Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A single-cell atlas of cancer-educated ecotypes across high-risk pediatric sarcomas

Christian Seitz

christian.seitz@kitz-heidelberg.de

University Childrens Hospital Tuebingen

Karla Catacora Castañeda

Hopp Children's Cancer Center (KiTZ)

Sophia Scheuermann

Hopp Children's Cancer Center (KiTZ)

Tom Fischer

Hopp Children's Cancer Center Heidelberg https://orcid.org/0000-0002-2138-9102

Fabienne Engelmann

Hopp Children's Cancer Center (KiTZ)

Anke King

Hopp Children's Cancer Center (KiTZ)

Roland Imle

Hopp Children's Cancer Center Heidelberg (KiTZ) / German Cancer Research Center (DKFZ) Heidelberg
https://orcid.org/0000-0001-9564-6952

Alice Nuernbergk

University Children´s Hospital Tübingen

Beate Kristmann

University Hospital Tübingen https://orcid.org/0009-0005-8299-5316

Jan-Philipp Mallm

German Cancer Research Center DKFZ https://orcid.org/0000-0002-7059-4030

Benjamin Mayer

University Children´s Hospital Tübingen

Stefanie Volz

Hopp Children's Cancer Center (KiTZ)

Kathrin Schramm

Hopp Children's Cancer Center (KiTZ); German Cancer Research Center (DKFZ)

David Jones

Hopp Children's Cancer Center Heidelberg (KiTZ) https://orcid.org/0000-0002-2036-5141

Olaf Witt

Hopp Children´s Cancer Center Heidelberg (KiTZ) https://orcid.org/0000-0003-3918-9157

Robert Autry
Division of Pediatric Neurooncology, German Cancer Research Center (DKFZ) https://orcid.org/0000-0002-6965-2942
Michaela Nathrath
Klinikum Kassel https://orcid.org/0000-0002-1584-1115
Monika Sparber-Sauer
University Children´s Hospital Tübingen
Uta Dirksen
University Hospital Essen https://orcid.org/0000-0002-5435-7860
Stefan Pfister
Hopp Children's Cancer Center Heidelberg (KiTZ) https://orcid.org/0000-0002-5447-5322
Jörg Fuchs
Department of Pediatric Surgery and Urology, University Children´s Hospital Tuebingen
Supat Thongjuea
Hopp Children's Cancer Center (KiTZ); German Cancer Research Center (DKFZ)
https://orcid.org/0000-0002-9129-4694
Iman Sadeghi
Clinic for Internal Medicine II
Thomas Grünewald
German Cancer Research Center https://orcid.org/0000-0003-0920-7377
Kristian Pajtler
German Cancer Research Center (DKFZ), Heidelberg, German https://orcid.org/0000-0002-3562-6121
Kendra Maass
German Cancer Research Center

Article

Keywords:

Posted Date: July 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10374394/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

A single-cell atlas of cancer-educated ecotypes across high-risk pediatric sarcomas

Authors:

Karla Catacora Castañeda 1,2,3#, Sophia Scheuermann 1,3,4,5,6,7,8#, Tom T. Fischer 1,2,3,4#, Fabienne Engelmann 1,4,5, Anke King 1,2,3, Roland Imle 1,2,3,4,9, Alice Nuernbergk 5, Beate Kristmann 5, Jan-Philipp Mallm 10,11, Benjamin F.B. Mayer 12, Stefanie Volz 1,2,3, Iman Sadeghi 13, Kathrin Schramm 1,3,14, David T.W. Jones 1,3,14, Olaf Witt 1,3,4,15 , Robert J. Autry 1,2,3,4, Michaela Nathrath 16,17, Monika Sparber-Sauer 5,18, Uta Dirksen 19,20,21, Stefan M. Pfister 1,2,3,4, Jörg Fuchs 12, Supat Thongjuea 1,2,3, Thomas G. P. Grünewald 1,3,22,23#, Kristian W. Pajtler 1,2,3,4#*, Kendra K. Maass 1,2,3,4#*, Christian M. Seitz 1,3,4,5,7,15 #**

1 Hopp Children's Cancer Center (KiTZ), Heidelberg, Germany.
2 Division of Pediatric Neurooncology, German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ), Heidelberg, Germany.
3 National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and Heidelberg University Hospital, Heidelberg, Germany.
4 Department of Pediatric Oncology, Hematology, and Immunology, Heidelberg University Hospital, Heidelberg, Germany.
5 Department of Haematology, Oncology, Gastroenterology, Nephrology, Rheumatology, University Children's Hospital Tübingen, Tübingen, Germany.
6 Cluster of Excellence iFIT (EXC 2180) "Image-Guided and Functionally Instructed Tumor Therapies", University of Tübingen, Tübingen, Germany.
7 Clinical Cooperation Unit Pediatric Oncology (B310), German Cancer Research Center (DKFZ), Heidelberg, Germany.
8 German Cancer Research Consortium (DKTK), Partner Site Tuebingen, German Cancer Research Center (DKFZ), Heidelberg, Germany.
9 Division of Soft-tissue Sarcoma Research, Hopp Children's Cancer Center, Heidelberg (KiTZ), German Cancer Research Center (DKFZ), Heidelberg, Germany.
10 German Cancer Research Center (DKFZ).
11 Single Cell Open Lab, German Cancer Research Center (DKFZ), Heidelberg, Germany.
12 Department of Paediatric Surgery and Paediatric Urology, University Children's Hospital Tübingen, Tübingen, Germany.
13 Clinic for Internal Medicine II, Freiburg University Medical Center, Freiburg im Breisgau, Germany.
14 Division of Pediatric Glioma Research (B360), German Cancer Research Center (DKFZ), Heidelberg, Germany.
15 Deutsches Konsortium für Translationale Krebsforschung (DKTK), DKFZ, core center Heidelberg, Deutschland.
16 Department of Pediatrics and Children's Cancer Research Center, Klinikum rechts der Isar, Technical University of Munich, School of Medicine, Munich, Germany.

17 Department of Pediatric Hemato-/Oncology, Psychosomatics and Systemic Diseases, Klinikum Kassel, Kassel, Germany.
18 Stuttgart Cancer Center, Zentrum für Kinder-, Jugend- und Frauenmedizin (Olgahospital), Pädiatrie 5 (Pädiatrische Onkologie, Hämatologie, Immunologie), Klinikum der Landeshauptstadt Stuttgart, Stuttgart, Germany.
19 Pediatrics III, West German Cancer Center, University Hospital Essen, University Duisburg-Essen, Essen, Germany.
20 National Center for Tumor Diseases (NCT), NCT West, Site Essen.
21 German Cancer Consortium (DKTK) Site Essen, Cancer Research Center (NCT) Cologne-, Essen, Germany.
22 Division of Translational Pediatric Sarcoma Research, German Cancer Research Center (DKFZ), German Cancer Consortium (DKTK), Heidelberg, Germany.
23 Institute of Pathology, Heidelberg University Hospital, Heidelberg, Germany

# equal contribution
*corresponding author

These authors contributed equally: Karla Catacora Castañeda, Sophia Scheuermann, Tom T. Fischer
These authors jointly supervised this work: Thomas G.P. Grünewald, Kristian W. Pajtler, Kendra K. Maass, Christian M. Seitz

Abstract

High-risk pediatric sarcomas have a dismal prognosis. Current precision oncology approaches have not substantially improved patient outcomes. Immunotherapy offers a unique opportunity to develop more effective treatment strategies, but its successful implementation is hindered by a limited understanding of the tumor microenvironment (TME). We generated a TME classification platform by performing single-nucleus RNA sequencing of 89 pediatric sarcoma samples representing the major entities and integrating these data with high-dimensional spatial imaging. We identified distinct stromal, myeloid, and lymphoid cell states, revealing both entity-specific and conserved functional programs. Compositional analyses demonstrated marked differences in immune and stromal infiltration, with inflammatory myeloid cells dominating most tumors. We defined twelve TME archetypes and identified convergent multicellular niches, including stroma-, vasculature-, and T cell-rich regions, validated by multiplex imaging. Cell-cell interaction analyses uncovered shared and entity-specific communication networks, providing a comprehensive atlas of pediatric sarcoma TMEs and a framework for rational immunotherapeutic development.

Main

Introduction

Soft tissue and bone sarcomas represent the most common extracranial solid tumors in childhood[1, 2] and remain the third leading cause of cancer-related mortality among pediatric and adolescent patient[3]. The current World Health Organization (WHO) classification of pediatric tumors recognizes a diverse spectrum of heterogeneous sarcoma entities based on integrated histological, molecular, and phenotypic features[4, 5, 6]. Among these, Ewing sarcoma, osteosarcoma, alveolar rhabdomyosarcoma (aRMS), embryonal rhabdomyosarcoma (eRMS), and non-rhabdomyosarcoma soft tissue sarcomas (NRSTS) constitute the major diagnosis groups, all being associated with high risk for metastasis and relapse[7, 8, 9, 10, 11]. Across most sarcoma entities, standard-of-care treatment relies on multi-agent chemotherapy combined with local control through surgery and/or irradiation. However, these approaches frequently fail to achieve durable tumor eradication. In patients with relapsed or metastatic disease, effective therapeutic options are limited, and long-term survival rates have stagnated below 30% for several decades. Emerging precision oncology programs, including MoleculAr Profiling for Pediatric and Young Adult Cancer Treatment Stratification (MAPPYACTS)[12] and the INdividualized therapy FOr high-Risk childhood Malignancies (INFORM) registry[13, 14, 15], have encountered a lack of actionable targets for clinically available, molecularly-targeted drugs in pediatric high-risk sarcomas[14, 15].

Immunotherapies, such as chimeric antigen receptor (CAR) T cell therapies, have transformed the treatment landscape of hematological malignancies such as leukemia and lymphoma[16]. In contrast, their efficacy in solid tumors has been limited[17, 18]. This has been attributed to multiple mechanisms, including impaired immune cell infiltration into a dense extracellular matrix, the presence of an immunosuppressive tumor microenvironment (TME), metabolic reprogramming,

T cell dysfunction, and T cell exhaustion[19, 20, 21]. Moreover, pediatric sarcomas have traditionally been considered immunologically “cold” tumors, characterized by low tumor mutational burden, a paucity of neoantigens, and limited host antitumor immunity[22, 23, 24]. In general, pediatric tumors, and especially pediatric sarcomas, are characterized by spatiotemporal tumor heterogeneity and evolution induced by current therapies[25].

Recent advances in single-nuclei RNA-sequencing (snRNA-seq) enabled to challenge this paradigm of “cold” pediatric sarcomas. Studies in individual sarcoma entities demonstrate that these tumors are infiltrated by diverse non-malignant cell populations, suggesting the presence of complex and dynamic cellular ecosystems[26, 27, 28]. Malignant cell transformation may actively shape and reprogram the TME, resulting in the exclusion and functional suppression of host or therapeutic anti-tumor immune responses[29]. However, a comprehensive, cross-entity characterization of the pediatric sarcoma TME at high resolution is lacking.

In this study, we establish a single-nucleus transcriptomic atlas comprising over 800,000 nuclei from a broad spectrum of pediatric sarcomas including samples from heavily pretreated patients, the likely target population of immunotherapeutic trials. By integrating snRNA-seq with representative spatial characterization of the TME, we delineate the cellular composition, transcriptomic states, and cell-to-cell interaction networks that define these tumors. This fine-granular and comprehensive resource enables the identification of both shared and sarcoma-entity-specific features of the tumor-educated TME and provides a framework to identify targets and patient groups for next-generation cellular and antibody-based immunotherapies.

Results

Generation of a comprehensive single-nucleus atlas across high-risk pediatric sarcomas

To gain comprehensive insights into the sarcoma TME, we implemented 10x Genomics 5' snRNA-seq and MACSima-based spatial proteomic profiling in a pan-sarcoma pediatric cohort (**Fig. 1a**). Leveraging the INFORM[13, 14] and associated biobanks, we obtained *n*=89 fresh frozen tissue samples from *n*=62 patients. Longitudinal tissue samples were available for *n*=23/62 (37.10%) patients. Our cohort comprised *n*=23/89 (25.84%) Ewing sarcoma, *n*=13/89 (14.60%) osteosarcoma, *n*=19/89 (21.35%) alveolar rhabdomyosarcoma (aRMS), *n*=12/89 (13.48%) embryonal rhabdomyosarcoma (eRMS), and *n*=22/89 (24.72%) non-rhabdomyosarcoma soft tissue sarcomas (NRSTS). Patients span age groups from infants to young adults, anatomical locations across organ sites, and disease timepoints from primary diagnosis to multiple relapse episodes (**Fig. 1b**). Bulk tissue profiling via the clinical genomics platform within INFORM enabled WHO gold standard histo-molecular diagnostics and revealed expected entity-defining fusions and driver mutations, while also highlighting the heterogeneity of sarcomas, with few recurrent alterations and limited actionable targets. As INFORM enrolls high-risk and relapsed malignancies, most samples in our cohort (*n*=66/86, 76.74%) were collected after initiation of front-line or salvage therapy. Accordingly, the TME is shaped by both tumor-intrinsic and therapy-induced programs, representing the high-risk, relapsed patient population in need of novel immunotherapeutic strategies and suitable for phase I/II evaluation.

To resolve sarcoma biology at single-cell resolution, we generated *n*=807,414 high-quality single-nucleus transcriptomes passing our pre-defined quality thresholds, with a median of *n*=9,316 nuclei per sample after filtering (**Supplementary Fig. 1a-g**). After preprocessing and batch-aware integration, Uniform Manifold Approximation and Projection (UMAP) visualization revealed that nuclei retained expected sample- and entity-associated structure, while recurrent cancer entity-agnostic populations also aligned across samples and entities (**Fig. 1c-d**). To distinguish malignant sarcoma cancer cells from non-malignant TME populations, we implemented complementary annotations strategies (**Supplementary Fig. 2a-g**). First, we inferred single-nucleus copy-number variation (CNV) profiles. In the majority of tumors (*n*=72), whole-chromosomal or sub-chromosomal alterations were detected and validated against matched bulk tumor CNV profiles (**Supplementary Fig. 2d-g**). Second, we employed marker-based annotation informed by published bulk and single-cell transcriptomic studies for the respective entities. Together, these approaches enabled assignment of *n*=713,332/807,414 (88.35%) to the malignant and *n*=94,082 (11.65%) to the non-malignant compartment (**Fig. 1e**). The expression of established entity-associated markers confirmed that our integrative atlas reflects biological hallmarks defined in previous entity specific studies, such as *RBM11, LOXHD1, and MAPT* in Ewing sarcoma[30], *MYO18B*, *TTN, and SPATS2L* in aRMS[27, 31], *IGF2* in eRMS, *RUNX2 and COL11A1* in osteosarcoma[26, 32], and *CACNA2D2* in desmoplastic small round cell tumors (DSRCTs)[33] (**Fig. 1f, Supplementary Fig. 2c, Supplementary Table 3**). The abundance of the non-malignant compartment varied between sarcoma entities, ranging from a median of approximately 6% in Ewing sarcoma to 14% in osteosarcoma (**Fig. 1g**). Across individual samples, non-malignant cell type proportions ranged from a minimum 0% to a maximum 85%, highlighting marked diversity in TME infiltration across our cohort and motivating deeper

phenotyping of non-cancer populations.

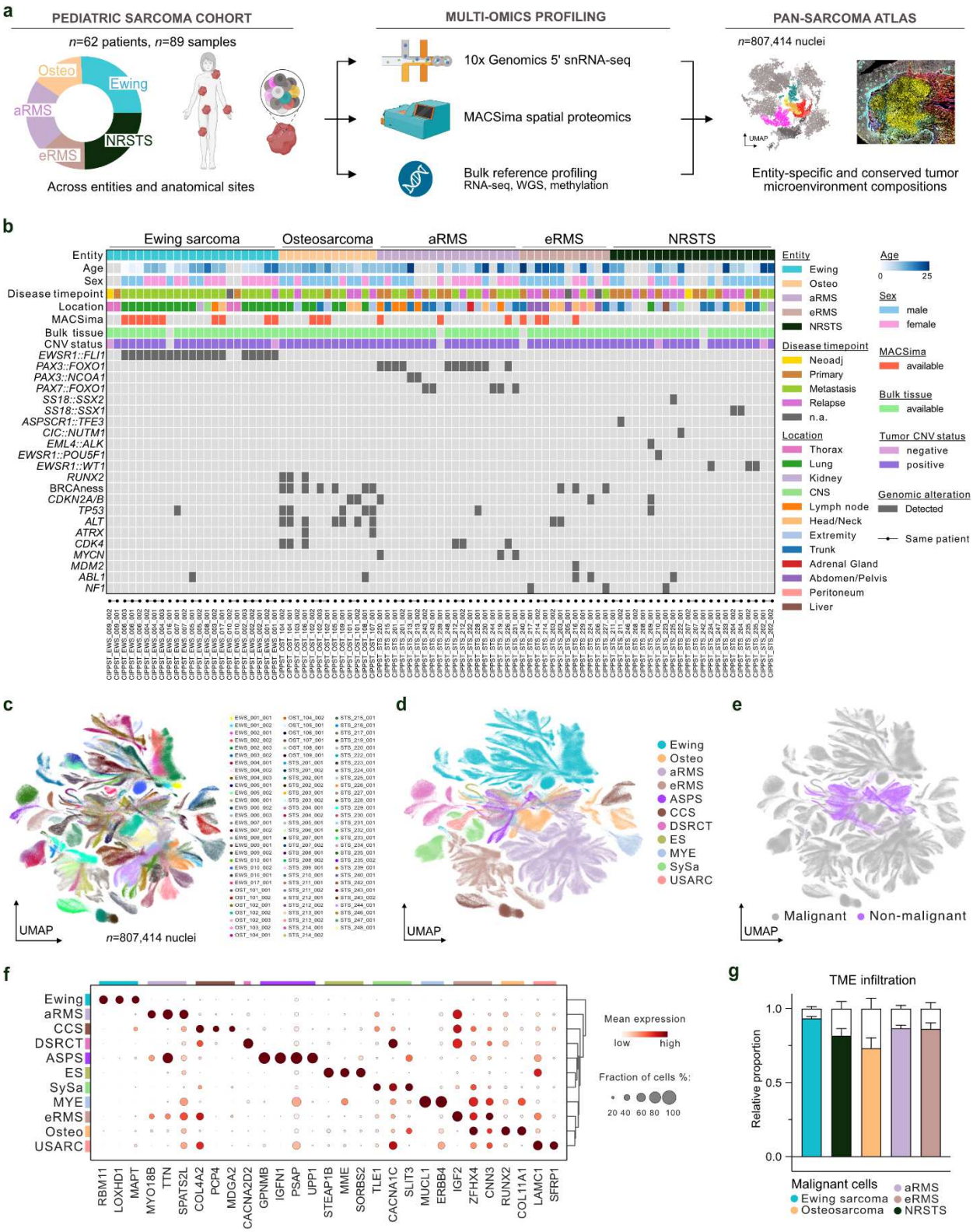


Figure 1: Generation of a comprehensive single-nucleus atlas across high-risk pediatric

sarcomas. (**a**) Overview of the study design to generate a single-nucleus RNA (snRNA) sequencing atlas of pediatric high-risk sarcomas. (**b**) Oncoplot summarizing clinical, diagnostic, and molecular annotation of the study cohort, grouped by sarcoma entity. Tumor tissue was characterized by the INFORM clinical genomics workflow, including DNA sequencing whole exome sequencing (WES) and shallow whole genome sequencing (WGS), RNA sequencing, and DNA methylation profiling. (**c**) Uniform Manifold Approximation and Projection (UMAP) embedding of all high-quality nuclei included in the integrated sarcoma atlas, colored by sample, *n*=807,414. (**d**) UMAP embedding colored by sarcoma entity, indicating sample-specific, entity-associated, and shared transcriptional structures. (**e**) UMAP embedding is colored by assignment of each nucleus to the malignant or non-malignant compartment. (**f**) Dot plot showing expression of representative entity-specific tumor cell markers. Dot color indicates mean gene expression and dot size indicates fraction of cells expressing each gene. (**g**) Proportion of nuclei assigned to the malignant and non-malignant compartment across major sarcoma entities, highlighting variations in TME composition.

Abbreviations: aRMS, alveolar rhabdomyosarcoma; ASPS, alveolar soft-part sarcoma; CCS, clear cell sarcoma; CNV, copy-number variation; CNS, central nervous system; DSRCT, desmoplastic small round cell tumor; eRMS, embryonal rhabdomyosarcoma; ES, epithelioid sarcoma; MYE, myoepithelial carcinoma; neoadj, neoadjuvant treatment; NRSTS, non-rhabdomyosarcoma soft-tissue sarcoma; Osteo, osteosarcoma; snRNA-seq, single-nucleus RNA sequencing; SySa, synovial sarcoma; TME, tumor microenvironment; UMAP, Uniform Manifold Approximation and Projection; USARC, undifferentiated sarcoma not otherwise specified; WGS, whole-genome sequencing.

Characterization of the sarcoma TME reveals heterogeneous immune and stromal infiltrates across subtypes

To characterize the sarcoma TME at high granularity, we extracted all *n*=94,082 non-malignant cells and re-clustered them using UMAP-based dimensionality reduction (**Fig. 2a**). Cell identities were assigned based on our multilayer annotation workflow, combining automated prediction by CellTypist[34], label transfer from published single-cell resources[35], gene set enrichment analysis (GSEA) of canonical pathways and cell signatures, and expert manual curation to generate the final consensus annotation (**Supplementary Fig. 3a-b**). This approach identified four major non-malignant compartments shared across all sarcoma entities: lymphocytes (*n*=17,531 nuclei, 2.17%), myeloid cells (*n*=26,720 nuclei, 3.31%), stromal cells (*n*=46,538 nuclei, 5.76%), and other tissue-resident cell types (*n*=3,293 nuclei, 0.41%) (**Fig. 2a, Supplementary Fig. 3c-d**). Subsequent high-resolution annotations of these four major compartments identified 19 distinct non-malignant cell types within the sarcoma TME (**Fig. 2b**).

Differential gene expression (DGE) analysis confirmed that the identified cell types exhibited hallmark transcriptional programs consistent with established markers (**Fig. 2c, Supplementary Fig. 3e**). The stromal compartment comprised fibroblasts, pericytes, and endothelial cells. Lymphoid populations included NK cells, B cells, plasma cells, and diverse T cell subsets, i.e. CD4+ T cells, CD8+ T cells, γδ T cells, mucosal-associated invariant T cells (MAIT), and proliferative T cells. The myeloid compartment was dominated by macrophages as well as less frequently dendritic cells (DCs), mast cells, and monocytes. In addition, we observed tissue-resident cell types consistent with the anatomical sampling sites represented in our cohort, such as osteoclasts, alveolar type II (ATII) cells, and ductal cells.

We next quantified the abundance of TME populations across major sarcoma entities and found substantial variation both between and within sarcoma entities (**Fig. 2d, Supplementary Table 4**). To quantify entity-specific differences in cell type composition, we applied scCODA[36] to estimate log2-fold changes in the abundance of each non-malignant population relative to the mean of the remaining sarcoma entities. We found that Ewing sarcoma was infiltrated by fibroblasts and CD4+ T cells, but less by macrophages. In osteosarcoma, macrophages were enriched together with endothelial cells and, as expected, osteoclasts, while fibroblasts were comparatively low in abundance. The TME of aRMS showed high proportions of fibroblasts and low proportions of endothelial cells and monocytes, whereas the TME of eRMS was low in fibroblasts and high in macrophages and dendritic cells. Across the heterogeneous group of NRSTS samples, no significant variations in cell types were observed except for low abundance of macrophages. Together, these data define diverse non-malignant TME landscapes in major pediatric sarcoma entities, adding a novel layer of sarcoma characterization beyond their malignant genotype.

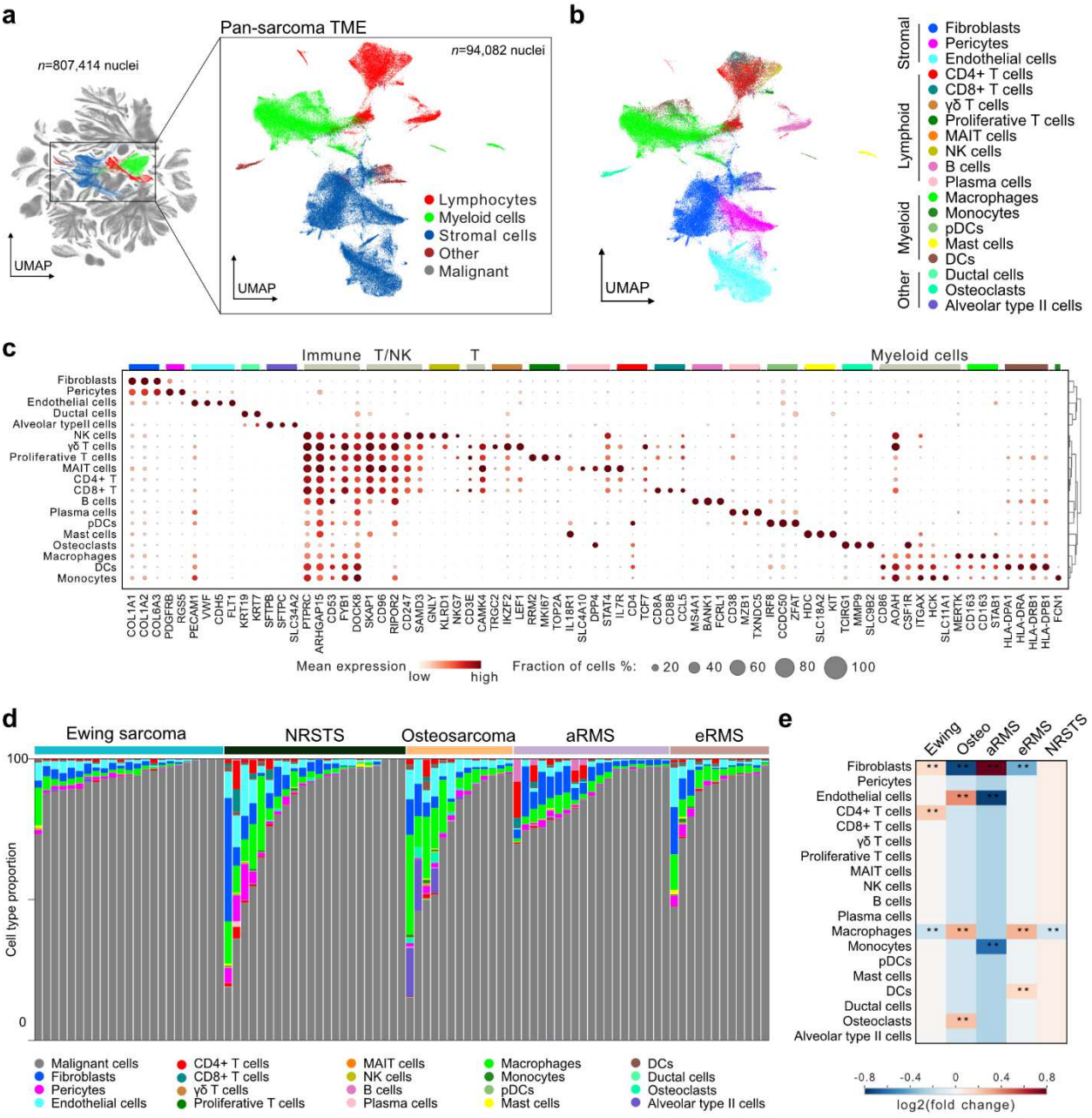


Figure 2: Annotation of the sarcoma TME defines stromal, lymphoid, and myeloid compartments. (**a**) UMAP embedding of the full sarcoma atlas highlighting the non-malignant compartment, which was re-projected for detailed annotation. The TME subset comprised *n*=94,082 nuclei and included stromal cells, lymphocytes, myeloid cells, and other non-malignant cell types. (**b**) UMAP embedding of the non-malignant TME compartment colored by detailed cell type annotation. Cell types identified include: fibroblasts, pericytes, endothelial cells (stromal compartment); NK cells, γδ T cells, proliferative T cells, mucosal-associated invariant T cells (MAIT), CD4+ and CD8+ T cells, B cells, plasma cells (lymphoid compartment); macrophages, monocytes, pDCs, mast cells, and DCs (myeloid compartment); ductal cells, osteoclasts, pulmonary ATII cells (other non-malignant cells). (**c**) Dot plot showing expression of canonical

marker genes used for cell type annotation. Dot color indicates mean gene expression and dot size indicates fraction of cells expressing each gene. (**d**) Stacked bar plots comparing proportions of malignant and non-malignant cell types across sarcoma entities. Samples are grouped by entity and sorted from diverse TME composition (left) to malignant cell dominant samples (right). (**e**) Differential enrichment of cell types across pediatric sarcoma diagnosis groups. Heatmap color shows the log2-fold change in cell-type abundance estimated by scCODA relative to the reference cell type (Pericytes) (red = enrichment, blue = depletion); this effect size is shown for all cell types. Cell types for which this change is statistically credible (scCODA, false discovery rate 0.2) are marked with **.

Abbreviations: ATII, alveolar type II; DC, dendritic cell; MAIT, mucosal-associated invariant T cells; NK, natural killer; pDC, plasmacytoid dendritic cell; TME, tumor microenvironment; UMAP, Uniform Manifold Approximation and Projection.

A conserved stromal framework with entity-specific CAF programs shapes the pediatric sarcoma microenvironment

Stromal cells are key components of the TME, shaping tumor architecture, vascularization, and cellular crosstalk and representing with 44.88% the largest cellular compartment within the pediatric sarcoma TME, accounting for and 5.76% of all cells (*n*=42,227; **Fig. 3a**). Based on canonical marker expression, stromal cells comprised pericytes (*PDGFRB*, 16.96%), fibroblasts (*COL1A2*, 48.24%), and endothelial cells (*VWF*, 34.79%) (**Fig. 3b**). Given the emerging role of CAFs in extracellular matrix (ECM) remodeling, angiogenesis, immune regulation, and tumor progression, we next investigated CAF heterogeneity (**Supplementary Fig. 4**).

Unsupervised clustering identified seven transcriptionally distinct CAF subpopulations (**Fig. 3c**). Vascular CAFs (vCAFs) shared features with pericytes, expressing *PDGFRB* together with *MYH11* and *ACTA2*. LIFR-CAFs were defined by high *LIFR* and DMD expression, suggesting roles in extracellular matrix organization and mechanosensory signaling. Antigen-presenting CAFs (apCAFs) expressed *CD74* and *PTPRC*, whereas tumor-like CAFs (tCAFs) showed moderate expression of canonical CAF markers together with elevated *CD99* and *RPS18*. Inflammatory CAFs (iCAFs) were characterized by *PLA2G2A*, *BICC1*, and *ABI3BP* expression, while matrix CAFs (mCAFs) exhibited a *POSTN*-high phenotype alongside broad expression of *COL1A1* and *COL3A1* (**Fig. 3d**).

Hallmark pathway enrichment analysis revealed distinct functional programs across CAF subsets (**Fig. 3e**). Pericytes were enriched for vascular homeostasis, angiogenesis, and oxidative phosphorylation, consistent with vessel stabilization. VCAFs shared vascular-associated programs but additionally displayed smooth muscle differentiation signatures. LIFR-CAFs preferentially enriched metabolic and cellular stress-response pathways, indicating a functionally distinct stromal state. ApCAFs showed enrichment of interferon signaling, inflammatory responses, and apoptosis, consistent with MHC class II-mediated immune interactions. ICAFs exhibited classical inflammatory programs, including IL6-JAK-STAT3, TNFα via NF-κB, IL2-STAT5, and interferon-γ signaling, supporting their role in cytokine secretion and immune modulation. In contrast, tCAFs were enriched for proliferative and metabolic pathways, including MYC targets, E2F targets, G2/M checkpoint, and oxidative phosphorylation, indicating a highly proliferative phenotype. Finally, mCAFs preferentially enriched extracellular matrix organization and mesenchymal programs, consistent with their role as the principal matrix-producing fibroblast population.

To assess the biological relevance of these stromal populations, we next examined their abundance across pediatric sarcoma entities (**Fig. 3f-h**). Endothelial cells represented the largest stromal population overall, followed by mCAFs, tCAFs, and pericytes, whereas vCAFs and LIFR-CAFs constituted comparatively small fractions (**Fig. 3f**). Although all stromal populations were detected across tumor entities, their relative abundance varied considerably (**Fig. 3g**). Osteosarcoma contained the highest proportion of endothelial cells, whereas aRMS was enriched for tCAFs. eRMS contained relatively more apCAFs, while Ewing sarcoma and NRSTS displayed a comparatively balanced stromal composition. Because aRMS and NRSTS contributed the largest stromal datasets (*n*=13,094 and *n*=15,658, respectively), these sample sizes should be

considered when interpreting relative abundances.

Differential abundance analysis further revealed tumor-specific stromal enrichment patterns (**Fig. 3h**). TCAFs were significantly enriched in aRMS and Ewing sarcoma but depleted in osteosarcoma and NRSTS, whereas endothelial cells exhibited the opposite distribution. Interestingly, the two rhabdomyosarcoma subtypes showed opposing mCAF abundance, with depletion in eRMS and enrichment in aRMS. ApCAFs were specifically enriched in eRMS, LIFR-CAFs showed modest enrichment in Ewing sarcoma, and iCAFs were depleted in osteosarcoma. Collectively, these findings demonstrate that while a common stromal framework exists across pediatric sarcomas, each entity exhibits a distinct stromal composition characterized by differential enrichment of specific CAF subsets and stromal populations.

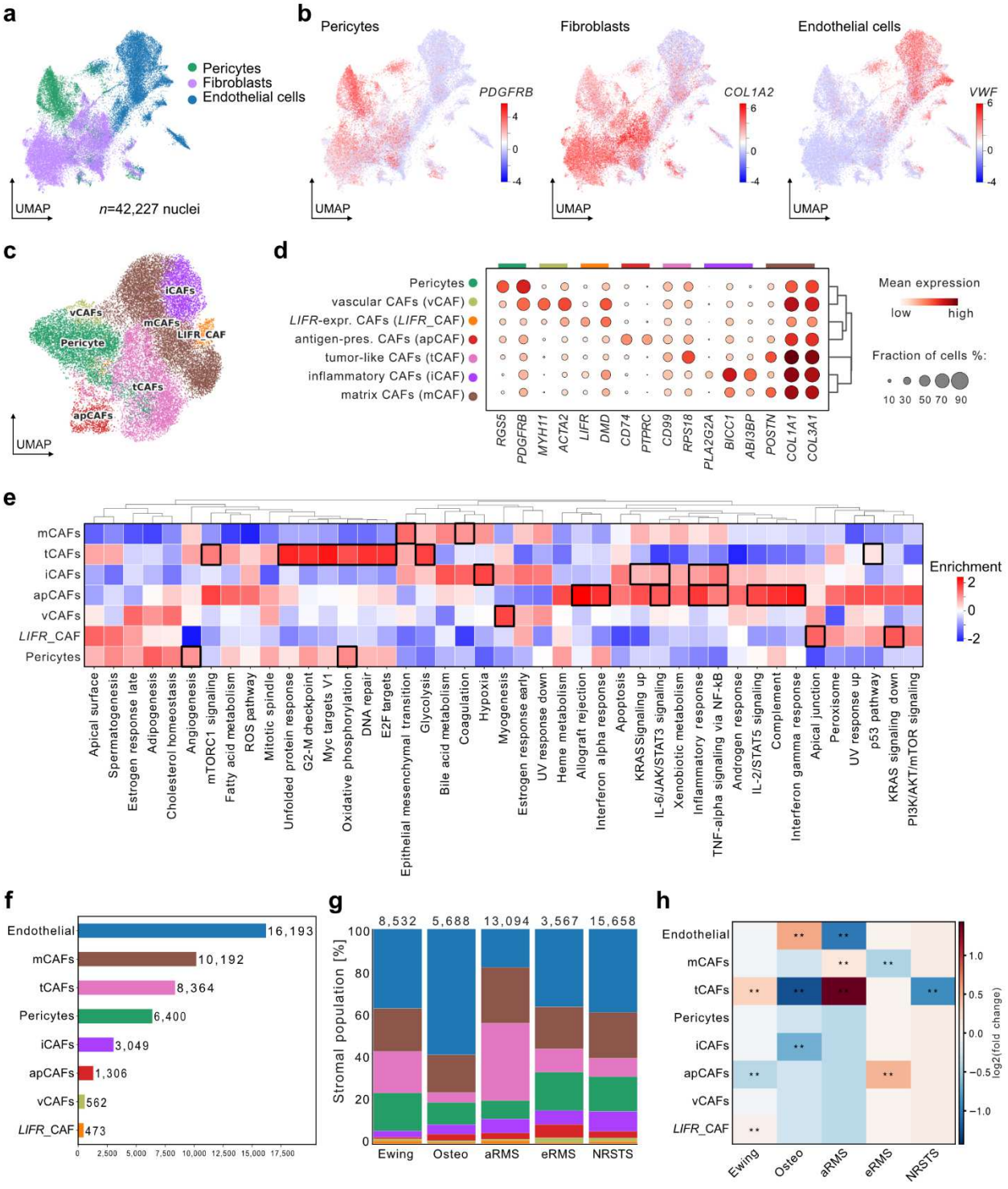


Figure 3: Stromal cell landscape and CAF functional heterogeneity in the sarcoma TME

(**a**) UMAP visualization of *n*=42,227 stromal nuclei identified by snRNA sequencing, showing the three major stromal cell populations: pericytes, fibroblasts, and endothelial cells. (**b**) UMAP feature plots showing the expression of canonical marker genes used to identify the major stromal populations, including *PDGFRB* (pericytes), *COL1A2* (fibroblasts), and *VWF* (endothelial cells). Color intensity represents normalized gene expression; (low, blue; high, red) (**c**) UMAP

visualization of fibroblast subclustering, identifying seven transcriptionally distinct stromal populations comprising pericytes, vascular CAFs (vCAFs), LIFR-expressing CAFs (LIFR-CAFs), antigen-presenting CAF (apCAFs), tumor-like CAFs (tCAFs), inflammatory CAFs (iCAFs), and matrix CAFs (mCAFs). (**d**) Dot plot showing the expression of representative marker genes across the identified stromal cell populations. Dot color indicates mean normalized gene expression, whereas dot size represents the percentage of cells expressing each gene. (**e**) Heatmap showing GSEA for the identified CAF subpopulations and pericytes. Red indicates positive enrichment and blue indicates relative depletion. Black boxes highlight key functional pathways used to characterize the respective CAF states. (**f**) Absolute numbers of stromal cell populations in decreasing order: Endothelial cells represented the largest stromal population, followed by mCAFs, tCAFs, and pericytes, whereas vCAFs and LIFR-CAFs were comparatively rare. (**g**) Relative abundance of stromal cell populations across pediatric sarcoma diagnosis, including Ewing, Osteo, aRMS, eRMS, and NRSTS. Numbers above the bars indicate the total number of stromal cells analyzed for each tumor diagnosis. (**h**) Differential enrichment of stromal cell populations across pediatric sarcoma diagnosis groups. Heatmap depicts the log2-fold change in cell-type abundance estimated by scCODA, with red indicating enrichment and blue indicating depletion. scCODA uses a Bayesian model to estimate the posterior probability of a credible effect for each cell-type-covariate combination while controlling the FDR at 0.2. Cell types with a credible effect under this threshold are marked with **. **Abbreviations**: CAF, cancer-associated fibroblasts; TME, tumor microenvironment; UMAP, Uniform Manifold Approximation and Projection. snRNA, single nuclei RNA; vCAFs,vascular CAFs; LIFR-CAFs, LIFR-expressing CAFs, apCAFs, antigen-presenting CAFs; tCAFs, tumor-like CAFs; iCAFs, inflammatory CAFs; mCAFs, matrix CAFs, Ewing, Ewing sarcoma; Osteo, Osteosarcoma; aRMS, alveolar rhabdomyosarcoma; eRMS, embryonal rhabdomyosarcoma; NRSTS, non-rhabdomyosarcoma soft tissue sarcomas.

Activated, effector, and exhausted lymphoid states dominate the immune microenvironment of high-risk pediatric sarcomas

Lymphocytes are highly specialized immune effector cells, orchestrating adaptive immune responses within the TME by balancing cytotoxic anti-tumor activity with immunoregulatory and immunosuppressive function. To characterize the lymphoid compartment, we extracted *n*=9,771 lymphoid cells, comprising 10.39% of the TME, and performed unsupervised subclustering, identifying eight major populations comprising B cells, NK cells, plasma cells, and six T cell lineages, including CD4+, CD8+, MAIT, γδ T, and proliferating T cells (**Fig. 4a**). Broad lineage assignment was based on canonical marker expression and gene signature scoring, including *CD3E* for T cells, *MS4A1* for B cells, and *NCAM1* for NK cells (**Fig. 4b**). B cells were further classified into memory or naïve B cells and plasma cells. NK cells were subdivided according to *CD16* expression, distinguishing functionally distinct NK cell subsets (**Supplementary Fig. 5a-b**).

To resolve functional heterogeneity, T cell clusters were annotated using marker genes and reference signatures from the pan-cancer T cell atlas by Chu et al. Interestingly, unsupervised clustering primarily reflected functional state rather than CD4/CD8 lineage, with activation, exhaustion, and effector differentiation driving most transcriptional variation. Final annotations integrated lineage prediction, canonical marker expression, reference-based label transfer, and unsupervised clustering.

Among the *n*=4,956 CD4+ T cells, we resolved ten transcriptionally distinct cell states, including regulatory T cells (Tregs), follicular helper T cells (Tfh), central memory T cells (Tcm), naïve T cells (Tn), TCF7+ (Tn_TCF7+_CD4) and TCF7L2+ naïve-like populations (Tn_TCF7L2_CD4), interferon-responsive T cells (TISG), cytotoxic/effector T cells (TEFF), exhausted T cells (TEX), and early activated T cells (early_activ_T) (**Fig. 4c**). Likewise, the 2,434 CD8+ T cells segregated into eight states comprising TCF7+ and TCF7L2+ naïve-like populations, TISG, TEX, TEFF, transitional effector (t-TEFF), Tn, and early_activ_T_CD8 (**Fig. 4d**). Cell states were assigned through combined DGE, (gene ontology) GO enrichment, GSEA, and label transfer from the reference atlas, followed by manual curation (**Fig. 4e-f** and **Supplementary Fig. 5c-d**). Across all sarcoma samples, the lymphoid compartment was dominated by TEFF_CD4 T cells (*n*=1,217), early_activ_T_CD4 cells (*n*=1,159), TEFF_CD8 T cells (*n*=892), Tcm_CD4 T cells (*n*=866), and TEX_CD8 T cells (*n*=716). Beyond canonical TCRab T cell subsets, infiltrates of γδ T (n=738), MAIT (*n*=304) as well as NK cells (*n*=504) were detected across all sarcoma entities. In contrast, Tn_TCF7+_CD8 T cells (*n*=92), t-TEFF_CD8 T cells (*n*=87), Tn_TCF7L2_CD8 T cells (*n*=86), TEX_CD4 T cells (*n*=56), ISG_CD8 T cells (*n*=41), and Tn_CD4 cells (*n*=42) were comparatively rare across sarcoma entities. Infiltration of suppressive Tregs (*n*=579) was detected across sarcoma entities (**Fig. 4g**). Overall, the lymphoid compartment of high-risk pediatric sarcoma is predominantly composed of activated and effector T cell states, accompanied by a substantial population of TEX_CD8 T cells and Tregs. We next assessed entity-specific differences in relative and absolute abundance of lymphoid cell populations across sarcoma entities (**Fig. 4h-i**). In line with stromal cell quantities, NRSTS samples exhibited the highest overall abundance of lymphoid cells relative to the other sarcoma types, while depleted in early_activ_T_CD4 cells. In contrast, Ewing sarcoma was characterized by the lowest overall abundance, with a predominance of Tn_CD8 T cells, early_activ_T_CD8 cells, and a lower abundance of MAIT cells. Overall, B

(*n*=198) and plasma cells (*n*=366) were sparsely detected and particularly low in Ewing sarcoma. Activated CD16+ NK cells and MAIT cells were most abundant in osteosarcoma. TEX_CD8 were almost exclusively found in NRSTS while TISG_CD4 and t-TEFF_CD8 T cells were only detected in Ewing and eRMS. Aggregation of T cell signature scores by entity revealed broadly conserved functional programs across tumor types, indicating that T cell state was largely maintained irrespective of histology (**Supplementary Fig. 5f**). Together, these findings demonstrate that while core lymphoid populations are shared across pediatric sarcomas, each entity exhibits a distinct immune composition defined by differential enrichment of activated, effector, and exhausted T cell states.

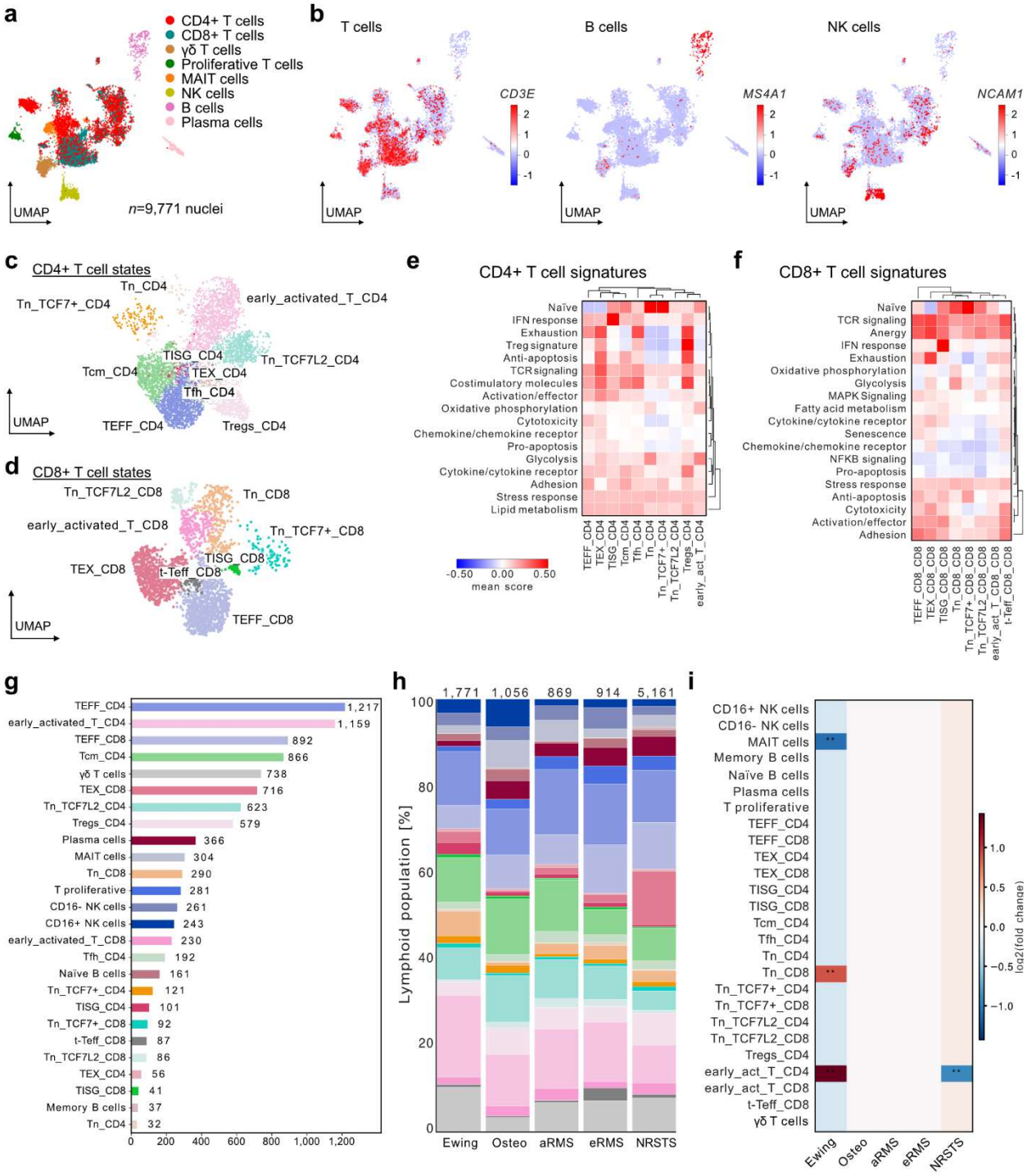


Figure 4: Single-cell characterization of the lymphoid compartment across high-risk sarcoma entities.

(**a**) UMAP visualization of all lymphoid and NK cells (*n*=9,771 nuclei), colored by the eight main cell type annotations: B cells, CD4+ T cells, CD8+ T cells, MAIT cells, NK cells, plasma cells, T proliferative cells, and γδ T cells. (**b**) UMAP feature plots showing the expression of canonical

lineage markers: *CD3E* (T cells), *MS4A1* (B cells), and *NCAM1* (NK cells). Color scale represents normalized gene expression (low, blue; high, red). (**c**) UMAP embedding of CD4+ T cells colored by cell states: regulatory T cells (Tregs_CD4), naïve T cells (Tn_CD4), TCF7+ naïve-like T cells (Tn_TCF7+_CD4), interferon-responsive T cells (TISG_CD4), cytotoxic/effector T cells (TEFF_CD4), follicular helper T cells (Tfh_CD4), central memory T cells (Tcm_CD4), TCF7L2+ naïve-like T cells (Tn_TCF7L2_CD4), and early activated CD4+ T cells (early_activ_T_CD4). (**d**) UMAP embedding of CD8+ T cells colored by cell states: TCF7+ naïve-like T cells (Tn_TCF7+_CD8), interferon-responsive T cells (TISG_CD8), exhausted T cells (TEX_CD8), cytotoxic/effector T cells (TEFF_CD8), transitional effector T cells (t-TEFF_CD8), naïve CD8+ T cells (Tn_CD8), TCF7L2+ naïve-like T cells (Tn_TCF7L2_CD8), and early activated CD8+ T cells (early_activ_T_CD8). (**e, f**) Heatmap of mean functional signature scores across CD4+ or CD8+ T cell states, respectively, with hierarchical clustering of both gene signatures and cell states. (**g**) Absolute numbers of global T cell populations in decreasing order. (**h**) Relative abundance of lymphoid cell populations across sarcoma diagnosis groups, including Ewing, Osteo, aRMS, eRMS, and NRSTS. Numbers above the bars indicate total lymphoid cell counts per entity (Ewing sarcoma, *n*=1,771; osteosarcoma, *n*=108; aRMS, *n*=869; eRMS, *n*=914; NRSTS, *n*=5,161). (**i**) Differential enrichment of lymphoid cell populations across pediatric sarcoma diagnosis groups. Heatmap depicts the log2-fold change in cell-type abundance estimated by scCODA relative to the reference cell type (CD16+ NK cells), with red indicating enrichment and blue indicating depletion. scCODA uses a Bayesian model to estimate the posterior probability of a credible effect for each cell-type-covariate combination while controlling the FDR at 0.2. Cell types with a credible effect under this threshold are marked with **.

Abbreviations: Tcm, central memory T cells; Tfh, follicular helper T cells, Tn, naïve T cells, Tregs, regulatory T cells; Tn_TCF7+, naïve-like TCF7+ T cells; Tn_TCF7L2, naïve-like TCF7L2+ T cells; TISG, interferon-responsive T cells; TEFF, cytotoxic/effector T cells, TEX, exhausted T cells; early_activ_T, early activated T cells; t-TEFF, transitional effector T cells, IFN, interferon; TCR, T cell receptor.

Macrophage heterogeneity shapes the myeloid microenvironment of high-risk pediatric sarcomas

The myeloid compartment comprises diverse innate immune cell populations that orchestrate inflammatory responses and regulate antitumor immunity within the TME. To characterize this compartment, we extracted *n*=24,790 cells, comprising 26.35% of the entire TME, and performed unsupervised clustering, identifying four major populations comprising macrophages, DCs, mast cells, and monocytes based on canonical marker expression (**Fig. 5a-b**). Macrophages were defined by expression of *CD163*, whereas mast cells and monocytes expressed *KIT* and *FCN1*, respectively (**Fig. 5b**).

The myeloid compartment was annotated using DEG and pathway enrichment analysis, identifying conventional dendritic cells (cDC1s and cDC2s), plasmacytoid dendritic cells (pDCs), monocyte-derived dendritic cells (moDCs), mature regulatory dendritic cells (mregDCs), and three monocyte subsets: classical, inflammatory, and *FCGR3A*+ monocytes (**Fig. 5c**). As macrophages constituted the predominant myeloid population and are key regulators of inflammation, tissue remodeling, and antitumor immunity within the TME, we next investigated their transcriptional heterogeneity. Macrophages were re-clustered and annotated using an integrative strategy combining canonical macrophage state markers reported in previous studies[37], DGE, GO enrichment, and pathway signature analyses. Rather than assigning macrophages according to the classical M1/M2 polarization paradigm, which insufficiently captures the complexity of tissue macrophages, annotation was guided by transcriptional cell states reflecting distinct functional programs (**Fig. 5c-e, Supplementary Fig. 6**). We identified six transcriptionally distinct macrophage states, comprising ECM-remodeling macrophages (Mac_ECM), inflammatory macrophages (Mac_Inf), lipid-metabolic macrophages (Mac_Lipid_Metabolic), proliferating macrophages (Mac_Prolif), scavenger/alveolar-like macrophages (Mac_Scavenger_Alveolar), and suppressive macrophages (Mac_Supp). Functional annotation revealed clear specialization of these populations. These states showed distinct functional programs, including ECM organization (Mac_ECM), inflammatory signaling (Mac_Inf), metabolic and lysosomal pathways (Mac_Lipid_Metabolic), cell-cycle activity (Mac_Prolif), and immunoregulatory pathways (Mac_Supp) (**Fig. 5e**). (**Fig. 5e**). Across all tumors, Mac_Supp represented the largest macrophage population (*n*=7,015), followed by Mac_Inf (*n*=6,486), Mac_Lipid_Metabolic (*n*=2,430), Mac_ECM (*n*=2,367), Mac_Prolif (*n*=2,309), and Mac_Scavenger_Alveolar (*n*=1,015) (**Fig. 5f**). Among non-macrophage myeloid cells, the analyzed tissues were dominated by moDCs (*n*=1,061), whereas mast cells, cDCs, pDCs, mregDCs, and monocytes were comparatively less frequent.

Comparison across sarcoma entities revealed marked differences in myeloid cell composition (**Fig. 5g-h**). To quantify entity-specific differences, we calculated the scCODA log2-fold changes for each myeloid population compared to the mean of the other entities. While Mac_Supp and Mac_Inf constituted the predominant myeloid populations across all entities, Ewing sarcoma contained a comparatively larger proportion of Mac_Prolif and less Mac_ECM. Both, aRMS and eRMS exhibited a strong enrichment of Mac_ECM, whereas aRMS comprised also more Mac_Lipid_Metabolic and Mac_Prolif. In contrast, NRSTS demonstrated less Mac_Prolifs. Together, these findings demonstrate that macrophage states are conserved across pediatric sarcomas but display entity-specific functional states, reflecting distinct programs of immune

regulation and tissue remodeling.

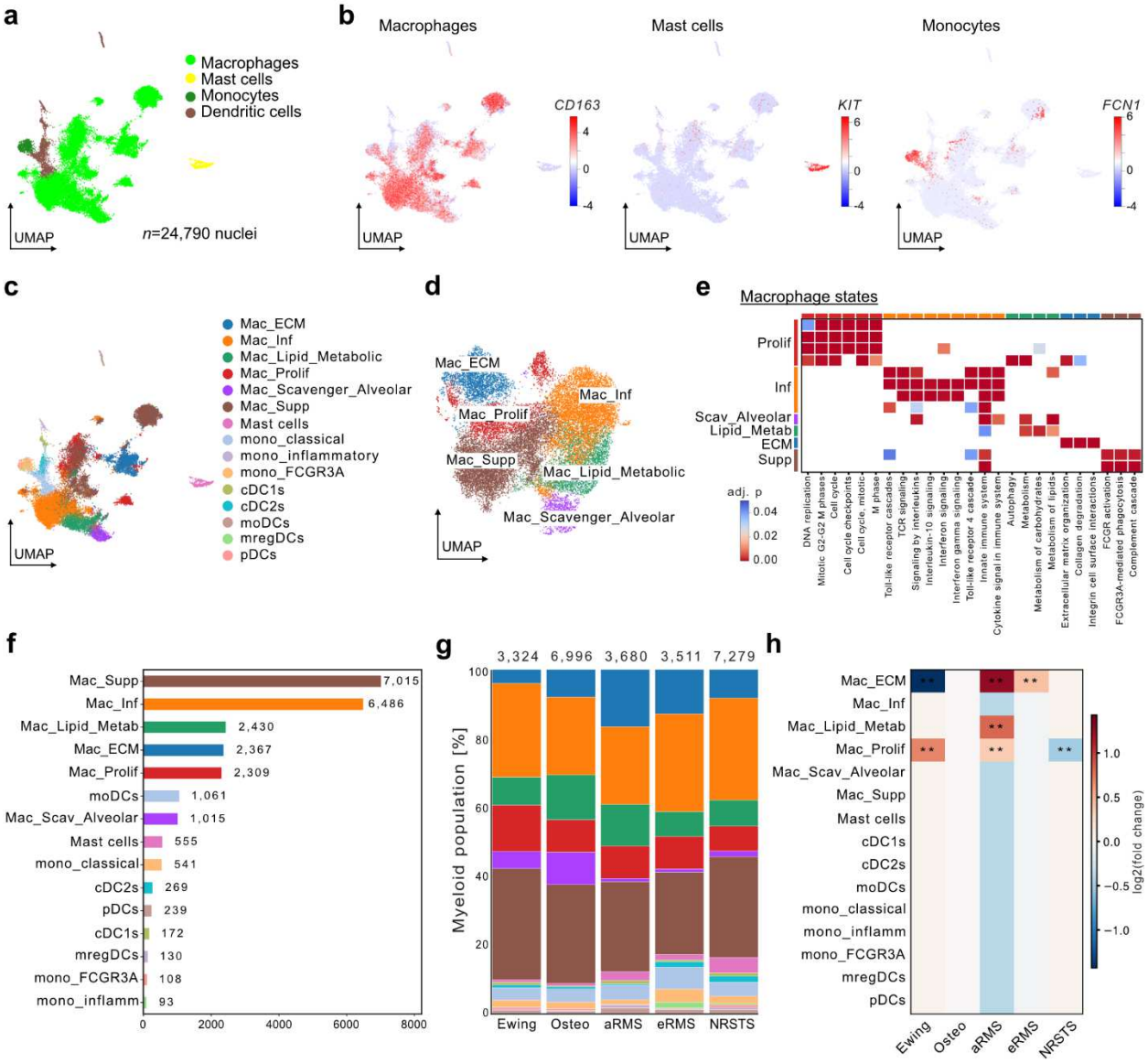


Figure 5: Single-cell characterization of the myeloid compartment reveals distinct macrophage states across pediatric sarcomas.

(**a**) UMAP visualization of all myeloid cells (*n*=24,790) showing major cell populations after subclustering. (**b**) Expression of representative lineage markers used for annotation, including *CD163* (macrophages), *KIT* (mast cells), and *FCN1* (monocytes). (**c**) UMAP embedding showing annotated myeloid cell populations, including mast cell, macrophages, DCs (conventional dendritic cells (cDC1s and cDC2s), mature regulatory dendritic cells (mregDCs), plasmacytoid DCs (pDCs)), and monocyte (classical, inflammatory, and *FCGR3A*+ monocytes) subsets. (**d**) Reclustering of macrophages identifies six distinct macrophage states: extracellular matrix

(ECM)-remodeling macrophages (Mac_ECM), inflammatory macrophages (Mac_Inf), lipid-metabolic macrophages (Mac_Lipid_Metabolic), proliferating macrophages (Mac_Prolif), scavenger/alveolar-like macrophages (Mac_Scavenger_Alveolar), and suppressive macrophages (Mac_Supp). (**e**) Heatmap summarizing significantly enriched functional pathways for each macrophage state based on GO and pathway enrichment analyses. (**f**) Bar plot of myeloid cell subtype distribution (cell counts) across the full dataset, ranked from most to least abundant. (**g**) Stacked bar plot showing the relative proportion (%) of each myeloid subtype within each sarcoma entity (Ewing sarcoma, osteosarcoma, aRMS, eRMS, and NRSTS). Numbers above the bars indicate the total number of myeloid cells per entity (Ewing,*n*=3,324; osteosarcoma, *n*=6,996; aRMS, *n*=3,680; eRMS, *n*=3,511; NRSTS, *n*=7,279). (**h**) Differential enrichment of myeloid cell populations across pediatric sarcoma diagnosis groups. Heatmap depicts the log2-fold change in cell-type abundance estimated by scCODA relative to the reference cell type (Mac_Supp), with red indicating enrichment and blue indicating depletion. scCODA uses a Bayesian model to estimate the posterior probability of a credible effect for each cell-type-covariate combination while controlling the FDR at 0.2. Cell types with a credible effect under this threshold are marked with **.

Abbreviations: cDCs, conventional dendritic cells; pDCs, plasmacytoid dendritic cells; moDCs, monocyte-derived dendritic cells; mregDCs, mature regulatory dendritic cells; Mac_ECM, ECM-remodeling macrophages, Mac_Inf, inflammatory macrophages, Mac_Lipid_Metabolic, lipid-metabolic macrophages; Mac_Prolif, proliferating macrophages; Mac_Scavenger_Alveolar, scavenger/alveolar-like macrophages; Mac_Supp, suppressive macrophages, FCGR, FC gamma receptor.

Convergent TME archetypes transcend pediatric sarcoma entities

Because immunotherapy responses may be driven more by the TME than by histology, shared TME states could reveal pan-sarcoma vulnerabilities. We therefore leveraged the cross-entity structure of our snRNA atlas to investigate whether high-risk pediatric sarcomas converge on common TME compositions. Hierarchical clustering of proportional cell type abundance per sample identified twelve discrete TME archetypes, including fibroblast-rich (C1, C2), macrophage-rich (C3), immune-rich (C5), T cell-rich (C10), vascularized (C7, C8, C9, C12), and balanced mixed-TME (C10) composition clusters (**Fig. 6a-b**). Cluster C4 was strongly enriched for alveolar ATII cells, reflecting the lung metastatic microenvironment of osteosarcoma, whereas cluster C6 was dominated by lymphocytes, consistent with lymph node metastases from Ewing sarcoma and aRMS (**Fig. 6a**). All remaining archetypes comprised samples from multiple entities, disease stages, and anatomical sites, indicating conserved TME architectures. Together, these findings indicate that TME archetypes represent conserved and biologically meaningful microenvironmental states.

To explore clinical relevance of TME archetypes, downstream analyses were restricted to archetypes with >3 clinically annotated samples: C2 (fibroblast-rich), C5 (immune-rich), C6 (lymph node), C9 (vasculature_3), C10 (T cell-rich), and C11 (balanced). As several patients contributed serial samples that transitioned between TME archetypes, survival analyses were performed at the tumor episode level, with time measured from tumor sampling (INFORM registry enrollment) to disease progression or death. These transitions highlight the dynamic nature of TME states, suggesting that they represent adaptive microenvironmental features rather than fixed properties of tumor origin (**Supplementary Fig. 7e**). Despite the limited power of this high-risk and relapsed/refractory cohort, the largest archetypes, C2 (*n*=19 tumor episodes), C5 (*n*=9), C11 (*n*=7), and C9 (*n*=6), the balanced archetype C11 showed a trend toward prolonged event-free survival (EFS) (median 23 months) compared with the fibroblast-rich C2 (15 months) and the immune-rich C5 and vasculature-rich C9 archetypes (10 months each) (**Fig. 6c, Supplementary Fig. 7d**). Long-term EFS also appeared more favorable in C11 and C5, whereas patients assigned to the fibroblast-rich archetype C2 tended to experience inferior outcomes. Although these observations suggest that TME archetypes may contribute to clinical outcome, larger, prospectively annotated cohorts will be required to validate these associations.

To further resolve the biological basis of the composition-based archetypes, we assessed the distribution of functional cell states across the major TME archetypes (**Fig. 6d-f**, for all archetypes: **Supplementary Fig. 7a-c**). Compositional analysis revealed that archetypes differ not only in cellular composition but also in coordinated functional programs conserved across sarcoma entities. In C2, stroma was dominated by matrix-producing mCAFs and tCAFs, with few endothelial cells and pericytes. Consistent with this desmoplastic microenvironment, Mac_ECM and Mac_Lipid_Metabolic macrophages were significantly enriched in C2. In contrast, C5 exhibited reduced stromal content and enrichment for endothelial cells, pericytes, and immune-engaging apCAFs, suggesting a more permissive immune niche. As expected for lymph node metastases, C6 was enriched for naïve B- and T cells, consistent with a physiological lymphoid microenvironment rather than a tumor-educated TME. C9 was dominated by endothelial cells and pericytes with sparse infiltration of CAFs, lymphocytes, and myeloid cells, and lacked enrichment for specific functional cell states. In contrast, C10 was significantly enriched for exhausted CD8+

T cells, effector CD4+ T cells, Tregs, and γδ T cells, together with depletion of Mac_ECM macrophages, consistent with an immunologically inflamed ("hot") microenvironment. Finally, the balanced archetype C11 was enriched for iCAFs and tCAFs together with a relative depletion of Mac_ECM macrophages. Collectively, these findings demonstrate that conserved TME archetypes are defined by coordinated cellular ecosystems rather than simple differences in cell-type abundance, providing a biologically informed framework for stratifying pediatric sarcomas according to microenvironmental states that may prove more relevant for immunotherapeutic intervention than histological classification alone.

Collectively, these findings show that TME archetypes reflect coordinated cellular ecosystems rather than isolated shifts in cell-type abundance, providing a framework for stratifying pediatric sarcomas by microenvironmental state, a classification that may prove more relevant to immunotherapy than histology alone.

To validate the identified TME archetypes at the spatial level, we analyzed MACSima imaging cyclic staining (MICS) data from representative examples of fibroblast-rich (C2), immune-rich (C5), and T cell-rich (C10) archetypes (**Fig. 6g**). Consistent with the snRNA-seq data, C2 exhibited dense and homogeneous fibroblast infiltration accompanied by extensive ECM deposition, including collagen III, collagen IV, and fibronectin, supporting the enrichment of mCAFs and Mac_ECM macrophages. T cells and macrophages were distributed uniformly throughout the tissue. In contrast, C5 displayed reduced ECM deposition, with spatially segregated immune infiltrates: T cell clusters localized predominantly within tumor cell-high regions, whereas macrophages accumulated in the peritumoral stroma. The T cell-rich archetype C10 exhibited a highly organized architecture characterized by prominent vasculature, perivascular macrophage accumulation, and abundant T cell infiltration. MICS recapitulated snRNA-seq-defined cell compositions and additionally captured populations, including granulocytes, that are known to be underrepresented in snRNA data (Supplementary Fig. 8). (**Supplementary Fig. 8**). Together, these complementary approaches integrate transcriptional identity with spatial organization, providing a framework for dissecting cross-entity TME archetypes and the tumor-microenvironment interactions that shape them.

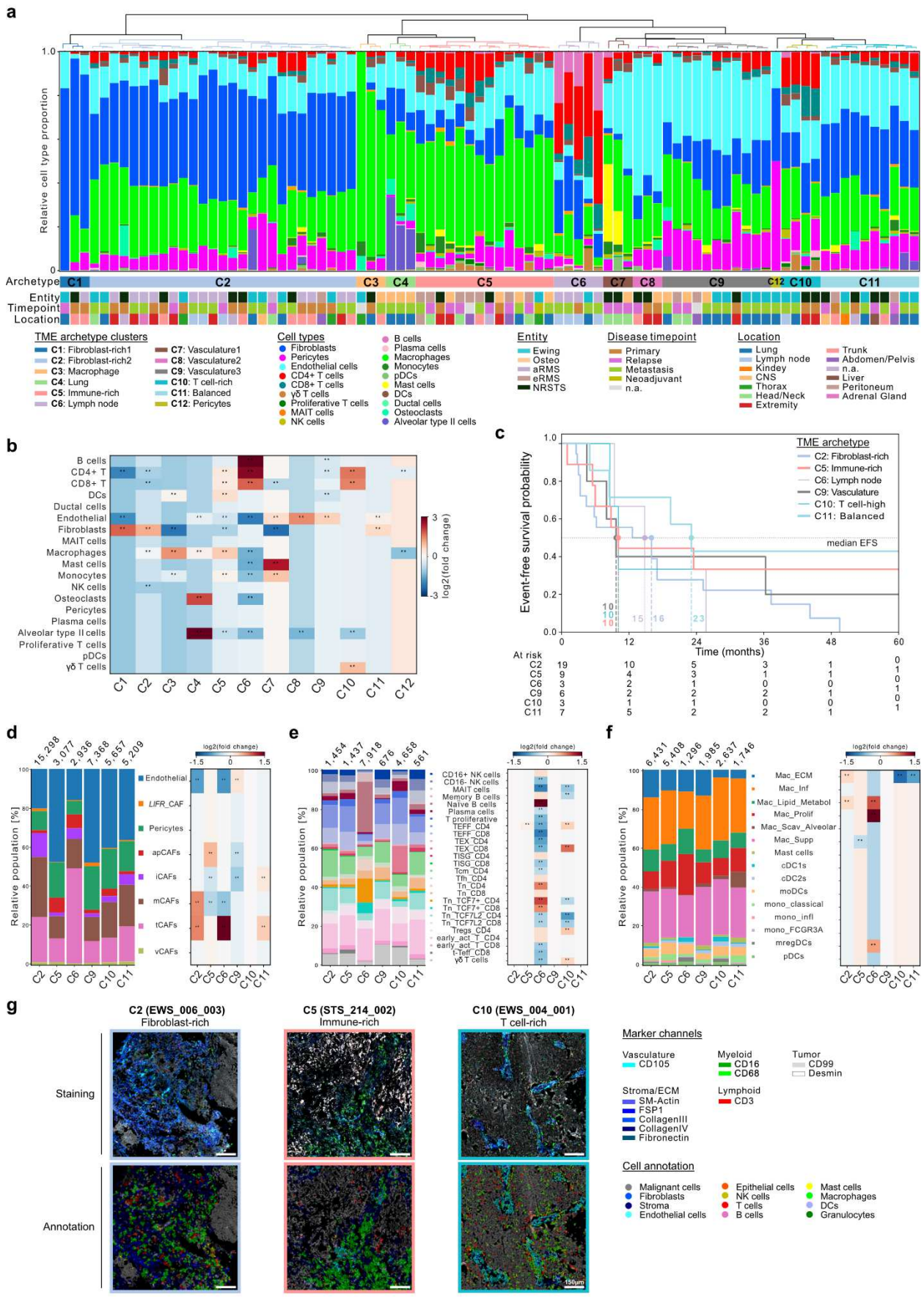

a
Relative cell type proportion
Archetype
Entity
Timepoint
Location
TME archetype clusters
C1: Fibroblast-rich1
C2: Fibroblast-rich2
C3: Macrophage
C4: Lung
C5: Immune-rich
C6: Lymph node
C7: Vasculature1
C8: Vasculature2
C9: Vasculature3
C10: T cell-rich
C11: Balanced
C12: Pericytes
Cell types
Fibroblasts
Pericytes
Endothelial cells
CD4+ T cells
CD8+ T cells
γδ T cells
Proliferative T cells
MAIT cells
NK cells
B cells
Plasma cells
Macrophages
Monocytes
pDCs
Mast cells
DCs
Ductal cells
Osteoclasts
Alveolar type II cells
Entity
Ewing
Osteo
aRMS
eRMS
NRSTS
Disease timepoint
Primary
Relapse
Metastasis
Neoadjuvant
n.a.
Location
Lung
Lymph node
Kindey
CNS
Thorax
Head/Neck
Extremity
Trunk
Abdomen/Pelvis
n.a.
Liver
Peritoneum
Adrenal Gland
b
log2(fold change)
c
Event-free survival probability
Time (months)
TME archetype
C2: Fibroblast-rich
C5: Immune-rich
C6: Lymph node
C9: Vasculature
C10: T cell-high
C11: Balanced
median EFS
At risk
d
Relative population [%]
Endothelial
LIFR_CAF
Pericytes
apCAFs
iCAFs
mCAFs
tCAFs
vCAFs
e
Relative population [%]
f
Relative population [%]
Mac_ECM
Mac_Inf
Mac_Lipid_Metabol
Mac_Prolif
Mac_Scav_Alveolar
Mac_Supp
Mast cells
cDC1s
cDC2s
moDCs
mono_classical
mono_infl
mono_FCGR3A
mregDCs
pDCs
g
C2 (EWS_006_003)
Fibroblast-rich
C5 (STS_214_002)
Immune-rich
C10 (EWS_004_001)
T cell-rich
Staining
Annotation
Marker channels
Vasculature
CD105
Myeloid
CD16
CD68
Tumor
CD99
Desmin
Stroma/ECM
SM-Actin
FSP1
CollagenIII
CollagenIV
Fibronectin
Lymphoid
CD3
Cell annotation
Malignant cells
Fibroblasts
Stroma
Endothelial cells
Epithelial cells
NK cells
T cells
B cells
Mast cells
Macrophages
DCs
Granulocytes

Figure 6: Convergent TME archetypes transcend pediatric sarcoma entities. (**a**) Hierarchical clustering of non-malignant cell type composition per sample identifying twelve recurrent TME archetypes across sarcoma entities. Stacked bar graphs show the relative abundance of the 21 non-malignant cell types per sample. Annotation tracks indicate sarcoma entity, disease time point, and anatomical location. (**b**) Heatmap of scCODA log2-fold changes in cell-state abundance across the twelve TME archetypes, highlighting archetype-specific enrichment (red) or depletion (blue) relative to the reference cell type. Cell states with a credible compositional effect (scCODA, FDR at 0.2) are marked with **. (**c**) Kaplan-Meier plot showing post-sampling EFS for an exploratory tumor episode-indexed analysis, evaluating time from tumor sampling to the next documented event. Numbers denote tumor episodes; for patients with serially profiled samples, each episode was included based on their respective archetype classification. (**d-f**) Relative composition of stromal (**d**), lymphoid (**e**), and myeloid (**f**) cell states across major TME archetypes with n>3 samples and corresponding enrichment heatmap showing scCODA log2-fold changes. (**g-h**) MACSima imaging cyclic staining (MICS) spatial proteomics results for EWS_006_003 (lung/pleura metastasis of a patient diagnosed with Ewing sarcoma), STS_214_002 (abdominal relapse of a patient diagnosed with eRMS), and EWS_004_001 (lung metastasis of a patient diagnosed with Ewing sarcoma). Selected markers for shown cell types are depicted within the graph. Scale bar: 150 µm.

Shared cell-cell interaction phenotypes reveal convergent communication networks across pediatric sarcomas

As the tumor interactome represents a key point of intervention to disrupt TME crosstalk, we sought to define both entity-specific and conserved patterns of cellular communication across pediatric sarcomas. Cell-cell interaction analyses were performed using LIANA [38] for each sarcoma entity and TME archetype. Entity-specific analyses identified distinct ligand-receptor networks in Ewing sarcoma (11 interactions), osteosarcoma (48), aRMS (15), and eRMS (67), highlighting disease-specific communication programs (**Supplementary Fig. 9a-g**). We next integrated inferred ligand-receptor interactions across entities, identifying seven conserved tumor-derived ligand-receptor interactions shared across pediatric sarcomas as well as entity-specific interactions (**Supplementary Fig. 9 and Supplementary Table 5)**. Conserved interactions converged on immune-regulatory, inflammatory, and stromal remodeling pathways, including LGALS1-PTPRC, HMGB1-TLR2, CD74-, NOTCH2-, and CADM1-associated signaling axes (**Supplementary Fig. 9h)**.

To characterize global communication programs within the TME, ligand-receptor interactions were inferred for each sample across all cell types (**Fig. 7a**). Samples were subsequently clustered according to their interaction profiles, revealing four distinct interaction phenotypes: I0, stromal-dominated; I1, immune-engaged; I2, vascular-myeloid; and I3, ECM-remodeling/mixed (**Fig. 7b-c**). The stromal-dominated phenotype, I0, was characterized by extensive CAF-mediated signaling through collagen-DDR2 and collagen-integrin interactions. The immune-engaged phenotype, I1, comprised endothelial- and myeloid-derived interactions with T cell populations, particularly involving effector T cells expressing *CD44*. The vascular-myeloid phenotype, I3, was dominated by reciprocal communication between endothelial cells and macrophages, whereas the ECM-remodeling phenotype, I3, was enriched for interactions between Mac_ECM macrophages and mCAFs mediated by signaling pathways including TLR4, PLAU, and LRP1.

As observed for the TME archetypes, sarcoma entities, anatomical sites, and disease episodes were broadly distributed across all interaction phenotypes, underscoring the existence of conserved communication networks that transcend histological subtype (**Fig. 7d-e**). Likewise, the major TME archetypes, including the fibroblast-rich (C2), immune-rich (C5), and T cell-rich (C10) archetypes, predominantly localized to their corresponding interaction phenotypes while exhibiting partial overlap (**Fig. 7f**). These findings indicate that TME archetypes and interaction phenotypes capture distinct, yet, complementary aspects of the pediatric sarcoma microenvironment, integrating cellular composition with functional intercellular communication.

Comparative analysis of paired patient samples revealed substantial temporal plasticity of interaction phenotypes during disease progression (**Fig. 7g)** , further supporting the notion that interaction states are driven by factors beyond sarcoma entities. Relapse samples exhibited a shift toward the immune-engaged (I1) and vascular-myeloid (I2) interaction phenotypes, accompanied by a reduced frequency of the stromal-dominated (I0) and, most prominently, the ECM-remodeling/mixed (I3) phenotypes. These findings suggest dynamic remodeling of intercellular communication networks during disease progression and therapy (**Supplementary Fig. 10**). To explore the clinical relevance of the identified interaction phenotypes, EFS was analyzed as described for the TME archetypes. Given the limited patient numbers, these analyses should be considered exploratory. Nevertheless, samples classified as the immune-engaged phenotype I1 (*n*=13) showed a trend toward improved outcome (median EFS, 19 months),

whereas the vascular-myeloid phenotype I2 (*n*=12) was associated with the shortest median EFS (9 months). Intermediate outcomes were observed for the stromal-dominated phenotype I0 (*n*=22; median EFS, 15 months) and the ECM-remodeling/mixed phenotype I3 (*n*=8; median EFS, 13 months). Notably, relapse samples preferentially mapped to the two phenotypes associated with the most divergent clinical outcomes, I1 and I2, suggesting that disease progression is accompanied by distinct trajectories of tumor-microenvironment communication that may reflect differential engagement of the host immune response. Together, these findings illustrate that interaction-based phenotyping provides complementary information to compositional analyses by capturing functional cell-cell communication programs, offering additional insight into tumor evolution, potential determinants of clinical outcome, and candidate therapeutic vulnerabilities.

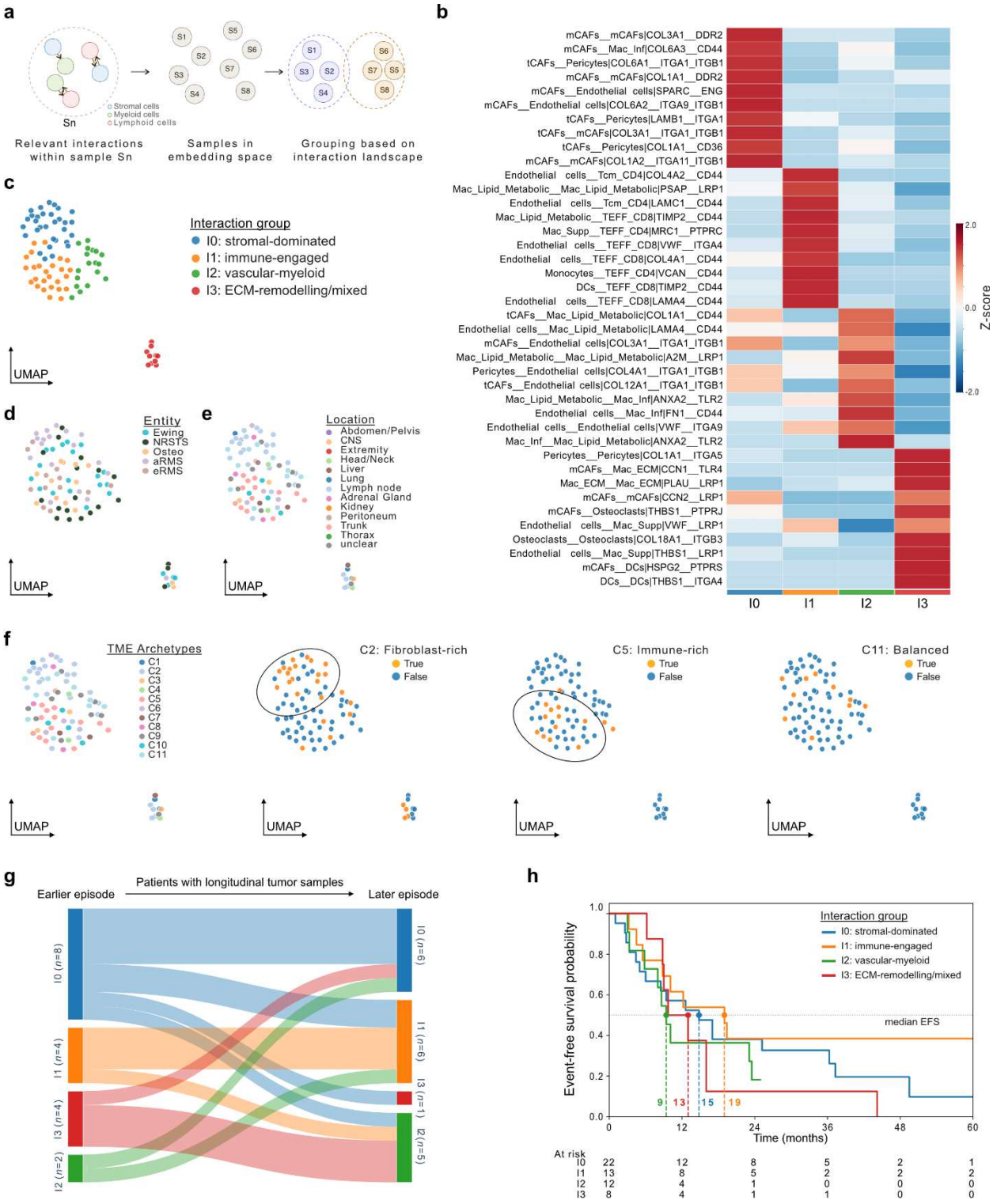


Figure 7: Interaction landscapes define cell-to-cell interaction networks across pediatric sarcomas. (**a**) Schematic of the interaction-based grouping approach: sample-specific ligand-receptor interactions were assembled into a shared interaction space in which each sample is represented by its interaction profile, and samples were clustered into interaction groups. (**b**) Heatmap

displaying the top 10 discriminating ligand-receptor interactions for each interaction phenotype. Row labels include legend cell type, receptor cell type, ligand, and receptor. Heatmap is colored by enrichment z-score, from blue (low) to red (high). (**c**) UMAP embedding of all samples based on their interaction analysis, identifying four interaction-based phenotypes: I0, stromal-dominated; I1, immune-engaged; I2, vascular-myeloid; I3, ECM-remodeling/mixed. (**d-f**) UMAPs of the interaction space, colored by sarcoma entity, location, and archetypes. (**h**) Sankey plot showing interaction group switches of the dominant interaction phenotype over time in patients with longitudinal sample pairs. (**h**) Kaplan-Meier plot showing the EFS probabilities stratified by interaction phenotypes.

Discussion

Here we present, to our knowledge, the largest single-nucleus atlas across high-risk pediatric sarcomas, integrating >800,000 single-nucleus transcriptomes from 89 tissue samples and 62 patients with spatial proteomic validation in selected representative cases. By combining copy-number inference, marker-based annotation, and cross-entity integration, we resolved malignant and non-malignant compartments across major sarcoma entities and defined 19 non-malignant cell types spanning stromal, lymphoid, and myeloid lineages. Beyond this granular cell type census, we identified recurrent, cross-entity TME archetypes and a mutual cell-to-cell interaction landscape that together suggest pediatric sarcomas share common microenvironmental factors irrespective of their underlying genomic drivers. Together, our findings establish a reference atlas and conceptual framework for dissecting the sarcoma TME, providing a foundation for biologically informed patient stratification and the discovery of therapeutic vulnerabilities.

Pediatric sarcomas have historically been considered immunologically "cold" and have shown limited responses to immune checkpoint blockade[39, 40]. Our data suggest more nuanced and diverse underlying TME biology than a uniformly cold phenotype, with immune-rich and T cell-rich archetypes present across entities. This raises several non-mutually-exclusive (immuno-)therapeutic opportunities that our atlas begins to delineate[41]: first, cellular therapies directed against tumor-associated antigens or TME constituents [42]. Second, strategies to "heat up" immunologically “cold” archetypes, for example through innate immune agonists or approaches that promote lymphocyte trafficking[43, 44]. Third, small-molecule or antibody-based inhibitors targeting suppressive or ECM-remodeling macrophage programs (e.g., CSF1R-directed agents) nominated by our macrophage state and interaction analyses[45]. Fourth, macrophage- or fibroblast-polarizing approaches aimed at shifting cells from tumor-permissive toward tumor-restrictive states[46]. The shared ligand-receptor axes identified across sarcoma entities provide a rational framework for prioritizing signaling pathways that may represent therapeutic vulnerabilities across the pediatric sarcoma spectrum, rather than being restricted to individual diagnoses.

More broadly, our study suggests that the pediatric sarcoma microenvironment is organized by principles extending beyond conventional histological classification. Despite entity-specific differences in TME composition and functional states, tumors converged on recurrent TME archetypes and intercellular communication programs shared across genetically and developmentally distinct sarcoma subtypes. This convergence supports the emerging concept that common ecological and immunological constraints, including stromal organization, immune adaptation and reciprocal tumor-host interactions, shape a finite number of conserved ecosystems in solid tumors that transcend tissue of origin [21, 47].

Moreover, growing evidence suggests that tumor behavior emerges from continuous interactions between malignant cells and their surrounding microenvironment rather than from tumor-intrinsic genetics alone[48]. Importantly, our interaction analyses extend the compositional view of the TME by demonstrating that tumors with similar cellular composition can nevertheless exhibit distinct intercellular communication states. This observation suggests that functional cell-cell

communication represents an additional layer of biological organization that may not be captured by cellular composition alone. By integrating both compositional archetypes and interaction phenotypes, our framework provides a more comprehensive description of the pediatric sarcoma microenvironment and may facilitate biologically informed patient stratification and the identification of candidate therapeutic vulnerabilities.

Several limitations of our atlas should be considered. The use of single-nucleus profiling of fresh-frozen archival material enabled retrospective analysis of a sizable cohort of rare pediatric sarcomas, but may bias recovery of certain cell populations and may incompletely capture cytoplasmic transcripts[49]. Spatial proteomic profiling provided orthogonal, spatially resolved validation of selected populations, including cell types less readily detected by snRNA-seq such as granulocytes, but was limited to representative samples rather than the full cohort. Ligand-receptor interactions were inferred computationally and motivate functional follow-up studies testing whether the shared and entity-specific interaction axes nominated here represent targetable dependencies to reprogram the sarcoma TME. Finally, survival analyses across TME archetypes and interaction phenotypes are hypothesis-generating and constrained by small subgroup sizes, heterogeneous prior disease course, and non-independence of serial tumor episodes from the same patient. Larger prospective cohorts will be needed to validate the stability, prognostic relevance, and treatment responsiveness of TME archetypes. Longitudinal molecular profiling from primary disease to relapse will test how tumors transition between archetypes and under therapeutic pressure.

Our relapse-predominant cohort based on INFORM represents a high-risk, pre-treated patient population as the most relevant target group for early-phase immunotherapeutic trials in the future. Our results argue for a shift in how TME-directed therapies for pediatric sarcomas are conceived and trialed. Whereas genomically targeted therapy has increasingly fractionated patients into smaller and smaller molecularly defined subgroups, our data suggest that TME-targeting strategies may instead be applicable across histological boundaries when patients are stratified by shared microenvironmental phenotype rather than by diagnosis alone. This raises the possibility of basket-style clinical trial designs in which TME archetype or interaction phenotype, rather than, or in addition to, tumor histology, guides enrollment and treatment selection[50], offering a route to develop microenvironment-directed therapies with relevance across the heterogeneous landscape of high-risk pediatric sarcomas.

Collectively, these findings suggest that the pediatric sarcoma TME is not merely a passive reflection of tumor histology, but an organizing axis of disease biology. Defining tumors by shared microenvironmental archetype and communication phenotype may enable rational selection of immune-, myeloid- and stroma-directed combinations across sarcoma entities. This atlas serves as a basis for developing TME-informed therapeutic strategies for children and adolescents with high-risk sarcoma.

Methods

Study design, eligibility, and participants

This study was performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. Retrospectively collected tumor tissue samples were obtained from the INFORM registry at Hopp Children's Cancer Center (German Clinical Trial Register: DRKS00007623; IRB approval: S-502/2013) and the biobank at University Hospital Tübingen (IRB approval: 819/2017BO1, 341/2024BO2). Written consent was obtained from each participant or their legal representatives. The study cohort included pediatric sarcoma patients with at least one banked tissue sample available for molecular profiling at single-cell resolution in the research laboratory (**Fig. 1a**). Where available, serial samples were included per patient for longitudinal analysis. Clinical data were collected independently of the molecular profiling.

Tissue processing

To prepare sample sections for single-nucleus RNA sequencing (snRNA-seq) and MACSima imaging cyclic staining (MICS), frozen tumor material embedded in Tissue-Tek OCT compound (Sakura Finetek™, 4583) were cut using a cryotome. To ensure close anatomical proximity of material used for the different analyses, we sampled tissue sections using an iterative protocol where 50 µm sections were cut for snRNA-seq and 4 µm sections were cut for MICS in an alternating manner. Consecutive cuts intended for snRNA-seq were combined into one microreaction tube and stored at -80 °C until isolation. MACSima sections were positioned on SuperFrost Plus slides (Epredia™, J1800AMNZ) and stored at -80 °C until experimental procedure.

Tumor bulk analyses

Tumor samples from patients enrolled in the INFORM trial were profiled by whole-exome-sequencing (WES), low-coverage (shallow) whole-genome sequencing (lcWGS), RNA-sequencing, and DNA methylation profiling based on Illumina EPIC array as previously described[13, 14]. Additional samples in the current cohort were subjected to the same analyses, if enough material was available after single-cell profiling.

Single-nucleus RNA sequencing (snRNA-seq)

Due to the hard osseous material in some samples, nuclei from frozen tissue were isolated using the Covaris cryoPREP workflow according to the published protocol "Single Nuclei Isolation using cryoPREP". In brief, scrolls were transferred into Covaris tissue tubes, flash frozen in liquid nitrogen, and dry pulverized at impact level 1/6 twice. The samples were resuspended in 700 µl of T-ST buffer and filtered through a 30 µm MACS SmartStrainer, which was washed once with additional 700 µl T-ST buffer. Samples were centrifuged at 500 g at 4°C for 5 minutes. After discarding the supernatant, samples were incubated with 500 µl Nuclei Extraction Buffer (Sigma Aldrich) on ice for 5 min. After centrifugation at 500 g at 4 °C for 5 min, samples were resuspended in ST buffer, and filtered through a 20 µm MACS SmartStrainer. The ST buffer contained NaCl (146 mM), Tris (pH 7.5, 10 mM), $CaCl_2$ (2 mM), and $MgCl_2$ (21 mM) in ultrapure water. To prepare T-ST buffer, ST buffer was supplemented with 0.03% Tween-20 and 0.01% bovine serum albumin

(BSA).

The final nuclei suspension was counted on the LUNA™ automated cell counter. 20,000 nuclei per sample were immediately processed for single-nuclei droplet formation, reverse transcription, and library preparation using the 10X Genomics Chromium Next GEM Single Cell 5' reagent kits v2 workflow (PN-1000286, 1000263, 1000190). The concentration and fragment distribution of the final libraries were assessed using Qubit HS DNA assay and Tapestation HS DNA D1000, respectively. Samples were pooled at equimolar amounts and subjected to paired-end sequencing (28+10+10+90 bp) on Illumina NovaSeq 6000, using either S1 or S2 lanes depending on the number of samples.

Spatial proteomics

Fixation, staining, and imaging of frozen patient samples using MICS were performed as previously described[51] and following a workflow comparable to that established for large-scale multiplexed tissue panels (ref). Briefly, sections were fixed in 4% paraformaldehyde (J19943.K2, Thermo Scientific) for 10 min at room temperature (RT) and washed three times with MACSima Running Buffer (130-121-565, Miltenyi Biotec). Slides were mounted onto a MACSwell Imaging Frame (One: 130-124-673; Two: 130-124-675, Miltenyi Biotec; working volumes 1,000 µl and 500 µl, respectively) and covered with Running Buffer prior to an initial nuclear counterstain. Immediately before MICS, sections were incubated with DAPI (130-111-570, Miltenyi Biotec; 1:10 in MACSima Running Buffer) for 10 min at RT in the dark and washed three times with Running Buffer. Fluorochrome-conjugated primary antibodies (APC, FITC, or PE; clones, dilutions, and fluorochromes listed in **Supplementary Table 1**) were diluted in MACSima Running Buffer in a MACSwell Deepwell Plate (130-126-865, Miltenyi Biotec) to the appropriate working volume, with DAPI (1:50) co-added to the antibody plate every eighth cycle.

Iterative cycles of antibody incubation, fluorescence imaging, and signal removal (photobleaching) were performed on a MACSima Imaging System (hardware v1.5.1, software v0.15.0), which automates reagent and sample handling via an xy-stage coupling a liquid-handling unit to a microscope with a monochromatic sCMOS camera. Each antibody was incubated for 10 min prior to fluorochrome excitation and image acquisition. Regions of interest (ROIs) were defined based on an overview DAPI scan, cross-referenced with an hematoxylin and eosin (H&E) stain of an adjacent section, and comprised overlapping fields of view (FoVs); autofocus was applied by default and refined manually per region based on DAPI signal.

Images were processed in the MACS iQ View software (v1.3.2): multiple exposures were merged into high-dynamic-range composites with automatic correction of overexposed regions. Flatfield and distortion correction were applied using instrument-specific calibration data, and adjacent fields were stitched based on the DAPI channel, with pre-stain bleaching images subtracted to yield final processed images (cropped 200 px top/bottom, 100 px left/right relative to the DAPI channel). For cell segmentation, we used the "Advanced Morphology for Tissue" algorithm within MACS iQ View for nuclear detection based on the DAPI signal, while cytoplasmic signal was assigned using the "Constrained Donut" approach with detection sensitivity and donut width optimized per region. Antibody performance was cross-validated across staining runs and against public reference data (Human Protein Atlas)[52] or conventional IHC. Marker intensities were

assigned to segmented cells for expert-guided gating and cell type annotation as described before[51], with additional markers overlaid as color maps to support annotation, and defined gates cross-verified with the corresponding staining images.

Single-nuclei RNA sequencing data processing

Count matrix generation

Raw reads were aligned to the human reference genome (GRCh38-2020-A) and quantified with 10x Genomics Cell Ranger v6.1.1[53] using cellranger count. Intronic reads were included (--include-introns=true) to capture the unspliced transcripts enriched in single-nuclei libraries, and samples sequenced across multiple lanes or libraries were merged within a single run. Filtered feature-barcode matrices were used for all downstream analyses, and nuclei are hereafter referred to as cells. Final CellRanger metrics for all samples are provided in **Supplementary Table 2**.

Ambient RNA removal

Because initial clustering revealed Ewing sarcoma-associated genes in otherwise non-malignant cells, ambient mRNA was estimated and removed per sample with SoupX v1.6.2[54] (R v4.2.0) from the paired raw and filtered matrices, using autoEstCont to estimate the contamination fraction (capped at 0.4). For Ewing sarcoma samples the automatic estimate underestimated contamination, so the fraction was set manually to 0.4. Corrected integer counts (adjustCounts) were used as input for all subsequent processing in Scanpy v1.11.5[55].

Quality control and doublet removal

Each sample was processed individually. Samples were excluded when Cell Ranger metrics indicated poor library quality (< 100 recovered nuclei, < 15% of reads confidently mapped to the transcriptome, or < 55% of reads in cells); low-quality nuclei and doublets were additionally removed per-nucleus. Cells with fewer than 200 detected genes and genes detected in fewer than 3 cells were removed, and per-cell quality-control metrics were computed (total UMI count, number of detected genes, and percentage of counts in mitochondrial and ribosomal genes). Cells were flagged as low-quality when their log1p total counts, log1p number of detected genes, percentage of counts in the top 20 genes, or mitochondrial percentage deviated by more than 5 median absolute deviations from the sample median. Doublets were identified with Scrublet v0.2.3[56] (expected doublet rate 0.07) and removed. Quality-control distributions were inspected before filtering, after doublet removal, and after final filtering to confirm that no coherent population was lost.

Normalization, feature selection, and clustering

Filtered counts were normalized to 10^4 counts per cell and log1p-transformed; this matrix was stored as the raw layer for differential expression and marker visualization, while the non-log-transformed matrix was retained for copy-number inference (see below). Highly variable genes were identified (flavor="seurat") and the matrix subset accordingly. Cell-cycle S and G2/M scores

were computed on the log-normalized data using the gene sets of Tirosh et al. (2016)[57]; total counts and mitochondrial percentage were then regressed out and values scaled to unit variance (clipped at 10 SD). PCA was performed on the scaled data (arpack solver), and a neighborhood graph was built from the first 15 principal components with 15 nearest neighbors. Cells were clustered with the Leiden algorithm[58] (resolution 0.5), and a Uniform Manifold Approximation and Projection (UMAP) embedding [59] was computed on the same graph for visualization.

Cell type annotation

Samples were analyzed individually and grouped by sarcoma subtypes. For each sample, a cluster-level automated annotation was generated to orient manual review, combining three complementary strategies followed by a consensus step: (i) marker-based annotation, matching cluster marker genes (differential expression against all other cells; scanpy.tl.rank_genes_groups, Wilcoxon rank-sum test) by gene overlap (scanpy.tl.marker_gene_overlap) to canonical markers from published references (including Hao et al., 2021)[60] and, independently, to PanglaoDB signatures (27 March 2020)[35]; (ii) reference-based classification with CellTypist v1.6.2[34] on the log-normalized data (Immune_All_High model, majority voting over Leiden clusters); and (iii) gene set enrichment of ranked cluster markers (adjusted $p < 0.05$) against MSigDB (v2023.1, human; collections C4, C5, C6, C8)[61] using gseapy v1.0.3[62]. Each strategy produced a per-cluster label and a coarse malignant/non-malignant call, and a per-cluster consensus was derived by majority vote.

Malignancy assessment and CNV inference

Malignant status was further assessed by inferring large-scale copy-number variation (CNV) with two complementary tools, inferCNV v1.20.0[63] and SCEVAN v1.0.1[64]. Both were run per sample on raw counts. For inferCNV, T cells were used as the normal reference, or other immune cell populations for samples lacking T cells, with a GENCODE v27 (hg38) gene-order file[65]; runs used a count cutoff of 0.1, per-group clustering, denoising, and HMM-based CNV prediction. SCEVAN was run with its automatically identified internal reference and returned a per-cell malignant/non-malignant classification. Clusters were called malignant when they showed consistent chromosome-arm-level alterations (inferCNV) and/or a consistent SCEVAN prediction, with single-cell profiles compared against matched bulk sequencing where available. Because some tumors (e.g. a subset of Ewing sarcomas) lack large-scale CNVs, malignancy was not defined by CNV alone but also by expression of a curated panel of entity-specific sarcoma markers (**Fig. 1f; Supplementary Table 3**). Cluster identity was assigned from the concordance of CNV status and marker expression, with markers taking precedence for genomically flat samples.

Integration across the cohort

To analyze the full cohort jointly, the 89 per-sample datasets were merged into a single object retaining the expert-verified per-sample annotations. Several integration approaches were compared (Scanorama, harmony, LIGER, limma, ComBat). For the full dataset, scANVI[66] was selected as it best separated the malignant from the non-malignant compartment while preserving sample-level heterogeneity and grouping samples of the same entity together. For the smaller

non-malignant subsets, Harmony[67] was used instead, as it resolved distinct cell states more clearly and scANVI performed less well on datasets of this size. Before integration, the merged dataset was renormalized (10^4 counts per cell, log-transformed) and 3,000 highly variable genes were selected in a batch-aware manner. Integration for the full dataset was then performed with scvi-tools v1.4.2[68]. A scVI model[69] was first trained on the SoupX-corrected counts, with sample and sequencing chip encoded as categorical batch covariates and total counts and mitochondrial percentage as continuous covariates. A scANVI model was then initialized from it using the per-sample cell type labels to yield the integrated latent representation. For the subsets, Harmony was run on the first 50 principal components with sample and sequencing chip as batch covariates.

Cell state annotation

Cell states were further resolved within the major immune and stromal compartments using an integrative strategy combining differential expression, GO/pathway and gene set enrichment analyses, and reference-based label transfer, followed by manual curation. T cell subsets were annotated using the pan-cancer T cell atlas (Chu et al.[70]) and fibroblast/stromal states using the cancer-associated fibroblasts (CAF) reference from Cords et al.[71], whereas macrophage states, poorly separated by canonical M1/M2 markers (**Supplementary Fig. 6**), were instead defined functionally through pathway enrichment combined with reference from Guimaraes et al.[37]. All annotations were validated both computationally and through manual review.

Cell type compositional analysis

Differences in cell-type composition across tumor entities, samples and archetypes were assessed using the single-cell compositional data analysis framework scCODA[36], run through the pertpy implementation. Cell-level annotations were aggregated into per-sample cell-type count tables, from which the proportion of each cell type per sample was derived. Group effects were modelled using sum contrast coding of the covariate, such that the effect of each group was expressed relative to the grand mean across all levels. Since scCODA measures each cell type relative to a fixed reference, the automatic reference selection was used, which selects the cell type with the lowest relative-abundance dispersion among those present in the majority of samples. Where indicated, the reference was instead fixed manually to a cell type verified to have high cross-sample presence and low relative-abundance variance. Parameter inference was performed with the No-U-Turn Hamiltonian Monte Carlo sampler using the default chain length. For each cell-type-covariate combination, scCODA's Bayesian model estimates the posterior probability of a credible effect on cell-type abundance, and effects were called credible at a false discovery rate of 0.2 following the scCODA guidelines. As scCODA structurally omits one group level, each analysis was run twice with the group ordering rotated so that a different level was dropped in each run while keeping the reference cell type fixed, and the two results were merged to obtain an effect estimate for every group. Effects were visualized as heatmaps of the log2-fold change relative to the reference cell type, with cell types showing a statistically credible effect highlighted (**) in the main figures.

Definition of tumor microenvironment (TME) archetypes in sarcoma

To define TME archetypes, each sample was characterized by the composition of its non-

malignant compartment. Malignant cells were excluded, and the relative abundance of each non-malignant cell type was computed per sample as its fraction among all non-malignant cells, yielding a sample-by-cell-type proportion matrix. Samples were grouped by agglomerative hierarchical clustering of this matrix using Ward's linkage on Euclidean distances between proportion profiles (scipy.cluster.hierarchy)[72]. The resulting dendrogram was inspected across a range of cut settings, and a partition into 12 clusters was selected as the configuration that best resolved biologically coherent groupings of non-malignant cell type distributions.

Cell-cell interaction analysis

Ligand-receptor interactions were inferred with LIANA[38, 73] on the single-nucleus RNA-seq dataset, using the OmniPath[74] Consensus ligand-receptor interaction database. The rank_aggregate method was applied on log-normalized counts, which aggregates the predictions of multiple individual scoring methods into consensus magnitude and specificity ranks via robust rank aggregation. Ligands or receptors expressed in fewer than 10% of cells per cluster were excluded, corresponding to the default expression proportion. For each interaction, the magnitude rank reflects the overall expression strength of the ligand and receptor, whereas the specificity rank reflects how specifically the pair is expressed in a given pair of cell types. For both metrics, lower ranks indicate more relevant interactions. Interactions were prioritized per group by retaining only those passing a threshold of 0.1 on both the magnitude and specificity ranks. The analysis was performed at multiple cell type annotation levels and was run separately for the full cohort, each tumor entity, each sample, and each TME archetype.

Sample grouping based on the cell-cell communication landscape

To group samples by their intercellular communication landscape, ligand-receptor interactions were inferred per sample with LIANA at the detailed cell-state level as described above. For each sample, interactions were filtered to retain only those passing a threshold of 0.1 on both the magnitude and specificity ranks, and each retained interaction was assigned a combined score defined as the product of (1 − magnitude rank) and (1 − specificity rank), so that interactions that are both strongly and specifically expressed received higher values. Each interaction was defined by the combination of the sender and receiver cell type and the corresponding ligand and receptor complex. The per-sample scores were assembled into a sample-by-interaction matrix, with absent interactions set to zero and interactions with no variance across samples removed. The matrix was scaled without mean-centering to account for its sparsity and reduced to 50 components by truncated singular value decomposition, of which the first 30 were retained. A nearest-neighbor graph (k = 10) was constructed on these components, and samples were grouped into clusters by Leiden clustering at a resolution of 1.0. To identify the interactions driving each cluster, each interaction was first tested for differential activity across all clusters with the Kruskal-Wallis test, followed by a one-versus-rest comparison of each cluster against all remaining samples using the Mann-Whitney U test, with p-values corrected for multiple testing using the Benjamini-Hochberg procedure.

Statistics and reproducibility

No statistical method was used to predetermine sample size; the cohort comprised all available pediatric sarcoma specimens across five entities that passed quality control (*n*=89). Analyses

were performed in Python (scanpy v1.11.5). Differential cell type composition was assessed with scCODA (v0.1.9); cell types with a statistically credible effect at the specified false discovery rate (FDR) (see "Compositional analysis") are marked ** in heatmaps. Cell-cell interactions were inferred with LIANA (v1.7.1), retaining interactions with magnitude and specificity ranks ≤ 0.1. Interaction-based sample groups were compared with the two-sided Mann-Whitney U test (scipy v1.17.1) with Benjamini-Hochberg FDR correction (statsmodels v0.14.6). For exploratory clinical association analyses, each sequenced tumor sample was treated as an index tumor episode. Post-sampling EFS was estimated using Kaplan-Meier analysis (lifelines v0.30.3)[75] and defined as the interval from tumor sampling, or INFORM enrollment if sampling date was unavailable, to the next documented progression, relapse, death or last follow-up. For patients with longitudinal samples, each tumor episode was included based on stratification according to the sample-level archetype and interaction phenotype classification. Unless otherwise specified, tests were two-sided with significance at $p < 0.05$ or the stated FDR thresholds.

References

1. Burningham Z, Hashibe M, Spector L, Schiffman JD. The epidemiology of sarcoma. *Clin Sarcoma Res* **2**, 14 (2012).
2. Wang M, Bi Y, Fan Y, Fu X, Jin Y. Global incidence of childhood cancer by subtype in 2022: a population-based registry study. *eClinicalMedicine* **89**, (2025).
3. Polfer EM, Davis W, Eckhoff MD, Wells ME, Singh V, Rajani R. Ewing Sarcoma in the Pediatric Population: Predictors of Survival Within the United States. *JAAOS: Global Research and Reviews* **8**, (2024).
4. Pfister SM*, et al.* A Summary of the Inaugural WHO Classification of Pediatric Tumors: Transitioning from the Optical into the Molecular Era. *Cancer Discovery* **12**, 331–355 (2022).
5. Board WCoTE. *Soft Tissue and Bone Tumours*, 5th edn. International Agency for Research on Cancer (2020).
6. Board WCoTE. *Paediatric Tumours*, 5th edn. International Agency for Research on Cancer (2023).
7. Perkins SM, Shinohara ET, DeWees T, Frangoul H. Outcome for children with metastatic solid tumors over the last four decades. *PLoS One* **9**, e100396 (2014).
8. Grünewald TGP*, et al.* Ewing sarcoma. *Nature Reviews Disease Primers* **4**, 5 (2018).
9. Beird HC*, et al.* Osteosarcoma. *Nature Reviews Disease Primers* **8**, 77 (2022).
10. Skapek SX*, et al.* Rhabdomyosarcoma. *Nature Reviews Disease Primers* **5**, 1 (2019).
11. Ferrari A*, et al.* Pediatric Non-Rhabdomyosarcoma Soft Tissue Sarcomas: Standard of Care and Treatment Recommendations from the European Paediatric Soft Tissue Sarcoma Study Group (EpSSG). *Cancer Management and Research* **14**, 2885–2902 (2022).
12. Berlanga P*, et al.* The European MAPPYACTS Trial: Precision Medicine Program in Pediatric and Adolescent Patients with Recurrent Malignancies. *Cancer Discov* **12**, 1266–1281 (2022).
13. Worst BC*, et al.* Next-generation personalised medicine for high-risk paediatric cancer patients – The INFORM pilot study. *European Journal of Cancer* **65**, 91–101 (2016).
14. van Tilburg CM*, et al.* The Pediatric Precision Oncology INFORM Registry: Clinical Outcome and Benefit for Patients with Very High-Evidence Targets. *Cancer Discovery* **11**, 2764–2779 (2021).
15. Heipertz A-E*, et al.* Outcome of Children and Adolescents With Relapsed/Refractory/Progressive Malignancies Treated With Molecularly Informed Targeted Drugs in the Pediatric Precision Oncology Registry INFORM. *JCO Precision Oncology*, e2300015 (2023).
16. June CH, O'Connor RS, Kawalekar OU, Ghassemi S, Milone MC. CAR T cell immunotherapy for human cancer. *Science* **359**, 1361–1365 (2018).
17. Majzner RG, Mackall CL. Clinical lessons learned from the first leg of the CAR T cell journey. *Nat Med* **25**, 1341–1355 (2019).
18. Chen T, Wang M, Chen Y, Liu Y. Current challenges and therapeutic advances of CAR-T cell therapy for solid tumors. *Cancer Cell Int* **24**, 133 (2024).
19. Yan T, Zhu L, Chen J. Current advances and challenges in CAR T-Cell therapy for solid tumors: tumor-associated antigens and the tumor microenvironment. *Experimental Hematology & Oncology* **12**, 14 (2023).
20. Lamplugh ZL, Wellhausen N, June CH, Fan Y. Microenvironmental regulation of solid tumour resistance to CAR T cell therapy. *Nature Reviews Immunology* **26**, 230–248 (2026).
21. Combes AJ, Samad B, Krummel MF. Defining and using immune archetypes to classify

and treat cancer. *Nature Reviews Cancer* **23**, 491–505 (2023).
22. Jeong S*, et al.* Sarcoma Immunotherapy: Confronting Present Hurdles and Unveiling Upcoming Opportunities. *Molecules and Cells* **46**, 579–588 (2023).
23. Fatahichegeni M, Ansarian MA, Xiao M, Gao W, Shi R. Immunotherapy for pediatric solid tumors: overcoming biological barriers through rational multimodal combinations. *Cancer Immunology, Immunotherapy* **75**, 146 (2026).
24. Majzner RG, Heitzeneder S, Mackall CL. Harnessing the Immunotherapy Revolution for the Treatment of Childhood Cancers. *Cancer Cell* **31**, 476–485 (2017).
25. Grünewald TGP, Oehme I, Wagner J, Fröhling S. Intratumor heterogeneity and clonal evolution in sarcomas: Molecular, cellular, and functional drivers of therapy resistance. *Seminars in Cancer Biology* **120**, 31–47 (2026).
26. Zhou Y*, et al.* Single-cell RNA landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. *Nature Communications* **11**, 6322 (2020).
27. DeMartino J*, et al.* Single-cell transcriptomics reveals immune suppression and cell states predictive of patient outcomes in rhabdomyosarcoma. *Nature Communications* **14**, 3074 (2023).
28. Visser LL, Bleijs M, Margaritis T, van de Wetering M, Holstege FCP, Clevers H. Ewing Sarcoma Single-cell Transcriptome Analysis Reveals Functionally Impaired Antigen-presenting Cells. *Cancer Research Communications* **3**, 2158–2169 (2023).
29. Zhang Z*, et al.* Integrated analysis of single-cell and bulk RNA sequencing data reveals a pan-cancer stemness signature predicting immunotherapy response. *Genome Medicine* **14**, 45 (2022).
30. Orth MF*, et al.* Systematic multi-omics cell line profiling uncovers principles of Ewing sarcoma fusion oncogene-mediated gene regulation. *Cell Reports* **41**, 111761 (2022).
31. Danielli SG*, et al.* Single cell transcriptomic profiling identifies tumor-acquired and therapy-resistant cell states in pediatric rhabdomyosarcoma. *Nature Communications* **15**, 6307 (2024).
32. Truong DD*, et al.* Mapping the Single-Cell Differentiation Landscape of Osteosarcoma. *Clinical Cancer Research* **30**, 3259–3272 (2024).
33. Geyer FH*, et al.* Comprehensive DSRCT multi-omics analyses unveil CACNA2D2 as a diagnostic hallmark and super-enhancer-driven EWSR1::WT1 signature gene. *Cancer Communications* **45**, 702–708 (2025).
34. Domínguez Conde C*, et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
35. Franzén O, Gan L-M, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, (2019).
36. Büttner M, Ostner J, Müller CL, Theis FJ, Schubert B. scCODA is a Bayesian model for compositional single-cell data analysis. *Nature Communications* **12**, 6876 (2021).
37. Guimarães GR*, et al.* Single-cell resolution characterization of myeloid-derived cell states with implication in cancer outcome. *Nature Communications* **15**, 5694 (2024).
38. Dimitrov D*, et al.* Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nature Communications* **13**, 3224 (2022).
39. Rytlewski J, Milhem MM, Monga V. Turning 'Cold' tumors 'Hot': immunotherapies in sarcoma. *Annals of Translational Medicine* **9**, 1039 (2021).
40. Petitprez F*, et al.* B cells are associated with survival and immunotherapy response in sarcoma. *Nature* **577**, 556–560 (2020).
41. Dyson KA*, et al.* Emerging trends in immunotherapy for pediatric sarcomas. *Journal of Hematology & Oncology* **12**, 78 (2019).
42. Sahin TK*, et al.* Cell therapy in sarcoma: current landscape and future directions. *J Immunother Cancer* **14**, (2026).

43. Wu B, Zhang B, Li B, Wu H, Jiang M. Cold and hot tumors: from molecular mechanisms to targeted therapy. *Signal Transduction and Targeted Therapy* **9**, 274 (2024).
44. Liu Y-T*, et al.* Turning cold tumors into hot tumors to ignite immunotherapy. *Molecular Cancer* **24**, 254 (2025).
45. Koo J, Hayashi M, Verneris MR, Lee-Sherick AB. Targeting Tumor-Associated Macrophages in the Pediatric Sarcoma Tumor Microenvironment. *Frontiers in Oncology* **Volume 10 - 2020**, (2020).
46. Yang D, Liu J, Qian H, Zhuang Q. Cancer-associated fibroblasts: from basic science to anticancer therapy. *Experimental & Molecular Medicine* **55**, 1322–1332 (2023).
47. Zhang W*, et al.* Non-invasive profiling of the tumour microenvironment with spatial ecotypes. *Nature* **654**, 1076–1087 (2026).
48. Wang Q*, et al.* Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell* **32**, 42–56.e46 (2017).
49. Ding J*, et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature Biotechnology* **38**, 737–746 (2020).
50. van Tilburg CM*, et al.* INFORM2 NivEnt: The first trial of the INFORM2 biomarker driven phase I/II trial series: the combination of nivolumab and entinostat in children and adolescents with refractory high-risk malignancies. *BMC Cancer* **20**, 523 (2020).
51. Scheuermann S*, et al.* Unveiling spatial complexity in solid tumor immune microenvironments through multiplexed imaging. *Frontiers in Immunology* **Volume 15 - 2024**, (2024).
52. Uhlén M*, et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
53. Zheng GXY*, et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017).
54. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience* **9**, (2020).
55. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).
56. Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems* **8**, 281–291.e289 (2019).
57. Tirosh I*, et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
58. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 5233 (2019).
59. McInnes L, Healy J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).
60. Hao Y*, et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology* **42**, 293–304 (2024).
61. Subramanian A*, et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).
62. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. *Bioinformatics* **39**, (2023).
63. T T, I T, C G, M B, B H. inferCNV of the Trinity CTAT Project.). Klarman Cell Observatory, Broad Institute of MIT and Harvard (2019).
64. De Falco A, Caruso F, Su X-D, Iavarone A, Ceccarelli M. A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. *Nature Communications* **14**, 1074 (2023).
65. Frankish A*, et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47**, D766–D773 (2019).
66. Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization

and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology* **17**, MSB20209620 (2021).
67. Korsunsky I*, et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* **16**, 1289–1296 (2019).
68. Gayoso A*, et al.* A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology* **40**, 163–166 (2022).
69. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* **15**, 1053–1058 (2018).
70. Chu Y*, et al.* Pan-cancer T cell atlas links a cellular stress response state to immunotherapy resistance. *Nature Medicine* **29**, 1550–1562 (2023).
71. Cords L*, et al.* Cancer-associated fibroblast classification in single-cell and spatial proteomics data. *Nature Communications* **14**, 4294 (2023).
72. Virtanen P*, et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).
73. Dimitrov D*, et al.* LIANA+ provides an all-in-one framework for cell–cell communication inference. *Nature Cell Biology* **26**, 1613–1622 (2024).
74. Türei D*, et al.* Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology* **17**, MSB20209923 (2021).
75. Davidson-Pilon C. lifelines: survival analysis in Python. *Journal of Open Source Software* **4**, 1317 (2019).

Acknowledgements

This work was mainly funded through a grant of the Deutsche Forschungsgemeinschaft (DFG GR3728/5-1, to T.G.P.G., K.W.P., and C.M.S.).

The INFORM program is financially supported by the German Cancer Research Center (DKFZ), several German health insurance companies, the German Cancer Consortium (DKTK), the German Federal Ministry of Education and Research (BMBF), the German Federal Ministry of Health (BMG), the Ministry of Science, Research and the Arts of the State of Baden-Württemberg (MWK BW); the German Cancer Aid (DKH), the German Childhood Cancer Foundation (DKS), RTL television, the aid organization BILD hilft e.V. (Ein Herz für Kinder) and the generous private donation of the Scheu family.

We would like to express our sincere thanks to Carsten Maus, Erjia Wang (Genomics and Proteomics Core Facility, DKFZ). Lena Weiser, Gregor Warsow (Omics IT and Data Management Core Facility, DKFZ) for their highly dedicated support in data management and processing. Gnanaprakash Balasubramanian, Christopher Previti, and Rolf Kabbe (Division of Pediatric Neurooncology, DKFZ) for their sincere and dedicated contribution to the bioinformatics analyses.

The authors would like to thank the Single-cell Open Lab at DKFZ Heidelberg for excellent services.

We thank the biobank at University Hospital Tübingen, for supporting this study and gratefully acknowledge the participating patients and their families.

Author Contribution

T.G.P.G, K.W.P, and C.M.S acquired funding for the study, T.G.P.G., K.W.P, K.K.M, and C.M.S conceived the study and K.W.P., K.K.M., and C.M.S. jointly supervised the study. T.T.F. and K.K.M. collected and processed the samples for sequencing experiments. K.C.C performed overall bioinformatic data analysis with contributions from A.K. . Bioinformatic support was provided by S.V., I.S. and R.J.A. and bioinformatic analyses were supervised by S.T. Results were interpreted together by K.C.C, T.T.F, S.S., T.G.P.G., K.W.P, K.K.M, and C.M.S. . S.S., F.E., A.N, and B.K. performed and analyzed MICS experiments. R.I., J.-P.M., B.F.B.M, K.S., D.T.W.J., O.W., M.N., M.S.-S., U.D., S.M.P., and J.F. contributed samples, collected data, or helped with experimental procedures and data interpretation. K.C.C., S.S., T.T.F., F.E., K.K.M., and C.M.S. wrote the manuscript. All authors have read and approved the manuscript.

Competing Interests

The authors declare no competing interests.

Data availability

Processed MICS images and analysis workflows for samples depicted in Fig. 6 are deposited at: **10.5281/zenodo.21292448**. Other processed imaging data are available from the authors upon reasonable request. The snRNA sequencing datasets generated during this study are available in the European Genome-phenome Archive (EGA) under accession number EGAS50000000393.

Code availability

Code related to the analyses in this study mainly reuse codes from other softwares (SoupX, Scanpy, Scrublet, scVI, scANVI, scvi-tools, leidenalg, umap-learn, CellTypist, gseapy, inferCNV, SCEVAN, scCODA, pertpy, LIANA, OmniPath, lifelines) and can be found on GitHub at **https://github.com/karlacat1/single_cell_atlas_sarcoma** under MIT License

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx
- SupplementaryTable3.xlsx
- SupplementaryTable4.xlsx
- SupplementaryTable5.xlsx
- SupplementaryFigures.pdf