-adjusted *p*-values to calculate GSEA scores for the Hallmark gene sets. Analyses and visualizations were generated in R<sup>34</sup> and using ggplot2<sup>35</sup> and ComplexHeatmap<sup>36</sup>.

## Immune group survival analysis

For survival analysis, the date of diagnosis was used as the left endpoint. For each patient, the sarcoma diagnosis most proximal in time to the date of sample collection was used to determine the diagnosis date (using a search window that extended five years prior to and up to one month beyond the date of sample collection). The date of death or date of last contact was used as the right endpoint (censored for date of last contact). Samples without a matching diagnosis record or unknown right endpoint were excluded. In total, 781 samples were included in the survival analysis. Cox proportional hazard models were fitted to assess the differences in survival of patients with sarcomas from each of the immune groups. In addition to testing for the effects of immune group assignment, the models included terms accounting for the sarcoma histology and tumor stage. To test for significant differences between the overall survival of each immune group, Likelihood ratio tests (LRTs) were conducted. Analyses were performed in R with the survival and survminer packages<sup>37</sup>.

## Statistics and Reproducibility

No statistical power analysis was completed to predetermine the study sample size. No data were excluded from analyses. The experiments were not randomized. All statistical analyses were completed using R and RStudio. Differential expression analysis was performed by fitting linear models using the limma (doi:10.1093/nar/gkv007) and edgeR (doi:10.1093/nar/gkaf018) R packages. The Kruskal-Wallis test was used to assess differences in immune cell enrichment scores across the five identified immune groups. For specific comparisons, the two-sided Wilcoxon Rank Sum test was applied to evaluate differences in NK cell infiltration between Group C and other groups. Two-sided Fisher’s Exact tests were used to determine the association between immune subgroup distribution and metastatic versus primary status for certain sarcoma subtypes. Survival analyses were conducted using Cox proportional hazard models, with likelihood ratio tests (LRTs) to assess differences in overall survival among immune groups. For all analyses, *p*-values were adjusted using the Benjamini-Hochberg method for multiple comparisons (i.e., False Discovery Rate or FDR), unless otherwise stated. All other statistical methods are described in their respective methods sections.

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The data used in this research was generated through private funding by Aster Insights ([www.asterinsights.com](http://www.asterinsights.com)) in collaboration with the Oncology Research Information Exchange Network (ORIEN, [www.oriencancer.org](http://www.oriencancer.org)). Requests for access to the data used in this study can be submitted to the corresponding author and ResearchDataRequest@AsterInsights.com.

## Code availability

Code for analyses can be found at: [https://github.com/FridleyLab/landscape\\_sarcomas\\_orien](https://github.com/FridleyLab/landscape_sarcomas_orien) and <https://doi.org/10.5281/zenodo.14851789>.

## References

- Kallen, M. E. & Hornick, J. L. The 2020 WHO Classification: What's New in Soft Tissue Tumor Pathology? *Am. J. Surg. Pathol.* **45**, e1–e23 (2021).
- von Mehren, M. et al. Soft Tissue Sarcoma, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology. *J. Natl Compr. Canc Netw.* **20**, 815–833 (2022).
- Gounder, M. M. et al. Clinical genomic profiling in the management of patients with soft tissue and bone sarcoma. *Nat. Commun.* **13**, 3406 (2022).
- Chang, W. et al. MultiDimensional ClinOmics for Precision Therapy of Children and Adolescent Young Adults with Relapsed and Refractory Cancer: A Report from the Center for Cancer Research. *Clin. Cancer Res.* **22**, 3810–3820 (2016).
- Boddu, S. et al. Clinical Utility of Genomic Profiling in the Treatment of Advanced Sarcomas: A Single-Center Experience. *JCO Precis Oncol.* **2**, 1–8 (2018).
- Nacev, B. A. et al. Clinical sequencing of soft tissue and bone sarcomas delineates diverse genomic landscapes and potential therapeutic targets. *Nat. Commun.* **13**, 3405 (2022).
- Petitprez, F. et al. B cells are associated with survival and immunotherapy response in sarcoma. *Nature* **577**, 556–560 (2020).
- Brohl, A. S. et al. The genomic landscape of the Ewing Sarcoma family of tumors reveals recurrent STAG2 mutation. *Plos Genet* **10**, e1004475 (2014).
- Bridge, J. et al. Clinicopathologic and Genomic Features of EWSR1-PATZ1 Fusion “Sarcoma”. *Lab Invest* 99 (2019).
- Cancer Genome Atlas Research Network. Electronic address, e. d. s. c. & Cancer Genome Atlas Research, N. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell* **171**, 950–965.e928 <https://doi.org/10.1016/j.cell.2017.10.014> (2017).
- Baird, K. et al. Gene expression profiling of human sarcomas: insights into sarcoma biology. *Cancer Res.* **65**, 9226–9235 (2005).
- Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
- Becht, E. et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).
- Lacuna, K., Bose, S., Ingham, M. & Schwartz, G. Therapeutic advances in leiomyosarcoma. *Front Oncol.* **13**, 1149106 (2023).
- Mao, X. et al. Single-cell transcriptome analysis revealed the heterogeneity and microenvironment of gastrointestinal stromal tumors. *Cancer Sci.* **112**, 1262–1274 (2021).
- Zhang, Q. et al. New insight on the correlation of immune landscapes with immune markers expression in different risk classification of gastrointestinal stromal tumors. *J. Gastroenterol.* **58**, 527–539 (2023).
- Delahaye, N. F. et al. Alternatively spliced NKp30 isoforms affect the prognosis of gastrointestinal stromal tumors. *Nat. Med.* **17**, 700–707 (2011).
- Menard, C. et al. Natural killer cell IFN-gamma levels predict long-term survival with imatinib mesylate therapy in gastrointestinal stromal tumor-bearing patients. *Cancer Res.* **69**, 3563–3569 (2009).
- Donald, F., Rafael, A., Jessica, A. W. & Jeremy, S. E. The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv*, 115717. <https://doi.org/10.1101/115717> (2017).
- Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
- Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
- Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. Fourth edn, (Springer, 2002).
- Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756 (2018).
- Tate, J. G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
- Marabelle, A. et al. Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study. *Lancet Oncol.* **21**, 1353–1365 (2020).
- Favero, F. et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
- CNTTools: Convert segment data into a region by sample matrix to allow for other high level computational analyses v. 1.30.0 (2023).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
- Haas, B. J. et al. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **20**, 213 (2019).
- Uhrig, S. et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* **31**, 448–460 (2021).
- Korotkevich, G. et al. Fast gene set enrichment analysis. *bioRxiv*, 060012 <https://doi.org/10.1101/060012> (2021).
- Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).
- R: A language and environment for statistical computing v. 4.1.2; 4.3.0 (R Foundation for Statistical Computing, 2023).
- ggplot2: *Elegant graphics for data analysis* (Springer-Verlag New York, 2016).
- Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
- survminer: Drawing Survival Curves using ‘ggplot2’ (2021).

## Acknowledgements

Funding for publication of this work was supported by the AmandaLee Sarcoma Fund, Moffitt Cancer Center.

## Author contributions

A.S. and O.E.O. drafted the manuscript and performed bioinformatics and statistical analysis. O.H., M.R., D.H., D.M., P.A., S.Z., N.S., and J.K.T.

provided bioinformatics analysis and support. M.C. and E.M.S. provided coordination efforts for the study. R.P., D.L., J.L.C., G.T., B.S., M.M., B.A.W., B.J.M., M.J.C., J.S.G., B.P.S., G.R., S.B.E., C.A.M., K.C., A.R.N., R.J.G., J.E.M., D.M.J., O.B., G.D.L., A.O.N., M.D., and D.R.R. contributed to patient recruitment for biospecimens. B.L.F. supervised bioinformatics analysis and contributed to study design. A.S.B. conceived and designed the study, drafted the manuscript, performed data analysis, and supervised the study. All authors critically reviewed and edited the manuscript and approved of the final version.

## Competing interests

O.H., M.C., M.R., D.H., D.M., P.A. and S.Z. are or were employed by Aster Insights. E.S. received research funding from the ORIEN Foundation. D.L. – advisory board with Aadi Biosciences and patent licensing (US 10,65,347 B2: *Method for Predicting Prognosis*) with MatchTx, Inc. J.L.C. is an employee of Tempus. G.T. – advisory board with SynOx, Deciphera, Daiichi Sankyo. B.A.W. – consulting or advisory role with Deciphera, Epizyme, Adcendo, Polaris, Boehringer Ingelheim, research funding from Exelixis, and travel expenses from Agenus. G.R. – advisory board with AstraZeneca. A.R.N. – advisory board with Foundation Medicine and NGM biosciences, travel compensation from Foundation Medicine and Caris Life Sciences, D.R.R. – data safety monitoring committee for Eisai and Springworks. M.D. – consulting AdaptImmune, Deciphera, and Aadi Biosciences. A.S.B. – advisory board with Deciphera, research funding (institution) from Merck. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-58678-6>.

**Correspondence** and requests for materials should be addressed to Andrew S. Brohl.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

<sup>1</sup>Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL, USA. <sup>2</sup>Aster Insights, Hudson, FL, USA. <sup>3</sup>Department of Biomedical and Health Informatics, School of Medicine, University of Missouri, Kansas City, MO, USA. <sup>4</sup>Department of Pharmacotherapy and Translational Research, College of Pharmacy, University of Florida, Jacksonville, FL, USA. <sup>5</sup>Division of Surgical Oncology, The Ohio State University Comprehensive Cancer Center, College of Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA. <sup>6</sup>Division of Medical Oncology, The Ohio State University Comprehensive Cancer Center, College of Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA. <sup>7</sup>Department of Biology, Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. <sup>8</sup>University of Colorado School of Medicine, Aurora, CO, USA. <sup>9</sup>Department of Orthopaedics and Rehabilitation, University of Iowa, Iowa City, IO, USA. <sup>10</sup>Department of Surgery, University of Kentucky, Lexington, KY, USA. <sup>11</sup>University of Utah, Huntsman Cancer Institute, Salt Lake City, UT, USA. <sup>12</sup>Indiana University Simon Comprehensive Cancer Center, Indianapolis, IN 46202, USA. <sup>13</sup>Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA. <sup>14</sup>Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA. <sup>15</sup>Department of Pathology, University of Virginia, Charlottesville, VA, USA. <sup>16</sup>Division of Surgical Oncology, Emory University, Atlanta, GA, USA. <sup>17</sup>Medical Oncology/Phase 1 program, Stephenson Cancer Center, University of Oklahoma Health Sciences, Oklahoma City, OK, USA. <sup>18</sup>Sarcoma Department, Moffitt Cancer Center, Tampa, FL, USA. <sup>19</sup>Department of Radiation Oncology, Moffitt Cancer Center, Tampa, FL, USA. <sup>20</sup>Department of Individualized Cancer Management, Moffitt Cancer Center, Tampa, FL, USA. <sup>21</sup>Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL, USA. <sup>22</sup>Division of Health Services and Outcomes Research, Children's Mercy, Kansas City, MO, USA. <sup>23</sup>These authors contributed equally: Alex Soupir, Oscar E. Ospina.

✉ e-mail: [andrew.brohl@moffitt.org](mailto:andrew.brohl@moffitt.org)