Università di Genova

Transcriptomic Landscape of Radiation-Induced Breast Angiosarcoma

Master of Science in
MEDICAL-PHARMACEUTICAL BIOTECHNOLOGY

Supervisors:
Prof. Paolo Malatesta
Prof. Lorenzo Ferrando

Co- Supervisor:
Prof. Paola Ghiorzo

Candidate:
Amirpasha Mahmoudi

Academic Year

2025-2026

Acknowledgments

Laudation to the God of majesty and glory! Obedience to him is a cause of approach and gratitude in the increase of benefits. Every inhalation of the breath prolongs life and every expiration of it gladdens our nature; wherefore every breath confers two benefits and for every benefit gratitude is due. Whose hand and tongue are capable of fulfilling the obligations of thanks to him? (Gulistan Sa'di, 1258 CE)

First and foremost, I want to thank my ***Mother*** and ***Father***. I have not been able to see them in person for two years, and every page of this thesis carries something of what they have given up so that I could be here. Their support has meant everything.

To ***Prof. Lorenzo Ferrando,*** thank you for trusting me and believing in me from the very beginning. Your constant support, the time you gave me, your guidance, and your patience carried me through every stage of this work. But more than anything, thank you for showing me what it means to truly love research that is something I will carry with me long after this thesis is finished.

I am also grateful to ***Prof. Gabriele Zoppoli***, Associate Professor of Internal Medicine at DIMI, for his role in leading our research group and for the environment and resources that made this project possible.

I would also like to thank the ***Italian people***, whose rich culture and warm, friendly nature made this transition far easier for me. Genoa has become a home I did not expect to find so far from my own.

Abstract

Breast angiosarcoma is a rare and aggressive cancer arising from endothelial cells, comprising two clinically distinct aetiological subtypes: radiation-induced (RTI) and non-radiation-induced (Non-RTI) angiosarcoma. The transcriptomic landscape differentiating these subtypes has not been systematically characterised.

Bulk RNA sequencing data from an internal discovery cohort (three European centres: Genova, Milan, Leuven) and an external validation cohort (The Angiosarcoma Project, cBioPortal) were analysed using a discovery-validation design incorporating differential gene expression, pathway enrichment, and immune microenvironment analyses.

RTI patients were older at diagnosis than Non-RTI patients in both cohorts (internal cohort, n=32, 27 RTI/5 Non-RTI: median 73.5 vs 37.2 years, p=0.0004; external cohort, n=55, 24 RTI/31 Non-RTI: median 63.0 vs 38.0 years, p<0.0001). Twelve Hallmark pathways were significantly enriched in RTI angiosarcoma, including MYC targets, E2F targets, mTORC1 signalling, and oxidative phosphorylation, all of which were replicated in the external validation cohort. Non-RTI angiosarcoma showed enrichment of epithelial-mesenchymal transition and myogenesis pathways, consistent with a stromal and mesenchymal transcriptional program. RTI tumours exhibited significantly higher immune infiltration, including elevated T cell and myeloid dendritic cell abundance, a higher overall immune score, and higher antigen presentation signatures. In contrast, Non-RTI tumours showed significantly higher endothelial cell and fibroblast abundance, consistent with a stromal and immune-cold microenvironment.

These findings support a transcriptomic basis for the immune-hot and immune-cold phenotypes of RTI and Non-RTI breast angiosarcoma, providing a molecular rationale for stratifying these subtypes and informing immune checkpoint blockade strategies targeting CTLA-4 and TIGIT in RTI disease.

Table of Contents

CHAPTER I: Introduction and Literature Review

1.1 Transcriptomics in cancer research

Bulk RNA sequencing (RNA-seq) is a high-throughput sequencing approach that enables the simultaneous quantification of thousands of transcripts across the genome, providing a comprehensive view of gene expression, pathway activity, and immune microenvironment composition from a single experiment (1). A schematic overview of the bulk RNA sequencing workflow, from tissue collection to biological interpretation, is shown in Figure 1.1.1. Since its introduction, RNA-seq has transformed transcriptome profiling in the life sciences, combining transcript discovery and quantification in a single assay that has become a standard tool across the biomedical research community (1,2). In breast cancer, gene expression profiling of surgical tumour specimens demonstrated that tumours could be classified into subtypes distinguished by pervasive differences in their gene expression patterns, with each tumour displaying a distinctive molecular portrait reflecting its biological diversity (3). Building on this foundation, the 50-gene PAM50 classifier further defined four intrinsic subtypes, luminal A, luminal B, HER2-enriched, and basal-like, each with distinct prognosis and predicted response to chemotherapy, providing prognostic and predictive information beyond standard clinical parameters (4). Large-scale initiatives such as The Cancer Genome Atlas have demonstrated that gene expression profiling can identify molecularly distinct subgroups within histologically defined cancer types, revealing pathway-level differences with direct implications for prognosis and treatment selection in translational cancer research (5).

Soft tissue sarcomas represent a heterogeneous umbrella category encompassing more than 70 distinct histological subtypes with highly variable biological behaviour, molecular profiles, and clinical outcomes (6). In this setting, where small patient numbers and histological heterogeneity limit the resolution of clinical studies, transcriptomic profiling has proven particularly valuable, revealing prognostic signatures and pathway-level distinctions with direct implications for patient management (7). For rare malignancies, the application of RNA-seq to formalin-fixed paraffin-embedded (FFPE) archival tissue, an invaluable resource maintained at cancer care institutions worldwide and linked to

long-term patient clinical data, enables retrospective molecular analyses that would otherwise be unfeasible due to the scarcity of prospectively collected fresh tissue (8).

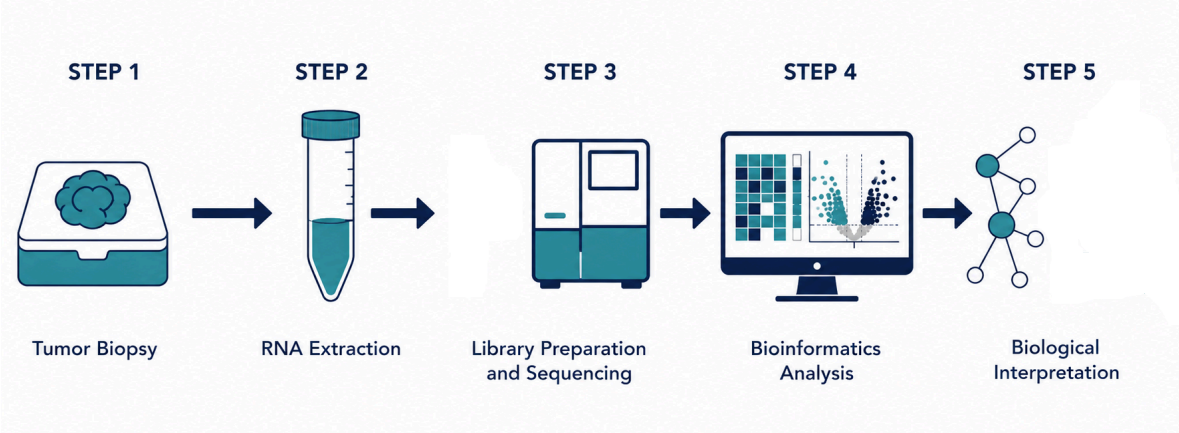


Figure 1.1.1 Schematic overview of the bulk RNA sequencing workflow, from tumour tissue collection through RNA extraction, library preparation and sequencing, bioinformatic analysis, to biological interpretation.

1.2 Angiosarcoma: definition, epidemiology, aetiology, and pathology

Angiosarcoma is a rare and aggressive cancer arising from endothelial cells, prone to recurrence and metastasis, which can arise in any part of the body, most commonly presenting as a cutaneous tumour of the head and neck, but also occurring in the breast, deep soft tissues, and visceral organs including the liver (9,10). Epidemiological data on angiosarcoma remain limited. In France, a population-based registry study reported an age-standardised incidence rate of 0.15 per 100,000 person-years over the period 1979 to 2016, rising to 0.26 per 100,000 person-years between 2007 and 2016, suggesting an increasing trend consistent with European estimates (Figure 1.2.1, 11). This increasing trend is particularly notable among older patients, most notably those older than 75 years, and among women with a prior history of breast cancer treatment, in whom the growing use of adjuvant radiotherapy following breast-conserving surgery has been associated with a rising incidence of radiation-induced (RTI) breast angiosarcoma (11,12). In the United States, the annual number of new angiosarcoma diagnoses doubled from 657 cases in 2001 to 1312 cases in 2019 (Figure 1.2.2, 12). This increase is largely attributed to the growing

number of breast cancer survivors who have received adjuvant radiotherapy following breast-conserving surgery, as prior radiation exposure represents the primary risk factor for the subsequent development of RTI angiosarcoma in the previously irradiated field (11,12).

Angiosarcoma arises through distinct aetiological pathways. Non-radiation-induced (Non-RTI) angiosarcoma develops de novo at various anatomical sites with no identifiable causative factor, while RTI angiosarcoma develops over areas of previously irradiated skin, particularly in the pectoral area of women who have undergone radiotherapy after breast cancer treatment (11). Additional forms include angiosarcoma associated with chronic lymphedema, known as Stewart-Treves syndrome, and those associated with UV light exposure (13).

Histologically, angiosarcoma is characterised by the proliferation of atypical endothelial cells forming irregular vascular channels, with a morphological spectrum ranging from vasoformative patterns in low-grade lesions to densely cellular and infiltrative growth in high-grade tumours (10). RTI breast angiosarcoma is typically high grade and poorly differentiated, often presenting as multifocal or diffuse lesions with epithelioid to spindle cell morphology arranged in solid nests intermixed with blood-filled spaces (14). Histopathological identification can be challenging as angiosarcoma may mimic other vascular tumours or carcinomas, and immunohistochemistry is useful in the diagnostic process (10,14). The endothelial markers CD31, ERG, CD34, and factor VIII, alongside vimentin and FLI1, are characteristically positive, and the proliferation index as assessed by Ki-67 is usually high (14). At the molecular level, MYC amplification has been identified as a hallmark of RTI angiosarcoma, observed in all RTI cases but absent in Non-RTI tumours (14,15).

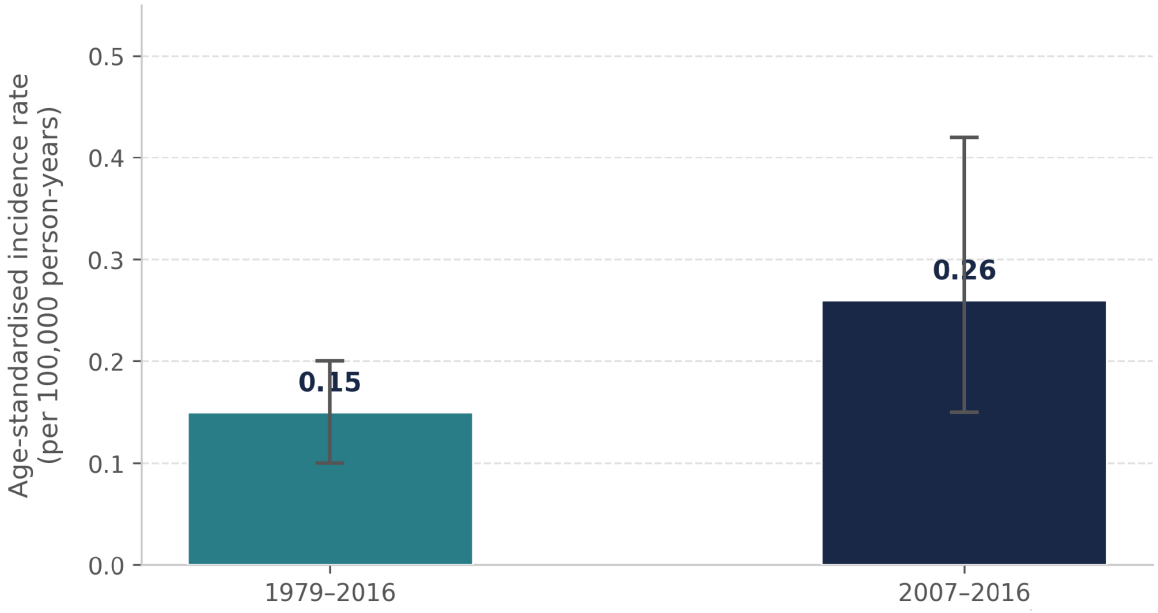


Figure 1.2.1 Age-standardised incidence rate of angiosarcoma in a French population-based registry (Doubs department), comparing the overall study period (1979–2016) and the most recent decade (2007–2016). Error bars represent 95% confidence intervals (11)

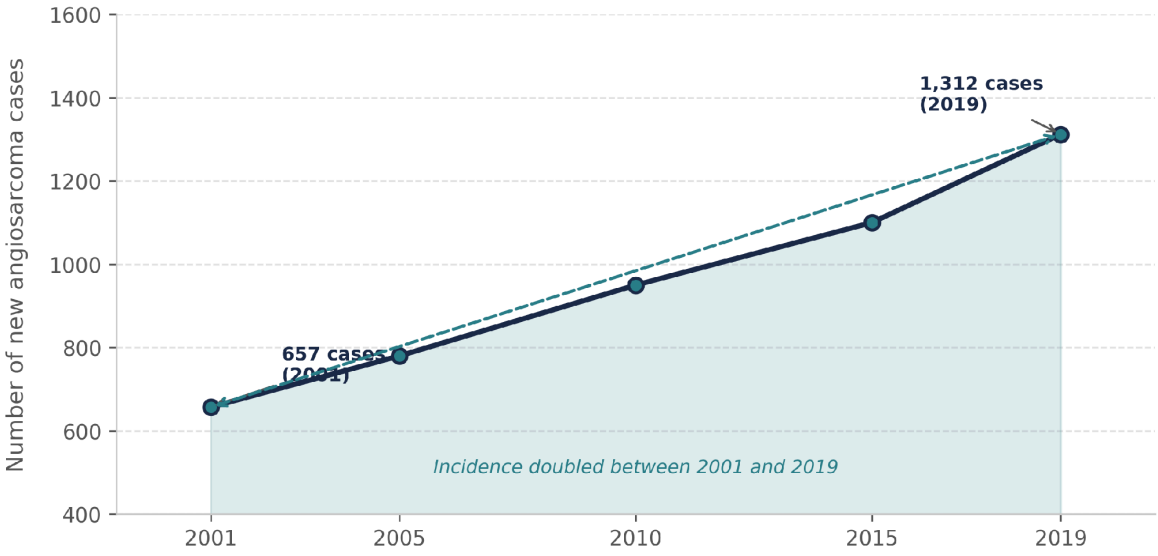


Figure 1.2.2 Trend in annual angiosarcoma diagnoses in the United States between 2001 and 2019, showing a doubling of incidence over this period (12).

1.3 Breast angiosarcoma: epidemiology and clinical subtypes

Breast angiosarcoma represents a distinct and particularly rare subset of angiosarcoma, accounting for approximately 13.4% of all angiosarcoma cases reported in the SEER database between 1975 and 2016 (Table 1, 16,17). Among all breast sarcomas, angiosarcoma is the predominant histological subtype and carries a particularly poor prognosis (16). The annual incidence of breast sarcomas overall is estimated at 4.48 cases per million women, with angiosarcoma representing the majority of this figure (16).

Existing literature describes breast angiosarcoma as comprising two clinically and biologically distinct subtypes based on aetiology, RTI and Non-RTI. Non-RTI angiosarcoma is reported to arise de novo within the breast parenchyma in women without a prior history of breast cancer or radiation exposure, typically affecting younger patients (16). RTI angiosarcoma is described as developing as a complication of prior radiotherapy administered as part of breast cancer treatment, typically arising in the cutaneous tissue of the previously irradiated field several years after the completion of radiation (14). Although the two subtypes share histological features, existing series report substantial differences in clinical presentation, age at diagnosis, and natural history, with RTI patients typically older at presentation and reported to have shorter survival and a greater propensity for local and distant recurrence compared to Non-RTI patients (16).

RTI angiosarcoma is reported to occur with a mean age at diagnosis of approximately 70 years, affecting up to 0.1% of all irradiated patients, with a median latency of approximately eight years from the completion of radiotherapy (14). Despite its relatively low incidence, RTI angiosarcoma carries a reported five-year overall survival of approximately 20%, reflecting a particularly aggressive clinical course attributed to high rates of local recurrence and metastatic spread (14).

Representative clinical and histopathological features of RTI breast angiosarcoma are shown in Figures 1.3 A and 1.3 B The key clinical differences between Non-RTI and RTI breast angiosarcoma are summarised in Figure 1.3 C.

Table 1 Anatomical distribution of angiosarcoma cases (17).

Anatomical site	**Proportion of cases *(%)***
Soft tissue	*35.7*
Head and neck	*25.8*
Visceral organs	*17.3*
Breast	***13.4***
Bone	*3.0*
Other	*2.4*
Unknown	*2.4*

(A) **(B)**

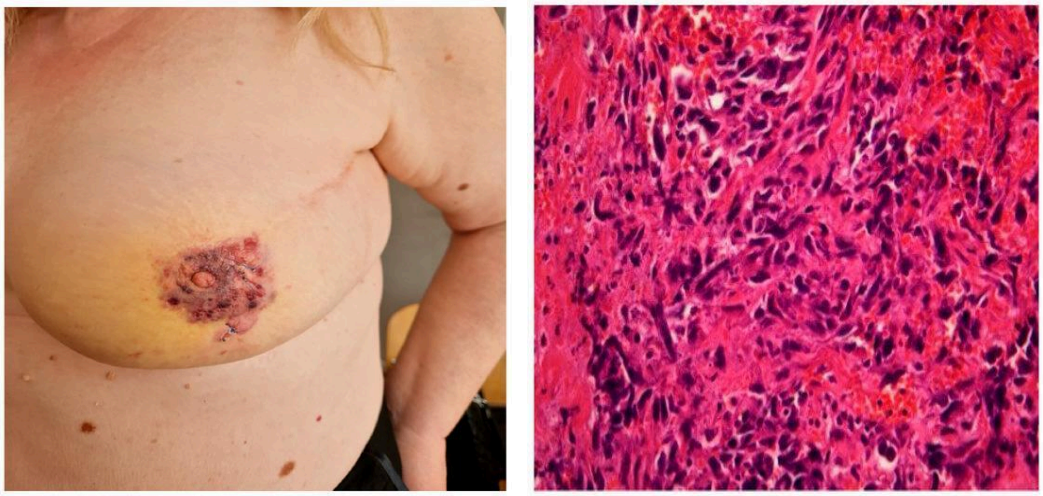


Figure 1.3 A Clinical presentation of RTI breast angiosarcoma (18) **Figure 1.3. B** Histopathological appearance of RTI breast angiosarcoma, haematoxylin and eosin staining, 40x magnification (18).

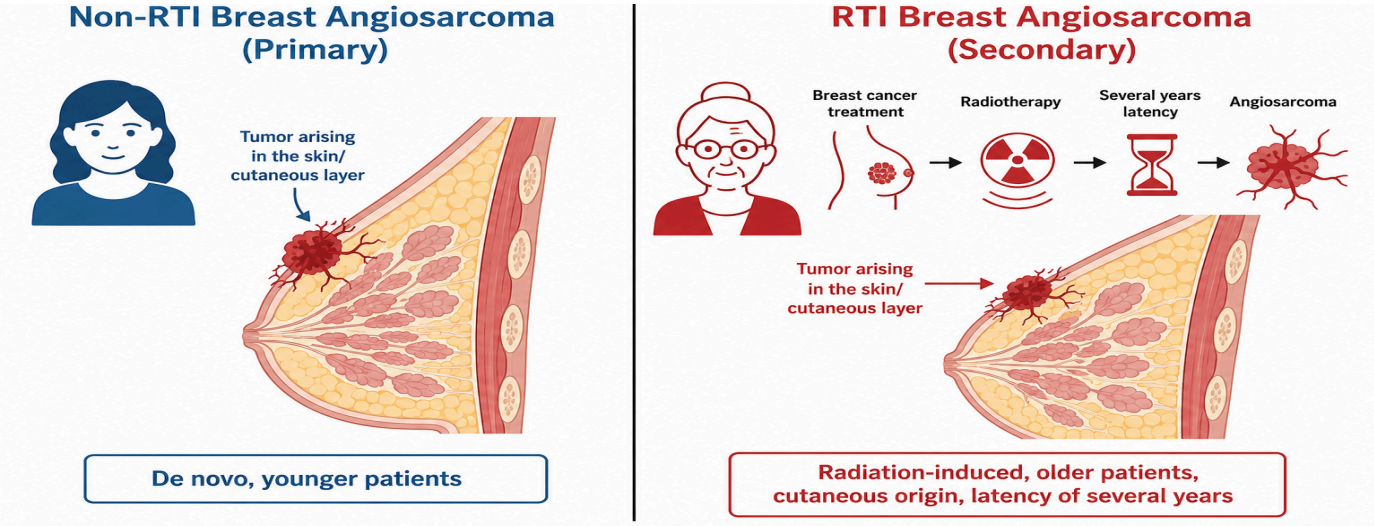


Figure 1.3.C Clinical distinction between Non-RTI and RTI breast angiosarcoma

1.4 Current treatment landscape

Treatment of angiosarcoma remains challenging, and treatment options for advanced disease remain limited (19). For localized disease, radical surgery represents the cornerstone of treatment, often combined with radiotherapy. In the metastatic setting, single-agent doxorubicin and weekly paclitaxel represent the most commonly used first-line systemic therapies, with both demonstrating comparable efficacy (20). Pazopanib, an oral multi-tyrosine kinase inhibitor, received FDA approval in 2012 as the non-cytotoxic agent approved for advanced soft tissue sarcoma, including angiosarcoma, following prior chemotherapy (21). Overall survival has not improved over the past decade despite these treatment options (13).

The biological features of RTI angiosarcoma have prompted interest in immune checkpoint inhibition. In fact, RTI angiosarcoma is characterised by a more T cell-infiltrated tumour microenvironment compared to Non-RTI disease, and by frequent mutations in DNA damage response pathway genes that induce genomic instability and an activated immune microenvironment, suggesting possible susceptibility to immune checkpoint inhibitors (19). The DART trial provided early prospective evidence of immune checkpoint inhibitor activity across angiosarcoma subtypes in a mixed patient population (22). More recently, the CEMangio phase II trial assessed cemiplimab a PD-1 inhibitor, in 18 patients with locally advanced or metastatic secondary angiosarcoma (12 radiotherapy-induced, 3 UV-associated, 3 lymphedema-associated), reporting a best overall response rate of 27.8% and a median overall survival of 13.1 months, concluding that cemiplimab warrants further investigation in this subgroup (19).

1.5 Molecular and immunological landscape

Gene expression profiling of angiosarcoma samples has identified two genomically distinct clusters correlating with anatomical location and radiation exposure (15). MYC amplification, driven by 8q24 chromosomal gain, has been reported in a high proportion of RTI angiosarcoma but not in Non-RTI disease, a finding also reported in the context of breast angiosarcoma specifically (14). In contrast, Non-RTI breast angiosarcoma is more commonly characterised by PI3K pathway mutations (12). RTI angiosarcoma is further

characterised by frequent mutations in genes involved in the DNA damage response (DDR) pathway (19).

The molecular alterations described above arise as a direct consequence of prior radiation exposure, a defining clinical feature distinguishing RTI from Non-RTI breast angiosarcoma. Ionising radiation induces DNA damage through both direct energy deposition and the generation of reactive oxygen species, producing a spectrum of DNA lesions of which double-strand breaks (DSBs) represent the most biologically consequential (23). The recognition and repair of DSBs are governed by the DNA damage response, a coordinated network of kinases including ATM, ATR, and DNA-PKcs (23). In tumours with impaired or overwhelmed DDR machinery, a common feature of radiation-exposed cells, DNA mis repair generates genomic alterations, mutations, and chromosomal aberrations, leading to persistent genomic instability (23). This instability increases the tumour mutational burden (TMB), defined by consensus as the total number of nonsynonymous somatic mutations per megabase of coding DNA sequenced, although reported thresholds and quantification methods vary across assays and cancer types (24). Increased TMB drives the generation of tumour-specific neoantigens, and radiation-induced DNA damage additionally triggers the release of damage-associated molecular patterns and activates the cGAS-STING pathway through cytosolic DNA sensing, stimulating type I interferon production and dendritic cell-mediated antigen presentation (Figure 1.5.1, 23). Together, these processes convert irradiated tumours into immunologically active environments capable of eliciting antitumour T cell responses, providing the mechanistic link between the MYC amplification and DDR mutations described above and the activated immune microenvironment associated with RTI angiosarcoma (19,23).

Sarcomas are generally considered immunologically cold tumours with a poorly immune-infiltrated tumour microenvironment, and angiosarcoma has historically been viewed within this paradigm (19). However, emerging evidence indicates that RTI and Non-RTI angiosarcoma differ substantially in their immune microenvironment composition. Comprehensive immunoprofiling has demonstrated significantly higher densities of *CD3*+, *CD8*+ cytotoxic, and *CD4*+ T helper cells in RTI compared to Non-RTI angiosarcoma, establishing a clear immunological distinction between the two subtypes. In

cutaneous angiosarcoma, transcriptomic studies have further identified the presence of tertiary lymphoid structures in a substantial proportion of cases, with their formation associated with active anti-tumour immune responses and a favourable prognosis, providing additional evidence that a subset of angiosarcomas display an immunologically hot phenotype (25).

Comprehensive transcriptomic characterisation of RTI versus Non-RTI breast angiosarcoma remains limited, and the pathway-level, immunological, and stromal differences between these subtypes are not yet fully understood.

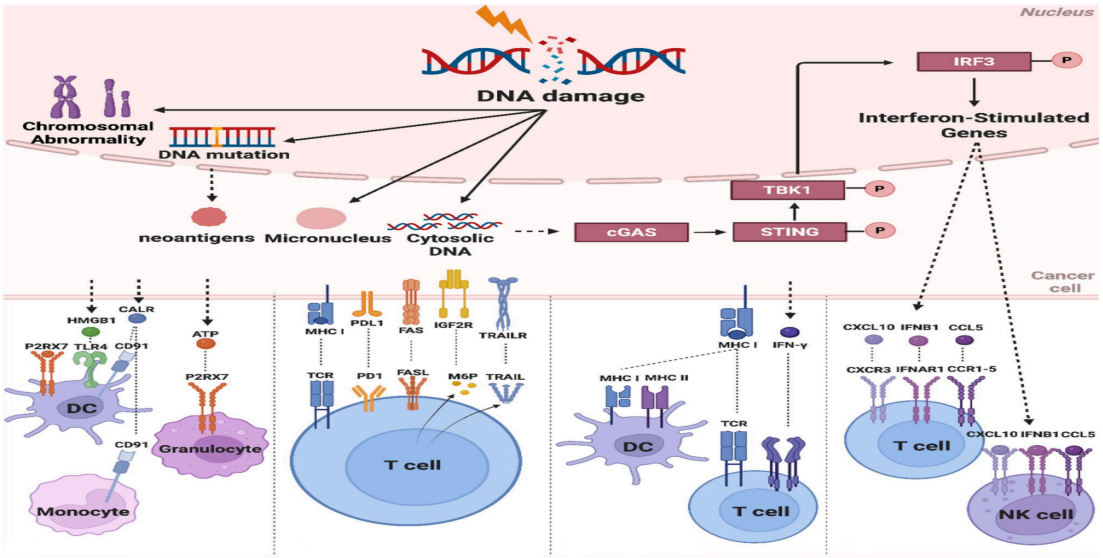


Figure 1.5.1 Immunogenic effects of DNA damage mis repaired in irradiated cancer cells, including immunogenic cell death, immune checkpoint modulation, neoantigen generation, and activation of cytosolic immunity pathways through the cGAS-STING axis (23).

1.6 Cancer-associated fibroblasts and immune exclusion in the tumour microenvironment

Cancer-associated fibroblasts (CAFs) are a key component of the tumour microenvironment with diverse functions, including matrix deposition and remodelling, extensive reciprocal signalling interactions with cancer cells, and crosstalk with infiltrating leukocytes (26). In normal tissue, fibroblasts are the major producers of connective tissue

extracellular matrix (ECM) and play a central role in wound repair, adopting a highly contractile, αSMA-expressing phenotype known as the myofibroblast state. Within tumours, resident fibroblasts undergo activation to become CAFs through a range of stimuli including transforming growth factor-β (TGF-β) signalling, interleukin-1 (IL-1) acting through NF-κB, and IL-6 acting through STAT transcription factors (26). CAFs are identified by markers including αSMA, fibroblast activation protein (FAP), vimentin, and platelet-derived growth factor receptor-α (PDGFRα), though no single marker is entirely specific to this cell population (26). Two functionally distinct CAF subtypes have been described myofibroblastic CAFs (myoCAFs), characterised by a matrix-producing contractile phenotype driven by high TGF-β-mediated αSMA expression, and inflammatory CAFs (iCAFs), located more distally from cancer cells and producing higher levels of IL-6 and other immunomodulatory cytokines (26).

CAFs are perhaps the most effective cell within the tumour microenvironment at depositing and remodelling the ECM, contributing to increased tissue stiffness that alters the migration of infiltrating leukocytes and impairs immune surveillance of the tumour (26). Beyond physical ECM remodelling, the predominant immunological effect of CAFs is immunosuppressive, with IL-6 and TGF-β playing well-established roles in reducing T cell responses. CXCL12 produced by CAFs further suppresses T cell-mediated tumour control, and interference with CXCL12 signalling has been shown to restore effective anti-tumour immune responses (26). Clinical analyses have confirmed an inverse association between CAF abundance and CD8+ T cell infiltration, with FAPhigh fibroblasts specifically correlated with regulatory T cell-mediated immunosuppression and poor clinical outcome (26). Together, these mechanisms establish CAFs as central drivers of an immune-excluded and immune-cold tumour microenvironment.

In the context of angiosarcoma, the distinct biological backgrounds of RTI and Non-RTI disease raise the possibility that CAF biology may play a differential role in shaping their respective immune microenvironments. Unlike RTI angiosarcoma, in which radiation-induced DDR mutations and increased TMB create an immunologically active microenvironment, Non-RTI angiosarcoma arises de novo without these genomic features (13). It is therefore plausible that, in the absence of radiation-driven immune stimulation,

the tumour microenvironment of Non-RTI angiosarcoma may instead be shaped by stromal programs in which CAF-mediated ECM remodelling and immunosuppressive signalling through TGF-β and CXCL12 restrict T cell access to the tumour. This interpretation is supported by transcriptomic pathway analyses comparing RTI and Non-RTI breast angiosarcoma, which have identified differential activation of TGF-β and Wnt signalling pathways (13), both well-established drivers of CAF activation and stromal remodelling (26). Taken together, this convergence of stromal biology and immunosuppressive signalling suggests, as a working hypothesis to be tested in this thesis, that CAF-mediated immune exclusion may represent a key mechanism underlying the immune-cold phenotype of Non-RTI breast angiosarcoma.

1.7 Tumour immune phenotypes: the hot, excluded, and cold classification

The concept of tumour immune phenotyping emerged from the observation that the type, density, and location of immune cells within the tumour site could predict patient survival more accurately than classical anatomical staging systems (27). This led to the development of the Immunscore, a standardised scoring system based on the quantification of CD3+ and CD8+ T cells at both the tumour centre and the invasive margin, classifying tumours across a spectrum from Immunoscore 0, representing absent immune infiltration, to Immunoscore 4, reflecting high immune cell densities at both locations (27). Building on this foundation, a comprehensive four-category classification was proposed, encompassing hot, altered-excluded, altered-immunosuppressed, and cold tumours, providing a practical framework to guide therapeutic decision-making and predict responses to immune checkpoint inhibition (27). The concept of cold and hot tumour immune phenotypes characterises the immune status of tumours and predicts the anticipated response to immune checkpoint blockade based on the type, quantity, and distribution of immune cell infiltration in the tumour microenvironment (28).

Immune-hot, or immune-inflamed, tumours are characterised by high T cell infiltration, increased PD-L1 expression, elevated tumour mutational burden, and the presence of a pre-existing antitumour immune response, features collectively associated with increased responsiveness to immune checkpoint inhibitors (27,28). Checkpoint activation on

tumour-infiltrating T cells, including upregulation of PD-1, CTLA-4, TIM-3, and LAG-3, is a hallmark of the immune-hot phenotype, reflecting functional impairment of otherwise abundant effector T cells (27). At the opposite end of the spectrum, immune-cold, or immune-desert, tumours are characterised by the absence of T cells both within the tumour and at its edges, failed T cell priming, low tumour mutational burden, poor antigen presentation, and low MHC class I expression, resulting in intrinsic insensitivity to immune checkpoint blockade (27). An intermediate phenotype, the immune-excluded tumour, is characterised by the accumulation of T cells at the tumour border without infiltration of the central tumour mass, driven by physical barriers including aberrant tumour vasculature, stromal fibroblasts, collagen-rich extracellular matrix, and TGFβ signalling (27). Both immune-excluded and immune-cold tumours show a deficiency in intrinsic anti-cancer immune capabilities, resulting in limited responses to immune checkpoint blockade (28).

A general schematic illustrating the cellular and cytokine composition that differentiates hot and cold tumour phenotypes Figure 1.6.1 (29).

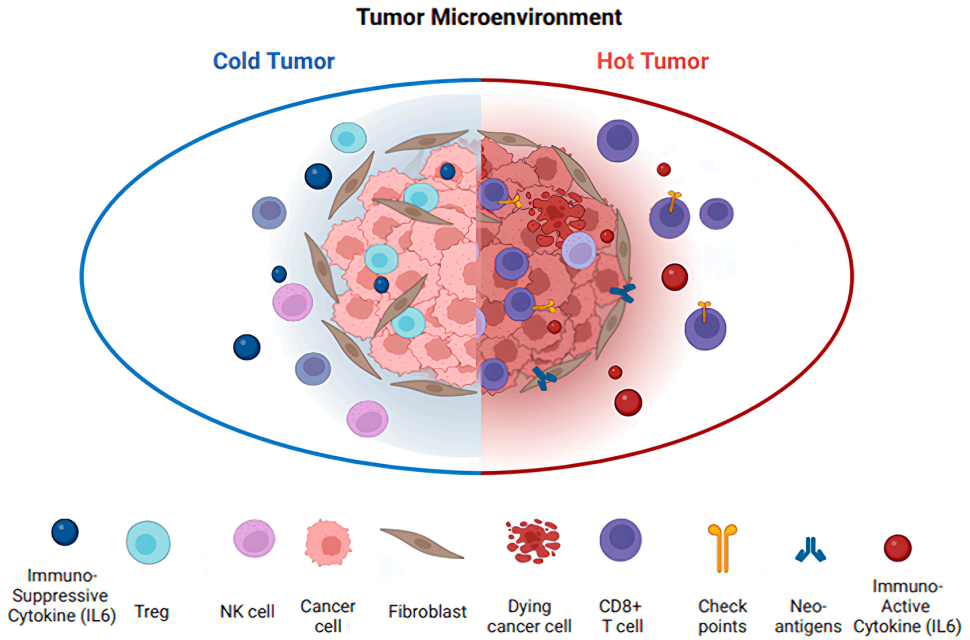


Figure 1.6.1 Cellular and cytokine composition of cold versus hot tumour microenvironments, including immunosuppressive and immune-active cytokines, regulatory and cytotoxic T cells, NK cells, and checkpoint molecules (29).

1.8 Epithelial-mesenchymal transition and tumour immune evasion

Epithelial-mesenchymal transition (EMT) is a biologic process through which polarised epithelial cells undergo biochemical changes that enable them to assume a mesenchymal cell phenotype, characterised by enhanced migratory capacity, invasiveness, elevated resistance to apoptosis, and increased production of extracellular matrix components (30). In cancer, type 3 EMT is initiated by signals from the tumour-associated stroma, including TGFβ, PDGF, EGF, and FGF-2, which activate a series of EMT-inducing transcription factors including Snail, Slug, ZEB1, and Twist (30). These transcription factors orchestrate the loss of epithelial markers such as E-cadherin and the acquisition of mesenchymal markers such as vimentin, enabling carcinoma cells at the invasive front to migrate into the bloodstream and disseminate systemically (30). EMT is now recognised as a continuum rather than a binary switch, with intermediate hybrid epithelial and mesenchymal phenotypes displaying high plasticity and diverse capacities for metastasis and therapy resistance (31).

Beyond its role in invasion and metastasis, EMT is increasingly recognised as a driver of immune evasion in the tumour microenvironment. EMT transcription factors including Snail, ZEB1, and Twist1 produce or attract immunosuppressive cells, including regulatory T cells, tumour-associated macrophages, and myeloid-derived suppressor cells, while simultaneously promoting the expression of immunosuppressive checkpoint molecules, leading to a tumour immunosuppressive microenvironment **(**31). In turn, the resulting immunosuppressive microenvironment produces TGFβ and other cytokines that further activate EMT transcription factors, creating a self-reinforcing feedback loop between EMT and immune evasion that promotes tumour progression and resistance to immune checkpoint inhibitors. Clinical evidence supporting this connection includes the observation that in non-small cell lung cancer, EMT status is strongly associated with increased regulatory T cell infiltration and upregulation of multiple immune checkpoint molecules, while in breast cancer, tumour cells with a mesenchymal gene expression profile express low levels of MHC class I and high levels of PD-L1, accompanied by exhausted CD8+ T cells within their stromal areas (31).

In the context of breast angiosarcoma, transcriptomic pathway analyses comparing Non-RTI and RTI breast angiosarcoma have identified differential enrichment of TGFβ and Wnt signalling pathways (32), both of which are well-established inducers of EMT transcription factors and drivers of the mesenchymal transcriptional program. Given that EMT is associated with reduced MHC class I expression, upregulation of immunosuppressive checkpoint molecules, and recruitment of immunosuppressive cell populations, the EMT-enriched transcriptional program observed in Non-RTI angiosarcoma may represent an additional mechanism through which immune evasion is established and maintained in this subtype, complementing the CAF-mediated stromal immune exclusion described in the preceding section (31).

1.9 Immune checkpoints: biology, regulation, and clinical relevance

Immune checkpoints are inhibitory immunoreceptors that act as gatekeepers of immune responses, maintaining self-tolerance and preventing excessive immune activation (33). In the tumour microenvironment, tumour cells exploit these regulatory pathways to suppress antitumour immunity by downregulating stimulatory signals such as MHC class I surface expression, while simultaneously upregulating inhibitory checkpoint ligands such as PD-L1, thereby dampening T cell-mediated tumour recognition. A number of inhibitory immunoreceptors have been identified and studied in cancer, including PD-1, CTLA-4, LAG-3, TIM-3, TIGIT, and BTLA. Blocking the activation of these inhibitory receptors has been shown to reinvigorate antitumour immune responses, with immune checkpoint blockade therapy frequently achieving more durable clinical responses than conventional chemotherapy or targeted therapies, reflecting the memory properties of the adaptive immune system. The likelihood of response to immune checkpoint inhibition is influenced by several tumour-intrinsic factors including tumour mutational burden, PD-L1 expression levels, interferon signalling pathway activity, and MHC class I expression (33).

CTLA-4 (cytotoxic T-lymphocyte associated protein 4) is one of the most extensively studied immune checkpoint receptors, functioning primarily through competitive inhibition of the co-stimulatory receptor CD28. CTLA-4 binds to its ligands CD80 and CD86 on antigen presenting cells with higher affinity than CD28, thereby blocking the

co-stimulatory signal required for full T cell activation. In addition to this competitive mechanism, CTLA-4-expressing T cells actively reduce CD80 and CD86 surface expression on antigen presenting cells through trans-endocytosis, resulting in decreased CD28 signalling and broader suppression of T cell responses (33). Regulatory T cells, which constitutively express CTLA-4, exploit this trans-endocytosis mechanism to mediate immunosuppression through dendritic cell modulation, a process required for their suppressive function. CTLA-4 further induces expression of indoleamine 2,3-dioxygenase (IDO) in dendritic cells, leading to tryptophan depletion within the tumour microenvironment and additional suppression of effector T cell activity (33).

The clinical validation of CTLA-4 as a therapeutic target was established with ipilimumab, a fully human monoclonal antibody against CTLA-4 approved by the United States Food and Drug Administration in 2011 for the treatment of metastatic melanoma (34). Ipilimumab was the first immune checkpoint inhibitor to receive regulatory approval, demonstrating in a phase III clinical trial that CTLA-4 blockade can improve overall survival compared to vaccine-alone control (34). This approval established a new paradigm in oncology and prompted extensive investigation of CTLA-4 inhibition across multiple tumour types, including in combination with anti-PD-1 therapy to achieve complementary dual checkpoint blockade (33).

TIGIT (T cell immunoreceptor with Ig and ITIM domains) is an emerging immune checkpoint receptor expressed on T and natural killer (NK) cells that shares its ligands CD155 (PVR) and CD112 (PVRL2) with the co-stimulatory receptor CD226 (33). TIGIT binds to these ligands with higher affinity than CD226, suppressing CD226-mediated co-stimulatory signalling through ligand competition, and can additionally bind directly to CD226 in cis to disrupt its homodimer formation and abrogate its co-stimulatory function. Trans ligation of TIGIT delivers inhibitory signals in T and NK cells and also suppresses T cell function indirectly by enhancing IL-10 production from dendritic cells via reverse CD155 signalling. At the intracellular level, TIGIT signalling inhibits PI3K and MAPK pathways as well as NF-κB activation, impairing multiple downstream arms of T cell activation and effector function. The co-expression of *CTLA-4* and *TIGIT* on tumour-infiltrating T cells reflects a state of active immune engagement in which multiple

inhibitory checkpoint pathways are simultaneously upregulated in response to ongoing T cell activity. In tumours characterised by substantial T cell infiltration and an immune-hot microenvironment, the investigation of *CTLA*-4 and *TIGIT* expression is therefore biologically motivated and may identify candidates for immune checkpoint-based therapeutic strategies (19).

1.10 Aims

Comprehensive transcriptomic characterisation of RTI versus Non-RTI breast angiosarcoma remains limited, leaving the pathway-level, immunological, and stromal differences between these two clinically distinct subtypes largely unexplored. The present study aims to address this gap using a discovery-validation design across two independent cohorts. The specific objectives are:

1. To identify differentially expressed genes and enriched biological pathways distinguishing RTI from Non-RTI breast angiosarcoma in an internal discovery cohort, and to independently validate these findings in an external cohort.
2. To characterise the tumour immune microenvironment composition of each subtype, including immune cell abundance, antigen presentation signatures, and immune checkpoint gene expression.
3. To assess stromal biology differences between RTI and Non-RTI tumours, with a focus on cancer-associated fibroblast activity and endothelial cell abundance.
4. To compare clinical and demographic features, including age at diagnosis, between RTI and Non-RTI patients across both cohorts.
5. To contextualise the transcriptomic findings in relation to existing clinical evidence and evaluate their potential implications for patient stratification.
6. To assess the relevance of these findings for therapeutic decision-making, including the biological rationale for immune checkpoint blockade in RTI disease.

CHAPTER II: METHODOLOGY

2.1 Study Design

The present study employed a two-cohort retrospective transcriptomic design to compare radiation-induced (RTI) and non-radiation-induced (Non-RTI) breast angiosarcoma. RTI was defined as angiosarcoma arising within a previously irradiated field following breast cancer treatment. Non-RTI was defined as angiosarcoma diagnosed in the absence of prior radiation history. Detailed inclusion and exclusion criteria for both cohorts are described in Section 2.2. The internal cohort served as the primary discovery cohort and the external cohort served as the independent validation cohort. Both cohorts were analysed separately with no batch correction applied between them.

The internal cohort comprised retrospectively collected archived tumour samples from a multicentre collaboration between Genova, Milan, and Leuven. Gene expression was profiled by bulk RNA sequencing.

The external cohort was obtained from The Angiosarcoma Project - Count Me In (Provisional, April 2025), accessed via cBioPortal for Cancer Genomics (35).

RNA sequencing-derived TPM expression data from the external cohort were downloaded and used for all downstream analyses. This dataset was generated and processed independently by the Count Me In consortium prior to public deposition. An overview of the study design, including patient numbers and the exclusion cascade applied to both cohorts, is presented in Figure 2.1.

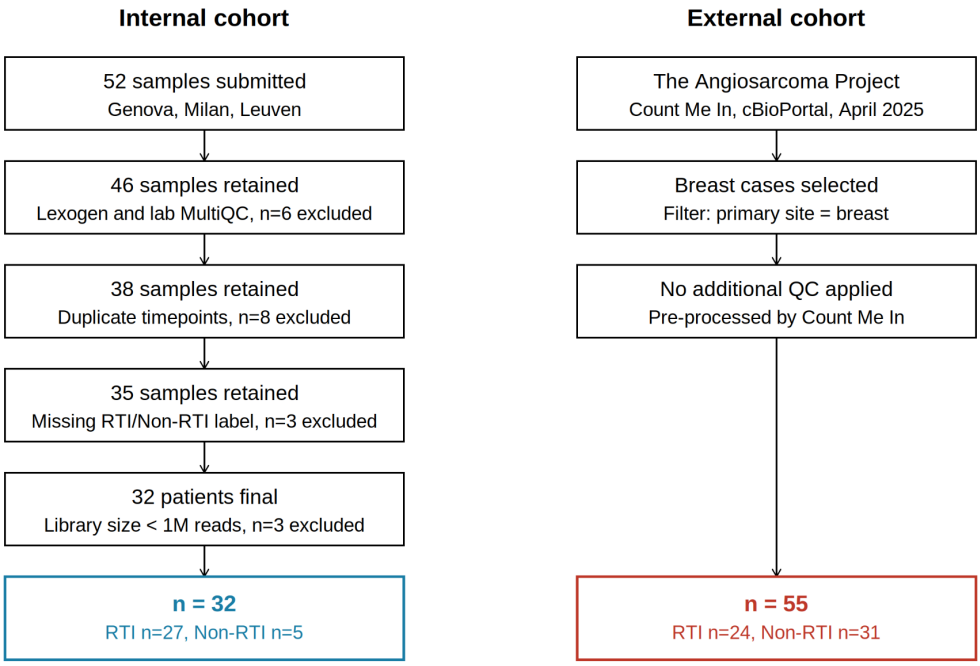


Figure 2.1 Study design overview. Schematic representation of the two independent cohorts used in this study.

2.2 Patient Selection and Data Sources

For the internal cohort, 52 samples from patients with a clinical diagnosis of breast angiosarcoma were submitted for RNA sequencing from three collaborating centres (Genova, Milan, Leuven). Sequencing and initial bioinformatic processing were performed by Lexogen (Experiment BB-3168). Following Lexogen sequencing and processing, raw sequencing data quality was assessed using FastQC and MultiQC (36), evaluating metrics including the percentage of uniquely mapped reads, sequence duplication levels, and adapter content. Based on this quality assessment, 6 samples were excluded from further analysis, yielding 46 samples retained for downstream processing. Of these, 8 samples were further excluded because they represented duplicate timepoints from the same patient, retaining 38 samples. An additional 3 samples were excluded due to missing RTI/Non-RTI clinical annotation, retaining 35 samples. Finally, 3 samples with a library size below 1 million uniquely mapped reads were excluded, yielding a final internal cohort of 32 patients (27 RTI, 5 Non-RTI).

For the external cohort, RNA sequencing-derived expression data were obtained from The Angiosarcoma Project – Count Me In (Provisional, April 2025), accessed via cBioPortal for Cancer Genomics (35,37,38). Samples were filtered to retain only patients with a primary tumour site annotated as breast (MRD_PRIMARY_SITE = breast), yielding a final external cohort of 55 patients (24 RTI, 31 Non-RTI). No additional quality control was applied to the external cohort, as the dataset had been processed and quality-checked by the Count Me In consortium prior to public deposition.

2.3 RNA-seq Data Processing

For the internal cohort, sequencing reads were processed using the Lexogen Data Analysis Pipeline (DAP) version 1.6.0 (processed 2023-11-06). Reads were first assessed for quality using FastQC, followed by adapter and poly(A) tail trimming using Cutadapt (39), which reduced raw 75 bp reads to a trimmed length ranging from approximately 40 to 68 bp depending on the sample. Trimmed reads were aligned to the Homo sapiens GRCh38 reference genome (Ensembl annotation release 107), supplemented with External RNA Controls Consortium (ERCC) and Spike-In RNA Variant (SIRV) synthetic control sequences, using the STAR aligner (40). ERCC and SIRV controls were included in the sequencing library at known concentrations to validate technical performance of the sequencing and quantification process. Alignment and quality metrics across all samples were aggregated using MultiQC (36). Across the 52 sequenced samples, the percentage of uniquely mapped reads ranged from 37.78% to 85.27%. The DAP pipeline generated two parallel gene-level expression outputs: a raw count matrix, used as input for DESeq2-based differential expression analysis.

For the external cohort, gene-level expression data were obtained directly from the Count Me In consortium via cBioPortal as two parallel files: a raw read count matrix, used as input for DESeq2-based differential expression analysis, and a TPM-normalized matrix, used for all downstream gene set scoring and immune deconvolution analyses.

2.4 Normalization

Library size normalization was performed using DESeq2 (v1.42.0) (41) independently for each cohort. DESeq2 was selected over alternative tools such as edgeR (42) due to its superior performance with small and unbalanced sample sizes a critical consideration for the internal cohort, which comprised only 5 Non-RTI samples. DESeq2 employs empirical Bayes shrinkage of dispersion estimates, which improves statistical power and controls false discovery rates in low-replicate settings.

For the internal cohort, size factors were estimated using the poscounts method via estimateSizeFactors(type = "poscounts"), which is appropriate for zero-heavy count matrices generated by 3' mRNA sequencing protocols such as QuantSeq. Prior to size factor estimation, genes with fewer than 10 total counts across all samples were removed to reduce noise from lowly expressed features (MIN_COUNT_FILTER = 10).

For the external cohort, size factors were estimated using the poscounts method via estimateSizeFactors(type = "poscounts"). Genes with fewer than 10 total counts across all samples were removed prior to model fitting. The same normalization approach was applied consistently across both cohorts to ensure comparability of the differential expression results.

2.5 Exploratory Analysis - PCA and Hierarchical Clustering

Principal component analysis (PCA) and hierarchical clustering were performed independently for each cohort to assess sample-level variation, identify potential batch effects, and verify grouping by radiation status prior to differential expression analysis. For the internal cohort, PCA was additionally used to assess whether samples clustered by centre of origin (Genova, Milan, or Leuven), which would indicate the presence of site-specific batch effects requiring correction. Variance stabilizing transformation (VST) was applied to the normalized count matrices using the vst() function with blind=FALSE from DESeq2 (41), which stabilizes variance across the range of mean expression values and is appropriate for downstream exploratory analysis.

For the internal cohort, PCA was performed on the VST-transformed count matrix using the prcomp() function in R (v4.x) (43). The first three principal components were computed and visualized as scatter plots of PC1 vs PC2 and PC2 vs PC3. Samples were coloured by radiation status (RTI or Non-RTI) and separately by centre of origin to evaluate potential site-specific clustering. The percentage of variance explained by each principal component was calculated and displayed on the axis labels. Hierarchical clustering was performed on the same VST-transformed matrix using Euclidean distance and Ward's D2 linkage method via the hclust() function in R (43). Results were visualized as a coloured dendrogram with sample labels coloured by radiation status using the dendextend R package (44).

For the external cohort, PCA was performed on the VST-transformed count matrix using prcomp() (1) in to assess sample-level variation and confirm grouping by radiation status. PC1 vs PC2 and PC2 vs PC3 were visualized with the same colour scheme and formatting for consistency across cohorts. Hierarchical clustering was performed using Euclidean distance and Ward's D2 linkage method (45), visualized as a coloured dendrogram using the dendextend package (44).

2.6 Differential Expression Analysis

For the internal cohort, a volcano plot was generated using ggplot2 (46) with apeglm-shrunken log2 fold change values on the x-axis and −log10 adjusted p-value on the y-axis (47). Genes were coloured by expression category: RTI-upregulated, Non-RTI-upregulated, and not significant. The top 20 most significant genes per direction were labelled using ggrepel. Threshold lines were drawn at $|\log2FC| = 1$ and $padj = 0.05$. For heatmap visualization, the top 25 RTI-upregulated and top 25 Non-RTI-upregulated genes ($padj < 0.05$, ranked by log2FC magnitude) were selected, yielding a matrix of 50 genes. VST-normalized expression values were row-wise z-score scaled and visualized using ComplexHeatmap (48), with columns split by radiation status and hierarchical clustering applied within each group. The color scale ranged from −4 to +4.

For the external cohort, the same procedures were applied using identical parameters. Prior to gene selection for heatmap visualization, a predefined list of 36 epidermal differentiation

complex genes with known high skin-specific expression, as defined by Fagerberg et al. (49), was excluded to minimize confounding from skin tissue contamination present in some cutaneous angiosarcoma samples.

2.7 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) was performed independently for each cohort using the clusterProfiler R package (v4.10.0) (50) with the Hallmark gene set collection from the Molecular Signatures Database (MSigDB v7.5.1) (51), retrieved via the msigdbr package.

For the internal cohort, genes were pre-ranked using the raw Wald statistic derived from DESeq2 standard results. The Wald statistic is defined as:

$W = \beta / SE(\beta)$

Where β is the estimated log2 fold change and $SE(\beta)$ is its standard error. This statistic integrates both the magnitude of differential expression and its statistical precision into a single continuous score. In datasets with small sample sizes such as the internal cohort with only 5 Non-RTI samples, fold change estimates can be highly unstable for lowly expressed genes. The Wald statistic penalizes unreliable estimates through the standard error term, producing a more stable and biologically meaningful ranking than log2 fold change alone (41). Furthermore, the raw Wald statistic was deliberately used in both cohorts rather than the apeglm-shrunken statistic, to ensure that the gene ranking metric was identical across cohorts and that cross-cohort NES values were directly comparable.

Ensembl gene identifiers were converted to HGNC symbols using the org.Hs.eg.db annotation package (52) prior to ranking. Duplicate gene symbols were resolved by retaining the first occurrence. GSEA was run using the following parameters: a minimum gene set size of 15 genes and a maximum of 500 genes. The minimum threshold ensures that pathways with too few genes are excluded, as small gene sets lack sufficient statistical power for reliable enrichment estimation. The maximum threshold excludes overly broad gene sets that are non-specific and may generate spurious enrichment signals.

Benjamini-Hochberg multiple testing correction was applied across all tested pathways (53). A fixed random seed was set prior to analysis (set.seed(42)) to ensure that the permutation-based p-value calculation produced identical results across runs, guaranteeing full reproducibility. All pathways were retained in the output object and filtered at padj less than 0.05 for reporting. Pathways with positive NES were classified as RTI-activated and those with negative NES as Non-RTI-activated.

For the external cohort, the same GSEA procedure was applied. Gene identifiers were already available as HUGO symbols and required no conversion prior to ranking. Identical parameters were used (minGSSize = 15, maxGSSize = 500, BH correction, seed = 42).

2.8 Tumour Microenvironment Estimation Assessment

Tumour microenvironment composition was assessed using the ESTIMATE algorithm (54), which deconvolutes bulk RNA-seq expression data into two per-sample scores: Immune Score, which reflects the degree of immune cell infiltration; Stromal Score, which reflects the abundance of stromal cells. ESTIMATE was selected over alternative deconvolution approaches for this analysis because it specifically quantifies the immune and stromal compartments as continuous scores rather than discrete cell type fractions, making it well-suited for comparing overall tumor microenvironment composition between RTI and Non-RTI groups at the cohort level.

Tumour purity was derived from the ESTIMATEScore using the formula published in the original ESTIMATE paper Yoshihara 2013 (54)

For the internal cohort, expression data were available as natural log (ln(TPM+1)) values. Prior to ESTIMATE input, the data were back-transformed to TPM using the formula TPM = exp(value) - 1, followed by log2(TPM+1) normalization, as ESTIMATE requires log2-transformed expression values calibrated to the Illumina platform scale (54). Ensembl gene identifiers were mapped to HGNC symbols using the org.Hs.eg.db annotation package (52). Duplicate gene symbols were resolved by retaining the maximum expression value across all samples, as maximum aggregation preserves the highest biological signal for

each gene. A GCT-formatted input file was generated manually for all 32 samples together and submitted to the estimateScore() function with Illumina platform settings.

For the external cohort, raw TPM values were log2(TPM+1) transformed prior to analysis. Hugo gene symbols were already available and required no conversion. Duplicate gene symbols were resolved by retaining the mean expression value. A GCT-formatted input file was generated for all 55 samples together and submitted to estimateScore() using identical Illumina platform settings.

Because ESTIMATE scores are computed independently per sample, they reflect the microenvironment composition of each individual tumour rather than being influenced by other samples in the dataset. However, absolute score values are not directly comparable across cohorts due to differences in the number of genes present in each expression matrix (internal: approximately 36,700 genes; external: approximately 54,300 genes), which affects the scale of the scores. All comparisons were therefore performed within each cohort independently using the Wilcoxon rank-sum test (43). Results were visualized as boxplots using ggplot2 (46) and patchwork.

2.9 Immune Cell Deconvolution Analysis

Immune and stromal cell type abundances were estimated using the MCP-counter algorithm (v1.2.0) (55), applied independently to each cohort. MCP-counter quantifies the abundance of 10 cell populations from bulk RNA-seq data using curated marker gene signatures: T cells, CD8 T cells, cytotoxic lymphocytes, B lineage cells, NK cells, monocytic lineage cells, myeloid dendritic cells, neutrophils, endothelial cells, and fibroblasts. MCP-counter was selected over alternative deconvolution methods such as CIBERSORT because it does not require a reference signature matrix to be fitted to the data, making it more robust for datasets where the cellular composition is unknown or highly variable, such as rare tumour types like angiosarcoma. The algorithm outputs a continuous score per cell type per sample, where higher scores reflect greater relative abundance of that cell population.

For the internal cohort, expression data were available as ln(TPM+1) values. Prior to MCP-counter input, Ensembl gene identifiers were mapped to HGNC symbols using the org.Hs.eg.db annotation package (52). Duplicate gene symbols were resolved by retaining the mean expression value across all samples. MCP-counter was applied to the resulting log2(TPM+1) matrix using HUGO symbols as the feature type (featuresType = "HUGO_symbols").

For the external cohort, raw TPM values were log2(TPM+1) transformed prior to analysis. Hugo gene symbols were already available and required no conversion. Duplicate gene symbols were resolved by mean value. MCP-counter was applied to the transformed matrix using identical settings (featuresType = "HUGO_symbols").

Group differences in MCP-counter scores were assessed for each cell type independently using the Wilcoxon rank-sum test (43). Results were visualized as boxplots using ggplot2 (46) and patchwork.

2.10 T Cell Exclusion Analysis

Immune evasion was assessed using the Tumour Immune Dysfunction and Exclusion (TIDE) computational framework (56), applied independently to each cohort. TIDE was selected over alternative immune evasion tools because it models two biologically distinct mechanisms by which tumours escape immune surveillance: T cell dysfunction, which occurs when cytotoxic T cells are present in the tumour but lose their killing ability due to inhibitory signals; and T cell exclusion, which occurs when physical or biological barriers prevent T cells from infiltrating the tumour altogether (56). This distinction is biologically important because it allows differentiation between tumours that are immune-infiltrated but dysfunctional versus tumours that are fundamentally immune-cold. TIDE generates four continuous scores per sample: TIDE score (overall immune evasion), Dysfunction score (reflecting T cell exhaustion within the tumour), Exclusion score (reflecting barriers to T cell infiltration), and CAF score (reflecting cancer-associated fibroblast activity, which is one of the main mediators of T cell exclusion through stromal remodelling). TIDE additionally predicts ICI response probability for each sample (Responder: TRUE or

FALSE) based on its immune profile. TIDE was run locally using the tidepy Python package with cancer type set to "Other" and pretreat set to False, as no prior immunotherapy treatment data were available for these patients.

For the internal cohort, expression data were available as ln(TPM+1) values. Ensembl gene identifiers were mapped to HGNC symbols using the org.Hs.eg.db annotation package (52). Duplicate gene symbols were resolved by retaining the mean expression value. The resulting log2(TPM+1) matrix was mean centred across genes prior to TIDE input, as required by the algorithm. Mean centring removes the overall expression level of each gene across all samples, so that TIDE focuses on relative differences in gene expression between samples rather than absolute expression levels. The centred matrix was written to a tab-separated input file and processed by tidepy.

For the external cohort, raw TPM values were log2(TPM+1) transformed. Hugo gene symbols were already available and required no conversion. Duplicate symbols were resolved by mean value. The same mean-centring transformation was applied prior to TIDE input using identical settings.

Group differences in TIDE scores were assessed for each score independently using the Wilcoxon rank-sum test. Results were visualized as boxplots and ICI response predictions were summarized as bar charts, both generated using ggplot2 (46) and patchwork.

2.11 Antigen Presentation Machinery Scoring

The activity of antigen presentation machinery (APM) and related immune gene programs was quantified using the singscore R package (57), applied independently to each cohort. singscore is a rank-based single-sample gene set scoring method that quantifies the enrichment of a predefined gene set within each individual sample, without requiring a reference dataset or population-level normalization. This is its key advantage over tools such as ssGSEA or GSEA: because singscore ranks genes within each sample independently, the score for one patient is not influenced by the expression values of other patients in the dataset. This makes it particularly appropriate for small and unbalanced cohorts such as the internal cohort, where population-level statistics can be distorted by

group size differences. The algorithm works by first ranking all genes within each sample from lowest to highest expression using the rankGenes() function and then computing an enrichment score for a given gene set using the simpleScore() function, which compares the average rank of genes in the set to what would be expected by chance.

Six immune gene signatures were scored per sample:

Antigen Processing: *TAP1, TAP2, TAPBP, PSMB8, PSMB9, CALR, CANX, B2M, HLA-A, HLA-B, HLA-C*

MHC Class I: *HLA-A, HLA-B, HLA-C, B2M, TAP1, TAP2*

MHC Class II: *HLA-DRA, HLA-DRB1, HLA-DQA1, HLA-DQB1, HLA-DPA1, HLA-DPB1, CD74*

T Cell Inflamed GEP: *CD8A, CD8B, GZMB, PRF1, IFNG, CXCL9, CXCL10, CXCL11, IDO1, TIGIT, CD274, PDCD1LG2* (58)

Immune Checkpoint: *CD274, PDCD1, CTLA4, LAG3, HAVCR2, TIGIT, PDCD1LG2* (22)

TLS Signature: *CCL19, CCL21, CXCL13, SELL, CCR7, LAMP3, BCL6, CXCR5, CD79B, MS4A1* (59)

For the internal cohort, expression data were available as log2(TPM+1) values. Ensembl gene identifiers were mapped to HGNC symbols using the org.Hs.eg.db annotation package (52). Duplicate gene symbols were resolved by retaining the mean expression value. Gene ranking and signature scoring were performed on the resulting log2(TPM+1) matrix using HUGO symbols.

For the external cohort, raw TPM values were log2(TPM+1) transformed prior to analysis. Hugo gene symbols were already available. Duplicate symbols were resolved by mean value. Gene ranking and signature scoring were performed using identical settings.

Group differences in singscore values were assessed for each signature independently using the Wilcoxon rank-sum test with Benjamini-Hochberg correction applied across all six comparisons (53). Results were visualized as boxplots using ggplot2 (46) and patchwork.

2.12 Microsatellite Instability Prediction

Microsatellite instability (MSI) status was predicted from bulk RNA-seq data using MSIsensor-RNA (60), applied independently to each cohort. MSI occurs when the DNA mismatch repair (MMR) system is defective, leading to the accumulation of mutations at short repetitive DNA sequences called microsatellites. Tumours with high MSI (MSI-H) are associated with higher tumour mutational burden and improved response to immune checkpoint inhibitors. MSIsensor-RNA was selected over DNA-based MSI detection methods because it derives MSI status directly from RNA-seq data using a machine learning model trained on TCGA data, eliminating the need for matched DNA sequencing. The algorithm outputs a continuous probability score between 0 and 1 per sample, where higher values indicate greater probability of MSI-H status, allowing quantitative comparison between groups rather than binary classification alone. The TCGA.MSIPopular model was used for prediction.

For the internal cohort, expression data were available as ln(TPM+1) values. Ensembl gene identifiers were mapped to HGNC symbols using the org.Hs.eg.db annotation package (52). Duplicate gene symbols were resolved by retaining the mean expression value. The resulting log2(TPM+1) matrix was then gene-wise z-score normalized across all samples prior to MSIsensor-RNA input, as required by the model. Gene-wise z-score normalization was applied by subtracting the mean expression of each gene and dividing by its standard deviation, transforming expression values into units of standard deviation from the mean. This normalization allows the model to evaluate the relative expression of each gene independently of its absolute expression level, which varies substantially across genes.

$$z = (x - \mu) / \sigma$$

Where:

x = expression value of gene g in sample i

μ = mean expression of gene g across all samples

σ = standard deviation of expression of gene g across all samples

Any genes required by the MSIsensor-RNA model that were absent from the expression matrix were added as zeros to ensure model compatibility. The processed matrix was transposed so that samples appeared as rows and genes as columns, as required by the MSIsensor-RNA input format, and written to a CSV file for Python-based processing. MSIsensor-RNA was run locally using the command-line tool with the detection mode.

For the external cohort, raw TPM values were log2(TPM+1) transformed. Hugo gene symbols were already available. Duplicate symbols were resolved by mean value. The same gene-wise z-score normalization and missing gene imputation procedures were applied prior to MSIsensor-RNA input using identical settings.

Group differences in MSI probability scores were assessed using the Wilcoxon rank-sum test (43). Results were visualized as boxplots using ggplot2 (46).

2.13 Immune Checkpoint Gene Expression

The expression of six key immune checkpoint genes was quantified and compared between RTI and Non-RTI groups independently for each cohort. Immune checkpoint proteins act as inhibitory regulators of T cell activity, and tumour cells exploit these pathways to evade immune destruction (61). Transcriptomic profiling of checkpoint gene expression from bulk RNA-seq data has been applied across cancer types to characterize immune escape patterns. The six genes analysed were: CD274 (encoding PD-L1), PDCD1 (encoding PD-1), CTLA4 (encoding CTLA-4), LAG3 (encoding LAG-3), HAVCR2 (encoding TIM-3), and TIGIT. These genes were selected because they represent the most clinically relevant and therapeutically targeted checkpoint axes currently evaluated in immunotherapy trials (61). Expression values were measured directly from the log2(TPM+1) normalized matrices

without additional transformation, as the purpose of this analysis was to compare expression levels of individual genes between two groups rather than to compute enrichment scores across gene sets.

For the internal cohort, checkpoint genes were extracted from the expression matrix by their Ensembl gene identifiers directly, as the internal dataset uses Ensembl IDs as row names. Expression data were already available as log2(TPM+1) values and required no further transformation.

For the external cohort, checkpoint genes were extracted by Hugo symbol directly from the raw TPM expression matrix. Raw TPM values were log2(TPM+1) transformed prior to analysis.

Group differences in checkpoint gene expression were assessed for each gene independently using the Wilcoxon rank-sum test (43). This non-parametric test was selected because expression data in small cohorts rarely follows a normal distribution, and the Wilcoxon test makes no assumption about the underlying data distribution an important consideration given the limited Non-RTI sample size in the internal cohort. Benjamini-Hochberg correction was applied across all six simultaneous comparisons (53) to control the false discovery rate. Results were visualized as boxplots using ggplot2 (46) and patchwork.

2.14 Drug Sensitivity Prediction

Drug sensitivity was predicted using the oncoPredict R package (62), applied independently to each cohort. oncoPredict predicts the half-maximal inhibitory concentration (IC50) for each drug in each patient tumour by training a ridge regression model on the Genomics of Drug Sensitivity in Cancer 2 (GDSC2) pharmacogenomics dataset **(**62), which comprises drug response data and gene expression profiles from cancer cell lines, and projecting the trained model onto patient tumour expression data. A lower predicted IC50 indicates greater predicted drug sensitivity for that drug in that patient.

oncoPredict was selected because it applies empirical Bayes batch correction via the ComBat algorithm to harmonize gene expression scale between the GDSC2 cell line training data and patient tumour expression data prior to IC50 prediction. This is necessary because cell line and tumour RNA-seq data are generated under different conditions and have systematically different expression distributions. The ComBat method estimates the mean and variance of batch effects from the data itself using a Bayesian framework, then removes them so that the patient expression values are on a comparable scale to the cell line training data before predictions are made. The training data comprised GDSC2 expression profiles (GDSC2_Expr_RMA_Log.rds) and drug response matrices (GDSC2_Res.rds), downloaded from the oncoPredict data repository (osf.io/c6tfx). The calcPhenotype() function was run with empirical Bayes batch correction (batchCorrect = "eb"), power transformation of phenotype values (powerTransformPhenotype = TRUE), removal of the lowest 20% of genes by variance (removeLowVaryingGenes = 0.2), and a minimum sample threshold of 10.

For the internal cohort, expression data were available as ln(TPM+1) values. Ensembl gene identifiers were mapped to HGNC symbols using the org.Hs.eg.db annotation package **(**52). Duplicate gene symbols were resolved by retaining the mean expression value. The resulting log2(TPM+1) matrix was used as input to calcPhenotype().

For the external cohort, raw TPM values were log2(TPM+1) transformed prior to analysis. Hugo gene symbols were already available. Duplicate gene symbols were resolved by mean value. The same calcPhenotype() parameters were applied.

Group differences in predicted IC50 values were assessed per drug using the Wilcoxon rank-sum test (43), with Benjamini-Hochberg correction applied across all drugs tested (53). For visualization, the top 10 drugs predicted to be more sensitive in RTI and the top 10 drugs predicted to be more sensitive in Non-RTI were selected based on uncorrected p-value and displayed as a lollipop plot using ggplot2 (46), where directional significance was expressed as signed −log10(p-value).

2.15 Statistical Analysis

Clinical characteristics were analysed independently for each cohort to describe the study population and assess potential differences between RTI and Non-RTI groups.

For the internal cohort, age at diagnosis was calculated for each patient as the difference in days between the recorded birth date and the angiosarcoma diagnosis date, divided by 365.25 to convert to years. Age distributions were compared between RTI and Non-RTI groups and visualized as boxplots with individual data points overlaid using ggplot2 (9)**.** Disease-specific survival was assessed using the Kaplan-Meier method (43), with the event defined as death due to disease (Last_known_vital_status = "Dead_(disease)"). Overall survival time was calculated in months as the difference between the last contact or death date and the angiosarcoma diagnosis date. Survival curves were estimated using the survfit() function from the survival R package and visualized using the survminer package, with log-rank p-values reported.

For the external cohort, age at diagnosis was available as binned age groups (e.g. 40-49, 50-59) rather than exact dates due to patient privacy requirements of the Count Me In dataset. Bin midpoints were used as numerical approximations for visualization purposes. Age distributions were visualized as boxplots per group using ggplot2 (46). No survival data were available in the external cohort. For RTI patients in the external cohort, treatment latency was calculated as the time in years between the most recent prior radiotherapy to a breast or chest wall location and the angiosarcoma diagnosis date, derived from the timeline treatment file (data_timeline_treatment.txt). Latency was computed as the absolute value of the radiotherapy start date divided by 365.25 and visualized as a lollipop plot per patient using ggplot2 (46). Latency was defined as the interval between the recorded start date of prior radiotherapy and the angiosarcoma diagnosis date. Because the cBioPortal timeline file records treatment starts rather than end date, latency values represent a slight overestimate.

CHAPTER III: RESULTS

3.1 Clinical Characteristics

The internal discovery cohort comprised 32 patients with breast angiosarcoma, of whom 27 were classified as radiation-induced (RTI) and 5 as non-radiation-induced (Non-RTI). All patients were female. RTI patients were significantly older at angiosarcoma diagnosis than Non-RTI patients (median age 73.5 years, range 43.1-89.9 vs median 37.2 years, range 26.3-60.8; Wilcoxon rank-sum test, $p<0.001$; Figure 3.1A).

The external validation cohort comprised 55 patients, of whom 24 were RTI and 31 were Non-RTI. All patients were female. Consistent with the internal cohort, RTI patients were significantly older at diagnosis than Non-RTI patients (median age 63.0 years, range 38.0-77.5 vs median 38.0 years, range 28.0-68.0; Wilcoxon rank-sum test, $p<0.001$; Figure 3.1B). Age was recorded as grouped intervals in the external dataset; midpoint values were used for descriptive analysis. Latency from the recorded start date of prior radiotherapy to angiosarcoma diagnosis was available for 22 of 24 RTI patients, with a median of 6.5 years (range 0.6-13.9 years; Figure 3.1C). Survival data were not available in the external validation cohort and disease-specific survival analysis was therefore not performed. The finding of significantly older age at diagnosis in RTI compared to Non-RTI patients was replicated in the external validation cohort.

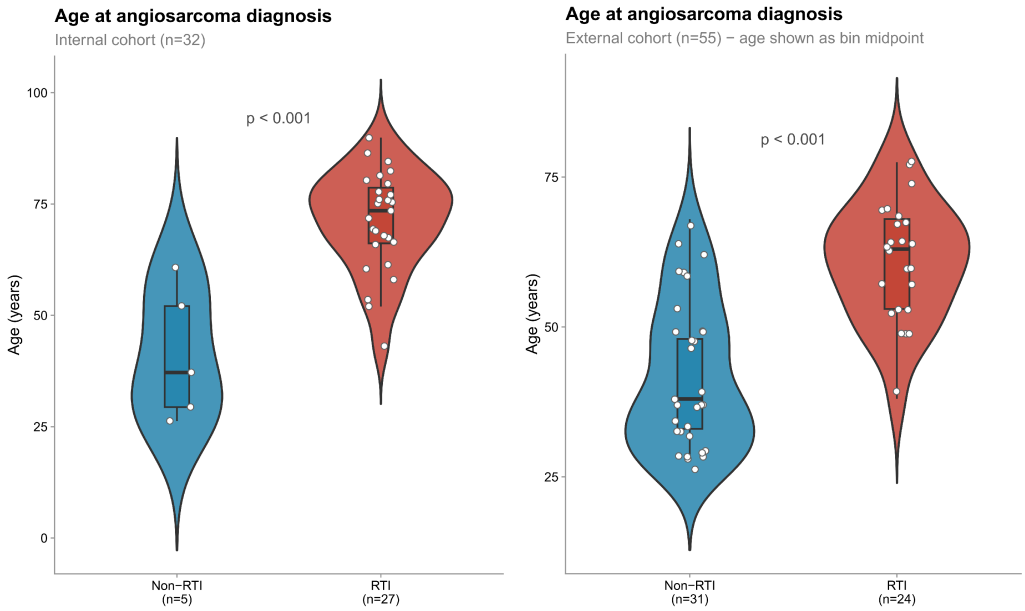


Figure 3.1A-B. Age at angiosarcoma diagnosis in the internal discovery and external validation cohorts. Violin plots showing the distribution of age at angiosarcoma diagnosis stratified by radiation status. Individual patient values are overlaid as jittered points; the central horizontal line represents the median and box edges represent the interquartile range. **(A)** Internal discovery cohort (n=32; RTI n=27, Non-RTI n=5). Exact age at diagnosis was available for all patients. **(B)** External validation cohort (n=55; RTI n=24, Non-RTI n=31). Age was recorded as grouped intervals; midpoint values of each interval were used for visualization.

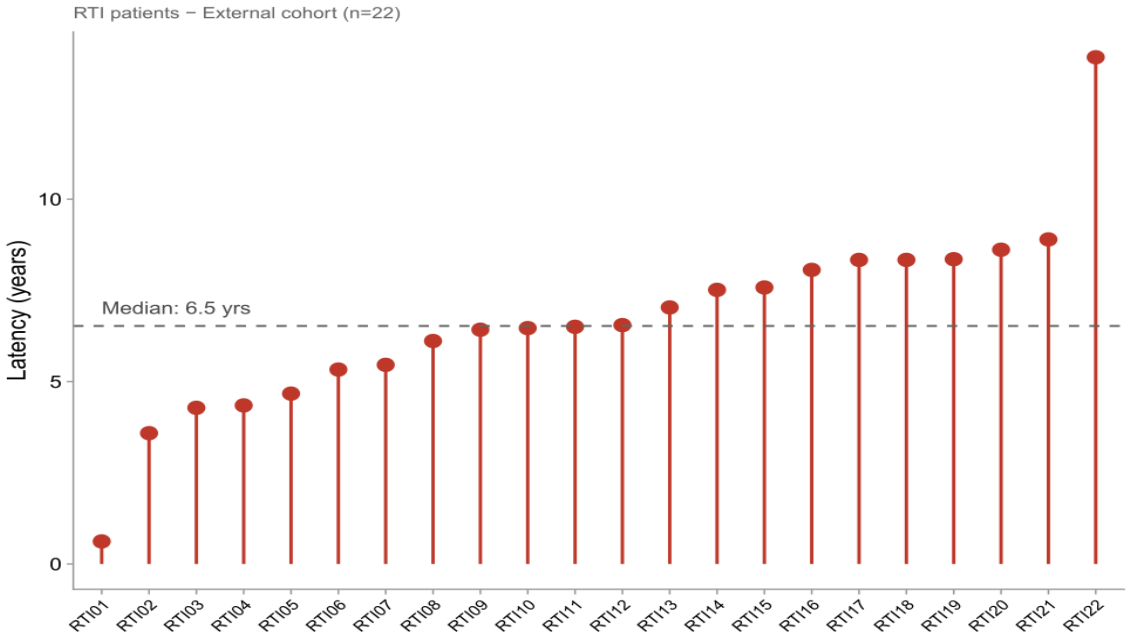


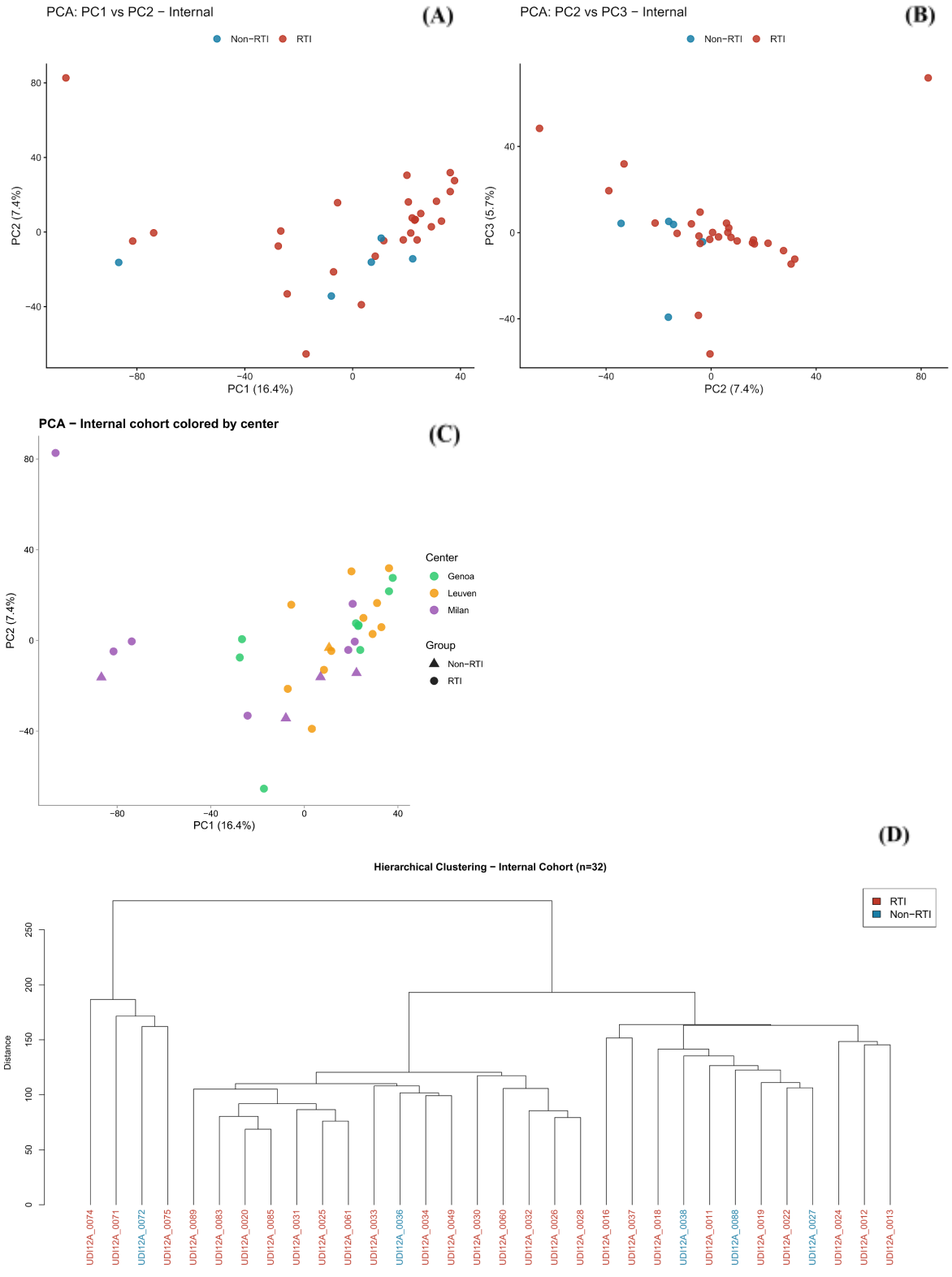
Figure 3.1C Latency from prior radiotherapy to angiosarcoma diagnosis, external cohort RTI patients (n=22). The dashed line indicates the median (6.5 years).

3.2 Quality Control and Batch Effect Assessment

Principal component analysis and hierarchical clustering were performed independently for each cohort to assess sample-level variation and evaluate potential sources of technical bias, including centre of origin, prior to differential expression analysis.

In the internal discovery cohort, PC1 explained 16.4% of total variance, PC2 explained 7.4%, and PC3 explained 5.7%. RTI and Non-RTI samples showed partial separation along PC1 in the PC1 vs PC2 projection (Figure 3.2A). No separation between RTI and Non-RTI samples was observed along PC2 vs PC3 (Figure 3.2B). No clustering by centre of origin was observed when samples were coloured by site (Genoa, Leuven, Milan), indicating the absence of a detectable site-specific batch effect (Figure 3.2C). Hierarchical clustering did not produce clear separation between RTI and Non-RTI samples, with the five Non-RTI samples distributed across RTI-dominated branches (Figure 3.2D). This pattern is consistent with the limited size of the Non-RTI group in the internal cohort.

In the external validation cohort, PC1 explained 34.9% of total variance, PC2 explained 7.4%, and PC3 explained 6.7%. RTI and Non-RTI samples separated along PC1 in the PC1 vs PC2 projection (Figure 3.2E). No separation between RTI and Non-RTI samples was observed along PC2 vs PC3 (Figure 3.2F). Hierarchical clustering revealed partial separation between RTI and Non-RTI samples, with both groups showing a tendency to cluster within distinct branches (Figure 3.2G).


PCA: PC1 vs PC2 − Internal
(A)
Non−RTI
RTI
PC2 (7.4%)
PC1 (16.4%)
PCA: PC2 vs PC3 − Internal
(B)
PC3 (5.7%)
PC2 (7.4%)
PCA − Internal cohort colored by center
(C)
Center
Genoa
Leuven
Milan
Group
Non−RTI
RTI
(D)
Hierarchical Clustering − Internal Cohort (n=32)
Distance
UDI12A_0074
UDI12A_0071
UDI12A_0072
UDI12A_0075
UDI12A_0089
UDI12A_0083
UDI12A_0020
UDI12A_0085
UDI12A_0031
UDI12A_0025
UDI12A_0061
UDI12A_0033
UDI12A_0036
UDI12A_0034
UDI12A_0049
UDI12A_0030
UDI12A_0060
UDI12A_0032
UDI12A_0026
UDI12A_0028
UDI12A_0016
UDI12A_0037
UDI12A_0018
UDI12A_0038
UDI12A_0011
UDI12A_0088
UDI12A_0019
UDI12A_0022
UDI12A_0027
UDI12A_0024
UDI12A_0012
UDI12A_0013

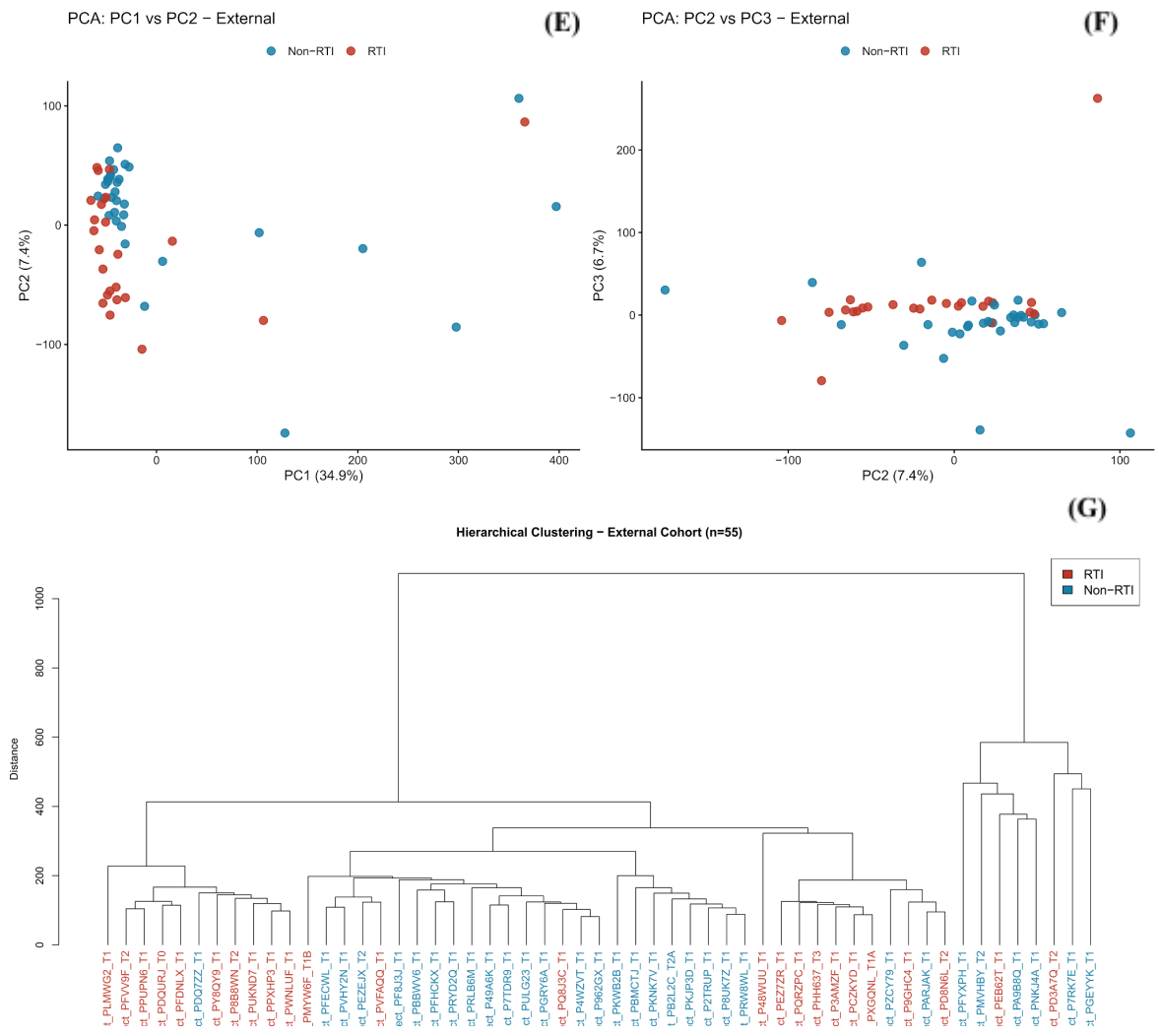


Figure 3.2. Transcriptomic variation and assessment of technical bias in the internal and external cohorts.

Each point in the PCA scatter plots represents one patient sample, plotted by principal component scores. Dendrograms show hierarchical clustering of samples, with branch height reflecting dissimilarity between samples or clusters. Percentages in axis labels indicate the proportion of total variance explained by each principal component.

(A) PCA of the internal cohort (n=32), PC1 (16.4%) vs PC2 (7.4%), coloured by radiation status.

(B) PCA of the internal cohort, PC2 (7.4%) vs PC3 (5.7%), coloured by radiation status.

(C) PCA of the internal cohort coloured by centre of origin (Genoa, Leuven, Milan), performed to assess site-specific batch effects. Point shape indicates radiation status (circle, RTI; triangle, Non-RTI).

(D) Hierarchical clustering dendrogram of the internal cohort (n=32). Sample labels are coloured by radiation status.

(E) PCA of the external cohort (n=55), PC1 (34.9%) vs PC2 (7.4%), coloured by radiation status.

(F) PCA of the external cohort, PC2 (7.4%) vs PC3 (6.7%), coloured by radiation status.

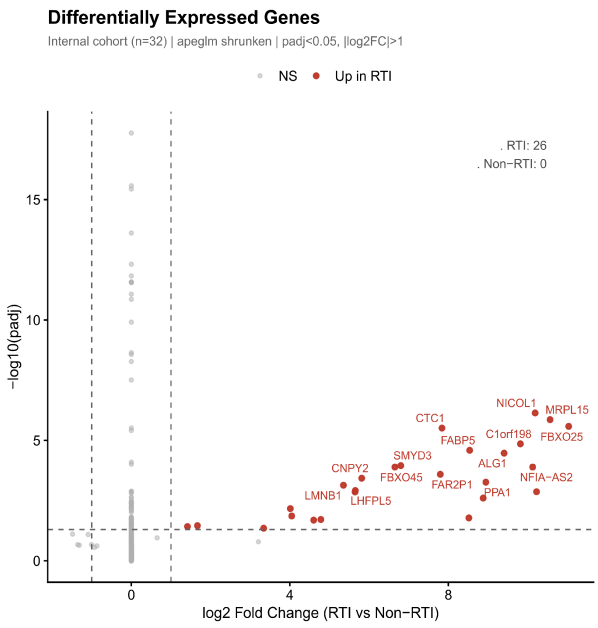
(G) Hierarchical clustering dendrogram of the external cohort (n=55). Sample labels are coloured by radiation status.

3.3 Differential Gene Expression Analysis

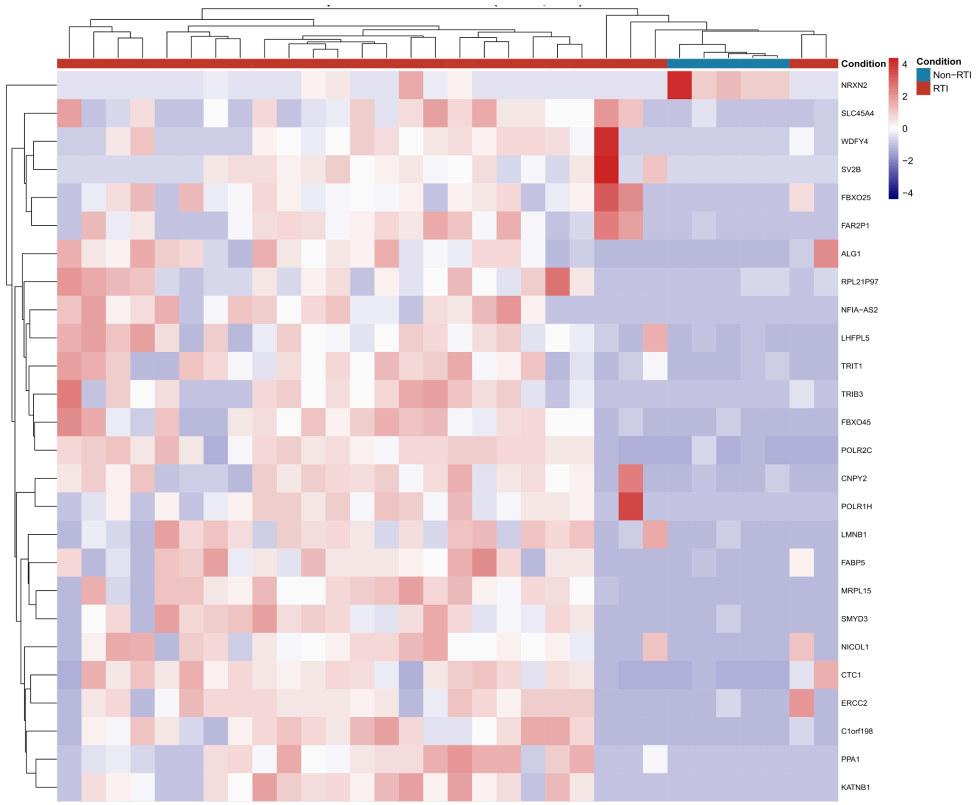
In the internal cohort, 26 genes were differentially expressed between RTI and Non-RTI groups, all of which were upregulated in RTI. No genes met the significance threshold in the Non-RTI-upregulated direction. The most significantly upregulated genes in RTI included *NICOL1* (log2FC=10.3, padj=2.5x10-7), *MRPL15* (log2FC=10.7, padj=5.2x10-7), *FBXO25* (log2FC=11.3, padj=1.0x10-6), *CTC1* (log2FC=7.9, padj=1.3x10-6), and *C1orf198* (log2FC=9.9, padj=5.8x10-6). Results are visualized as a volcano plot (Figure 3.3A) and a heatmap of all 26 DEGs (Figure 3.3B).

In the external cohort, 953 genes were differentially expressed, of which 702 were upregulated in RTI and 251 in Non-RTI. The most significantly RTI-upregulated genes included *C9orf163* (log2FC=2.8, padj=3.0x10-14), *CPNE7* (log2FC=4.1, padj=4.2x10-12), *BPIFC.1* (log2FC=3.0, padj=2.8x10-11), *TIMD4* (log2FC=3.2, padj=3.4x10-9), and *FGFR3* (log2FC=3.5, padj=1.0x10-7). The most significantly Non-RTI-upregulated genes included *PCDH17* (log2FC=-3.0, padj=2.8x10-11), *DPY19L2P1* (log2FC=-4.1, padj=6.4x10-11), *UBTFL6* (log2FC=-3.2, padj=8.6x10-9), *HSPG2* (log2FC=-1.5, padj=1.4x10-8), and *CLVS1* (log2FC=-3.5, padj=3.9x10-7). For heatmap visualization, 36 genes with known high expression in skin tissue were excluded from gene selection to minimize the influence of potential skin contamination in the external dataset. Results are visualized as a volcano plot (Figure 3.3C) and a heatmap of the top 50 DEGs after skin gene exclusion (Figure 3.3D). The external cohort yielded a higher number of DEGs than the internal cohort (953 vs 26), respectively.

(A)


Differentially Expressed Genes
Internal cohort (n=32) | apeglm shrunken | padj<0.05, |log2FC|>1
NS
Up in RTI
. RTI: 26
. Non−RTI: 0
−log10(padj)
log2 Fold Change (RTI vs Non−RTI)
NICOL1
MRPL15
CTC1
FABP5
C1orf198
FBXO25
SMYD3
ALG1
CNPY2
FBXO45
FAR2P1
NFIA−AS2
LMNB1
LHFPL5
PPA1

(B)


Condition
Non−RTI
RTI
NRXN2
SLC45A4
WDFY4
SV2B
FBXO25
FAR2P1
ALG1
RPL21P97
NFIA−AS2
LHFPL5
TRIT1
TRIB3
FBXO45
POLR2C
CNPY2
POLR1H
LMNB1
FABP5
MRPL15
SMYD3
NICOL1
CTC1
ERCC2
C1orf198
PPA1
KATNB1

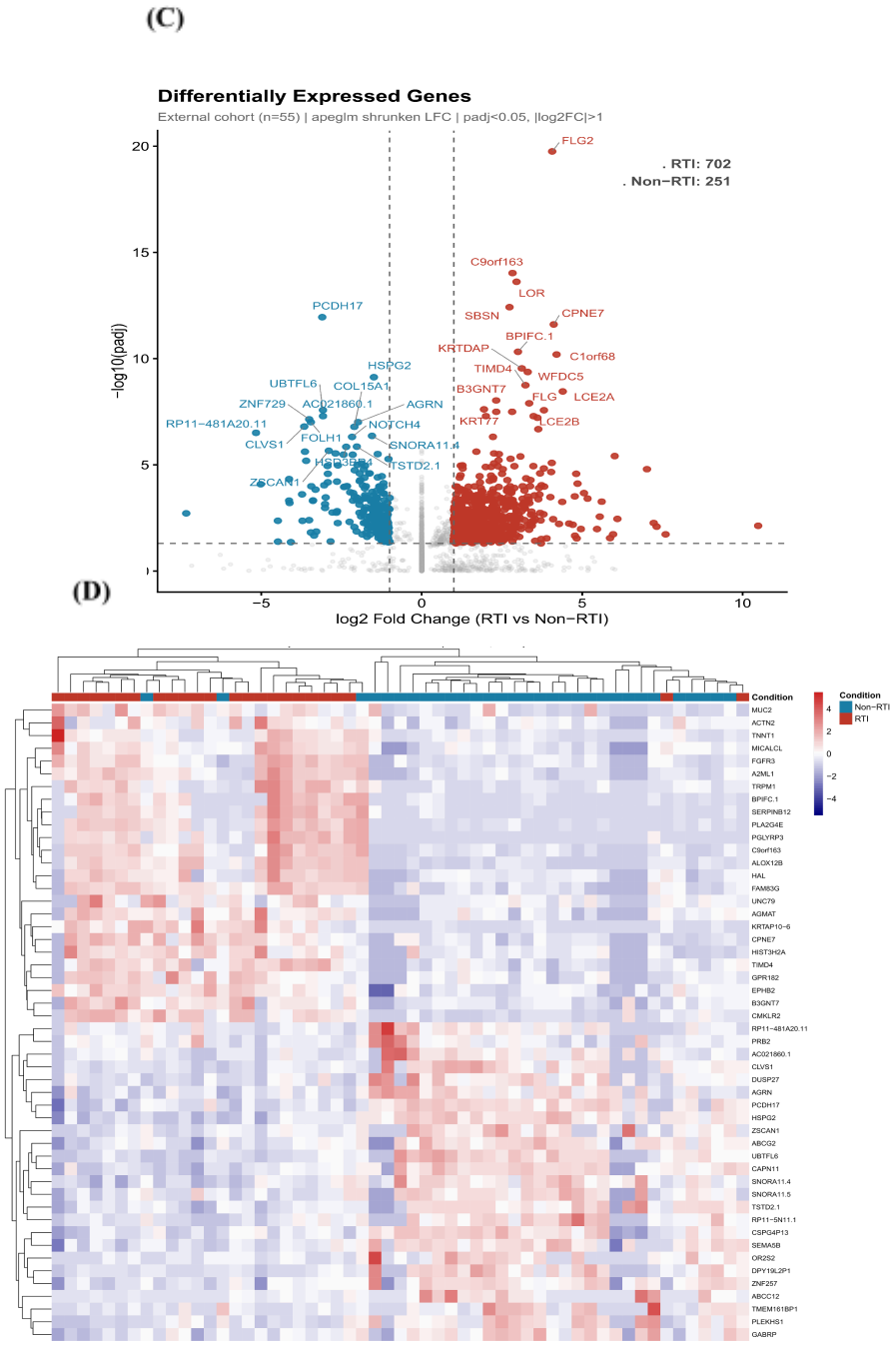


Figure 3.3. Differential gene expression analysis in the internal and external cohorts.
Volcano plots show apeglm-shrunken log2 fold change (x-axis) against -log10(adjusted p-value) (y-axis) for all tested genes. Genes are coloured by significance category: not significant, upregulated in RTI, or upregulated in Non-RTI, based on thresholds of padj < 0.05 and |log2FC| > 1 (dashed lines). The top differentially expressed genes per direction are labelled. Heatmaps show row-wise z-score scaled, VST-normalized expression values for selected DEGs. Columns are split by radiation status; rows (genes) and columns within each group are ordered by hierarchical clustering. Colour scale reflects relative expression (blue, low; red, high). **(A)** Volcano plot, internal cohort (n=32). Twenty-six genes met the significance threshold, all upregulated in RTI. **(B)** Heatmap of all 26 DEGs identified in the internal cohort. **(C)** Volcano plot, external cohort (n=55).953 genes met the significance threshold (702 upregulated in RTI, 251 in Non-RTI). Genes with known high skin expression are shown here but were excluded from heatmap gene selection in panel D.**(D)** Heatmap of the top 50 DEGs in the external cohort, ranked by adjusted p-value

3.4 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis was performed independently for each cohort using the Hallmark gene set collection from MSigDB, with genes pre-ranked by the Wald statistic derived from DESeq2. Pathways were considered replicated if they reached statistical significance in both cohorts (padj < 0.05) with concordant NES direction.

In the internal discovery cohort, 12 Hallmark pathways were significantly enriched (padj < 0.05), of which 10 were RTI-activated and 2 were Non-RTI-activated. The most significantly RTI-activated pathways included MYC Targets V1 (NES=2.093, padj=5.00e-09), MYC Targets V2 (NES=1.835, padj=6.21e-04), E2F Targets (NES=1.823, padj=2.93e-06), Oxidative Phosphorylation (NES=1.788, padj=4.04e-06), and mTORC1 Signalling (NES=1.697, padj=4.56e-05). The two Non-RTI-activated pathways were Epithelial Mesenchymal Transition (NES=-1.943, padj=5.34e-08) and Myogenesis (NES=-1.469, padj=2.80e-03). Results are visualized as a dot plot (Figure 3.4A).

In the external validation cohort, 34 Hallmark pathways were significantly enriched (padj < 0.05), of which 32 were RTI-activated and 2 were Non-RTI-activated. The most significantly RTI-activated pathways included MYC Targets V1 (NES=2.993, padj=7.14e-10), mTORC1 Signalling (NES=2.578, padj=7.14e-10), Oxidative Phosphorylation (NES=2.575, padj=7.14e-10), MYC Targets V2 (NES=2.415, padj=7.14e-10), and E2F Targets (NES=2.372, padj=7.14e-10). The two Non-RTI-activated pathways were Epithelial Mesenchymal Transition (NES=-1.475, padj=2.27e-03) and Myogenesis (NES=-1.545, padj=5.15e-04). Results are visualized as a dot plot (Figure 3.4B). Significant Hallmark pathways were also ranked by NES in a bar plot for each cohort (Figure 3.4C). Pathway concordance between cohorts was highest for Non-RTI-activated pathways, with 100% of significant Non-RTI pathways shared between the internal and external cohorts, compared to 31.2% for RTI-activated pathways (Figure 3.4 D).

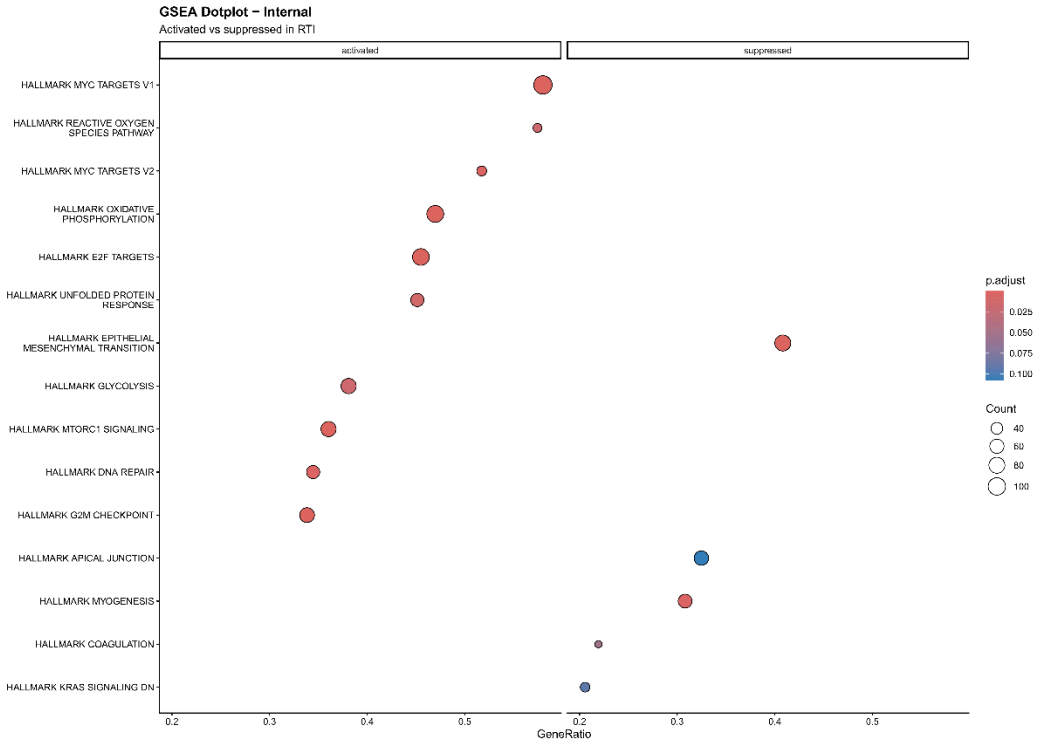


Figure 3.4 (A) Dot plot of significant Hallmark pathways in the internal cohort (n=32). Dot position reflects normalized enrichment score (NES); dot size reflects gene set size; dot colour reflects statistical significance.

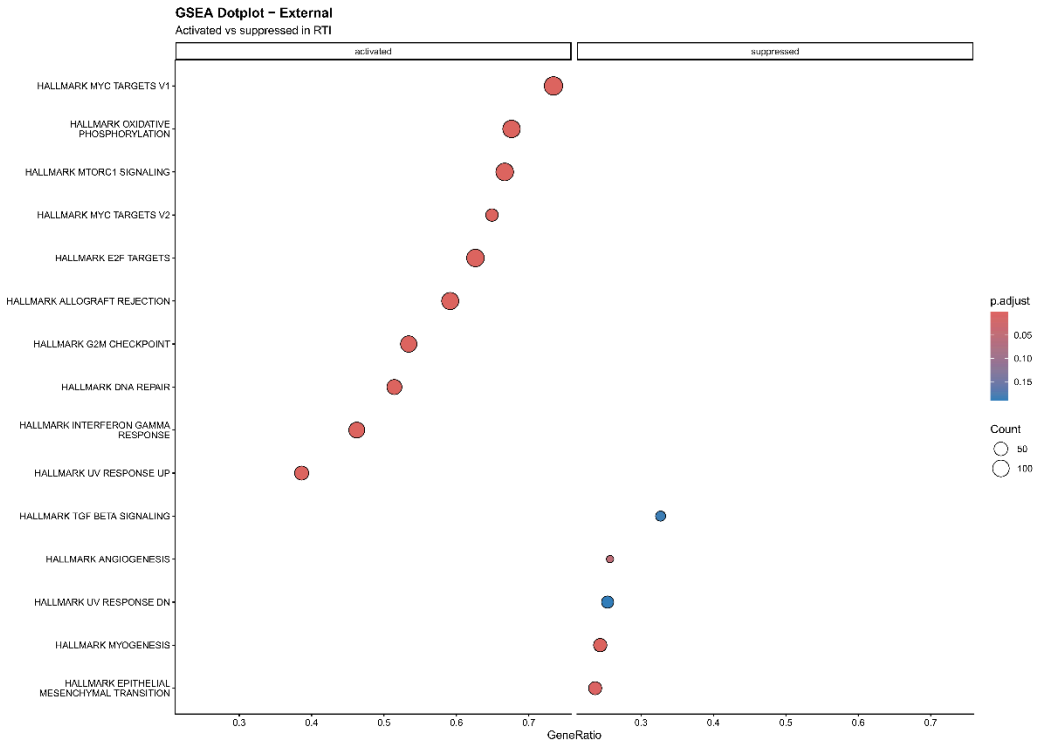


Figure 3.4 (B) Dot plot of significant Hallmark pathways, external cohort (padj < 0.05).

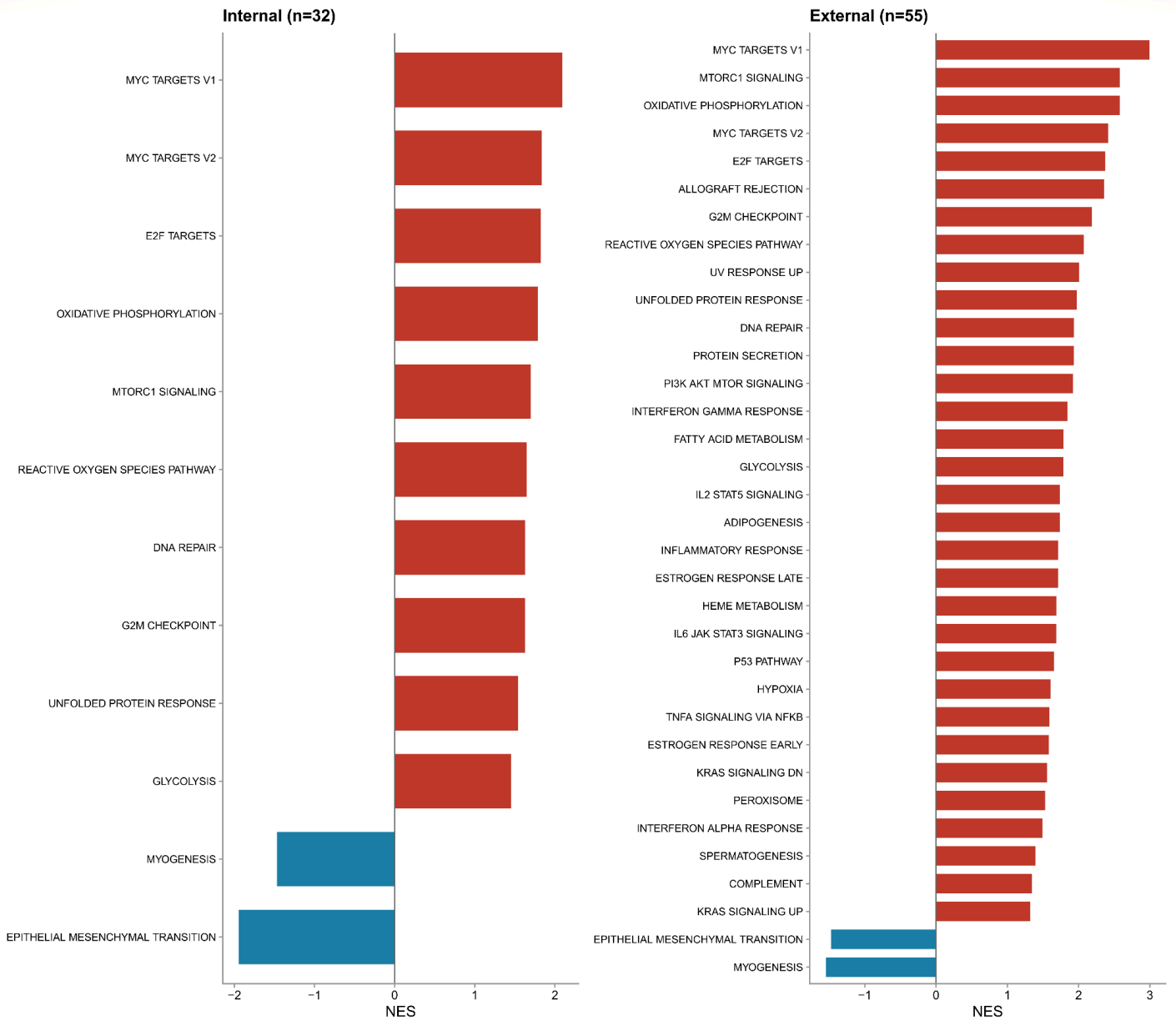


Figure 3.4 (C) Bar plots of significant Hallmark pathways (padj < 0.05) ranked by normalized enrichment score (NES), shown separately for the internal cohort (n=32, left) and external cohort (n=55, right). Red bars indicate RTI-activated pathways (positive NES); blue bars indicate Non-RTI-activated pathways (negative NES).

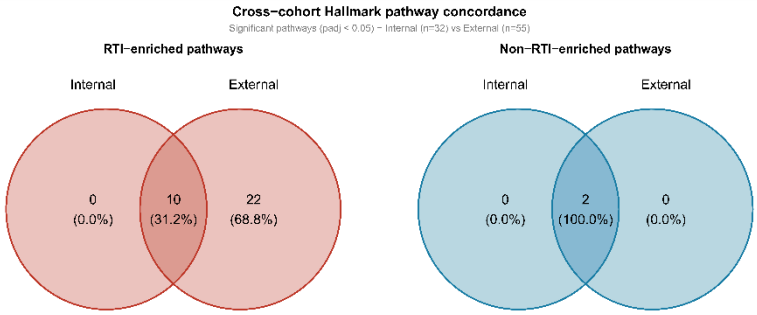


Figure 3.4 (D) Venn diagrams showing cross-cohort concordance of significant Hallmark pathways, separated by enrichment direction. Left, RTI-enriched pathways; right, Non-RTI-enriched pathways. Numbers indicate pathways significant in each cohort alone or in both (overlap); percentages are relative to the total number of unique significant pathways in that direction across both cohorts.

3.5 Tumour Microenvironment Estimation

Tumour microenvironment composition was assessed using the ESTIMATE algorithm, which generates per-sample Immune Score, Stromal Score, and Tumour Purity estimates from bulk RNA-seq data.

In the internal discovery cohort, no statistically significant differences were observed between RTI and Non-RTI groups in Immune Score (median RTI 9959.65 vs Non-RTI 8706.23; Wilcoxon rank-sum test, p=0.335) or Stromal Score (median RTI 8191.21 vs Non-RTI 9574.35; p=0.122). Tumour Purity estimates were not informative in the internal cohort, as the cosine transformation produced values outside the valid range for all samples, likely reflecting the limited dynamic range of expression data in this dataset. Results are visualized as boxplots (Figure 3.5A).

In the external validation cohort, RTI tumours showed significantly higher Immune Score than Non-RTI tumours (median RTI 20880.36 vs Non-RTI 19249.65; Wilcoxon rank-sum test, p=0.004, FDR=0.014). No significant difference was observed in Stromal Score

(median RTI 19279.05 vs Non-RTI 19814.71; p=0.794) or Tumour Purity (median RTI 0.93 vs Non-RTI 0.97; p=0.203). Results are visualized as boxplots (Figure 3.5B). The finding of higher Immune Score in RTI tumours was not replicated in the internal cohort, though the direction of effect was consistent.

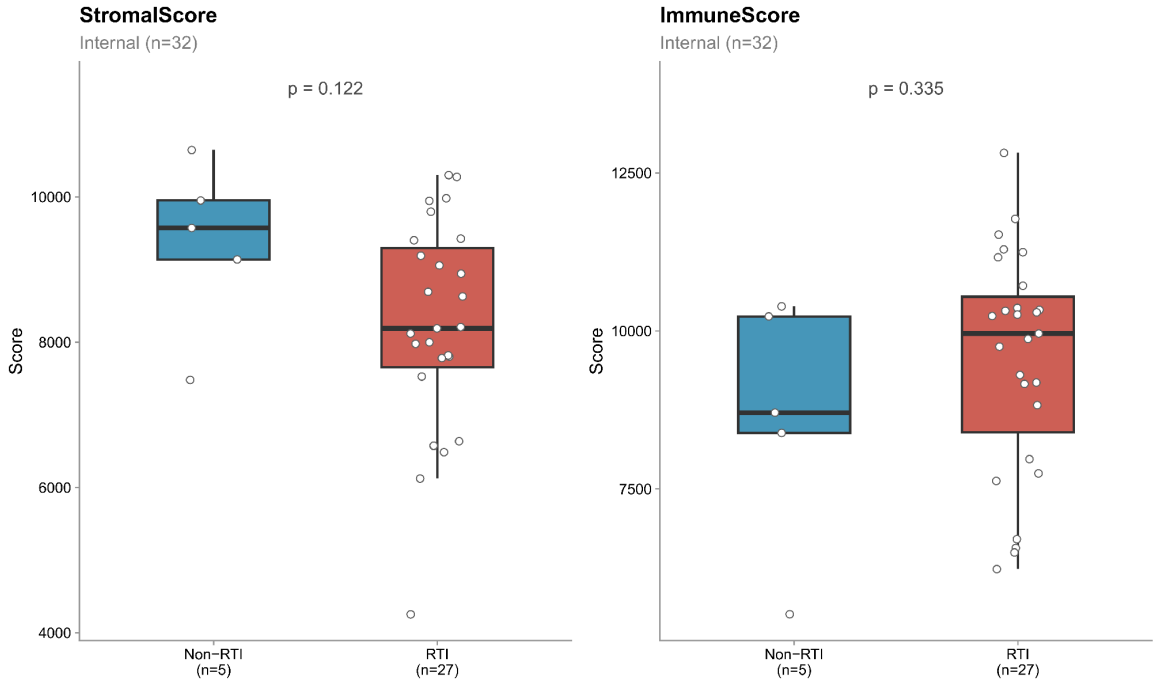


Figure 3.5 (A) Immune Score and Stromal Score per group, internal cohort (n=32).

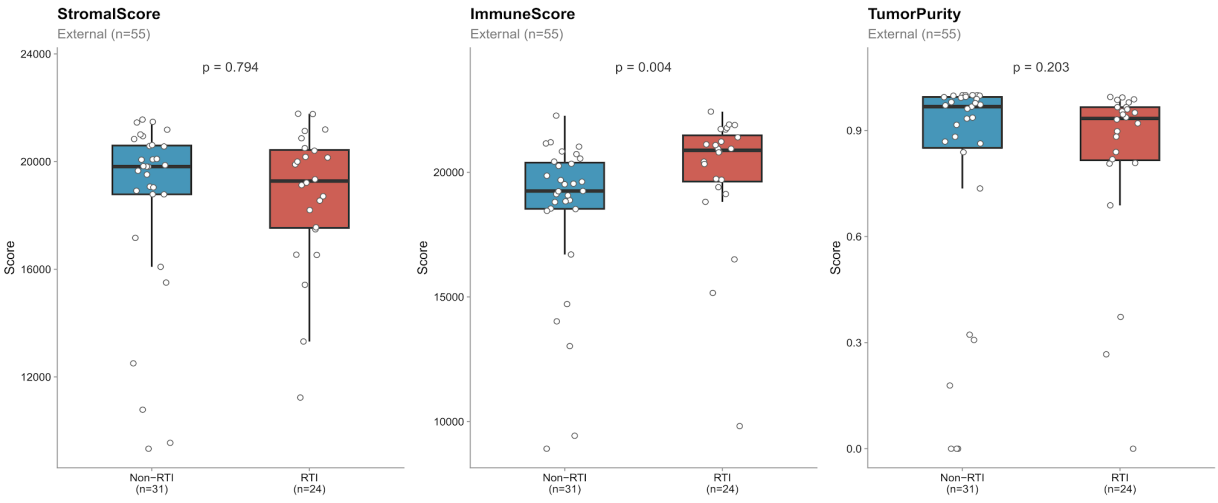


Figure 3.5 (B) Immune Score, Stromal Score, and Tumour Purity per group, external cohort (n=55).

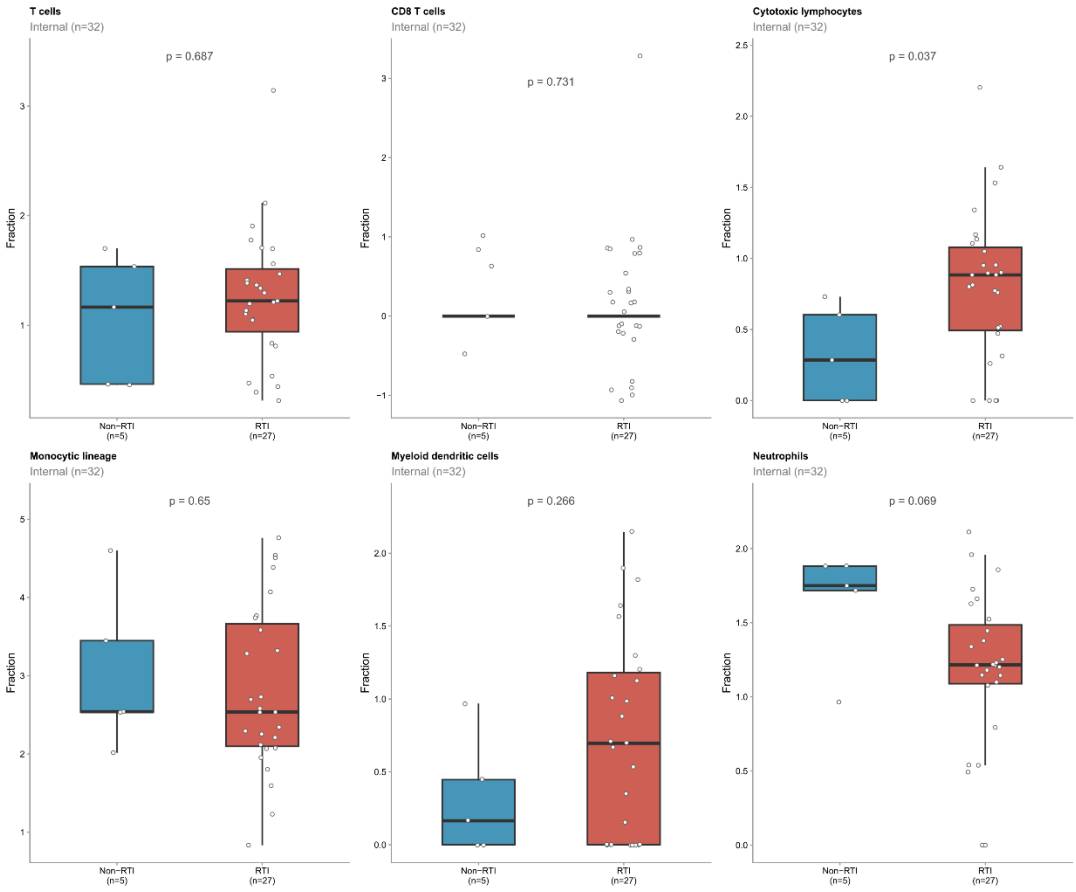
3.6 Immune Cell Deconvolution

Immune and stromal cell type abundances were estimated using MCP-counter, applied independently to each cohort. Ten cell populations were quantified per sample.

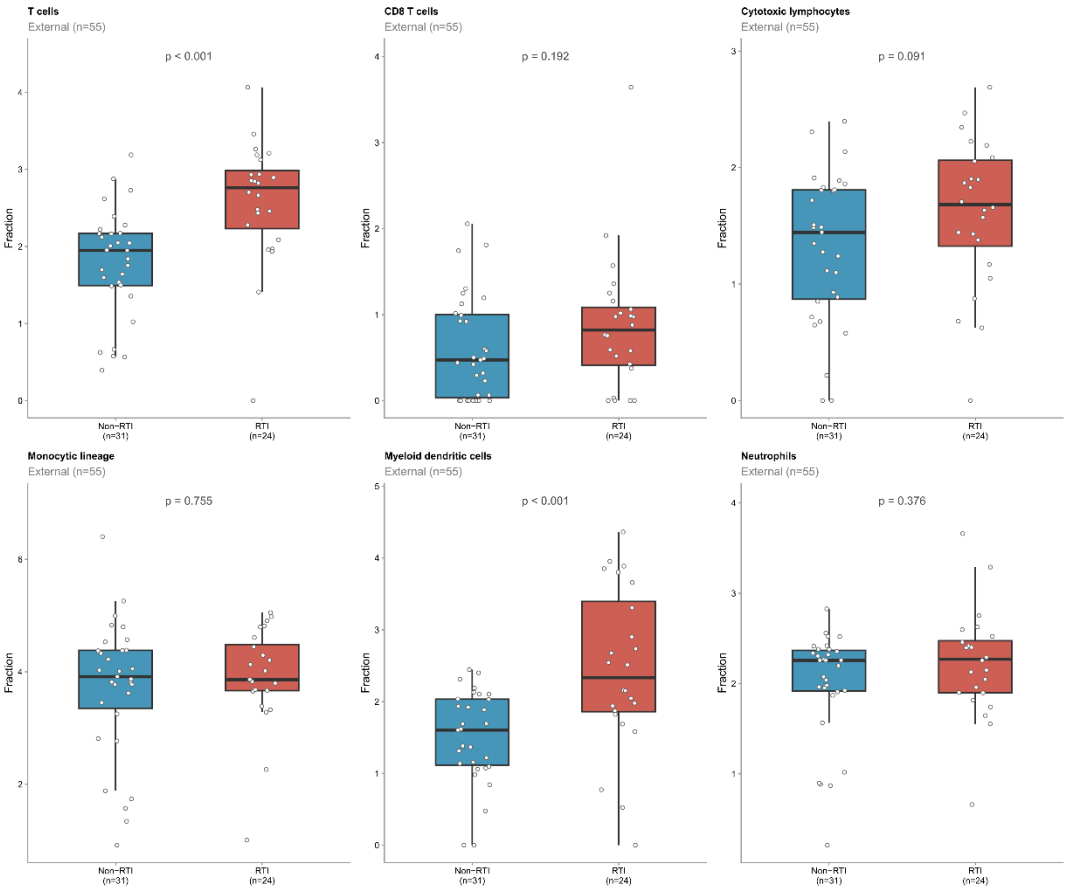
In the internal discovery cohort, Fibroblasts were significantly higher in Non-RTI tumours compared to RTI (median Non-RTI 7.296 vs RTI 6.162; Wilcoxon rank-sum test, p=0.022). Cytotoxic lymphocytes showed a nominally higher score in RTI tumours (median RTI 0.884 vs Non-RTI 0.285; p=0.037), though this did not survive FDR correction (FDR=0.187). No significant differences were observed in the remaining eight cell populations. Results are visualized as boxplots (Figure 3.6A).

In the external validation cohort, three cell populations showed statistically significant differences after FDR correction. T cells were significantly higher in RTI tumours (median RTI 2.763 vs Non-RTI 1.948; Wilcoxon rank-sum test, p=0.0001, FDR=0.001). Myeloid dendritic cells were also significantly higher in RTI (median RTI 2.335 vs Non-RTI 1.604; p=0.001, FDR=0.004). Endothelial cells were significantly higher in Non-RTI tumours (median Non-RTI 5.324 vs RTI 4.873; p=0.01, FDR=0.037). No significant differences were observed in the remaining seven cell populations. Results are visualized as boxplots (Figure 3.6B). The finding of higher fibroblast abundance in Non-RTI tumours was not replicated in the external cohort. The higher T cell and myeloid dendritic cell abundance in RTI tumours observed in the external cohort was not statistically significant in the internal cohort, though the direction of effect was consistent.

(A)


T cells
Internal (n=32)
p = 0.687
CD8 T cells
Internal (n=32)
p = 0.731
Cytotoxic lymphocytes
Internal (n=32)
p = 0.037
Monocytic lineage
Internal (n=32)
p = 0.65
Myeloid dendritic cells
Internal (n=32)
p = 0.266
Neutrophils
Internal (n=32)
p = 0.069
B lineage
Internal (n=32)
p = 0.201
NK cells
Internal (n=32)
p = 0.152
Endothelial cells
Internal (n=32)
p = 0.96
Fibroblasts
Internal (n=32)
p = 0.019
Fraction
Non−RTI (n=5)
RTI (n=27)

(B)



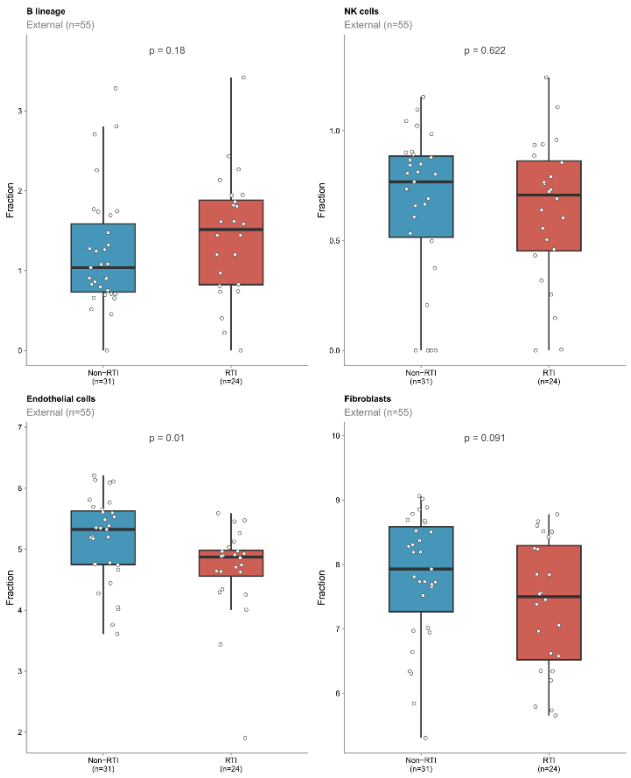


Figure 3.6 MCP-counter immune cell deconvolution scores in the internal and external cohorts.

(A) MCP-counter scores for all 10 cell populations, internal cohort (n=32). Wilcoxon rank-sum test p-values shown.

(B) MCP-counter scores for all 10 cell populations, external cohort (n=55). Wilcoxon rank-sum test p-values shown.

3.7 T Cell Exclusion Analysis

Immune evasion was assessed using the TIDE framework, which quantifies T cell dysfunction, T cell exclusion, cancer-associated fibroblast activity, and overall immune evasion per sample.

In the internal discovery cohort, no significant differences were observed between RTI and Non-RTI groups in TIDE score (median RTI -0.014 vs Non-RTI 0.270; Wilcoxon rank-sum test, p=0.299), Dysfunction score (p=0.233), or Exclusion score (p=0.299). CAF score was significantly higher in Non-RTI tumours compared to RTI (median Non-RTI 0.093 vs RTI -0.022; p=0.006, FDR=0.030), indicating greater cancer-associated fibroblast activity in Non-RTI tumours. Results are visualized as boxplots (Figure 3.7A).

In the external validation cohort, no significant differences were observed in TIDE score (median RTI -0.075 vs Non-RTI 0.092; p=0.172) or Dysfunction score (p=0.570). Exclusion score was nominally higher in Non-RTI tumours (p=0.035) but did not survive FDR correction (FDR=0.087) but still is trend. CAF score was significantly higher in Non-RTI tumours (median Non-RTI 0.063 vs RTI -0.051; p=0.010, FDR=0.049). Results are visualized as boxplots (Figure 3.7B). The finding of significantly higher CAF score in Non-RTI tumours was replicated in the external validation cohort.

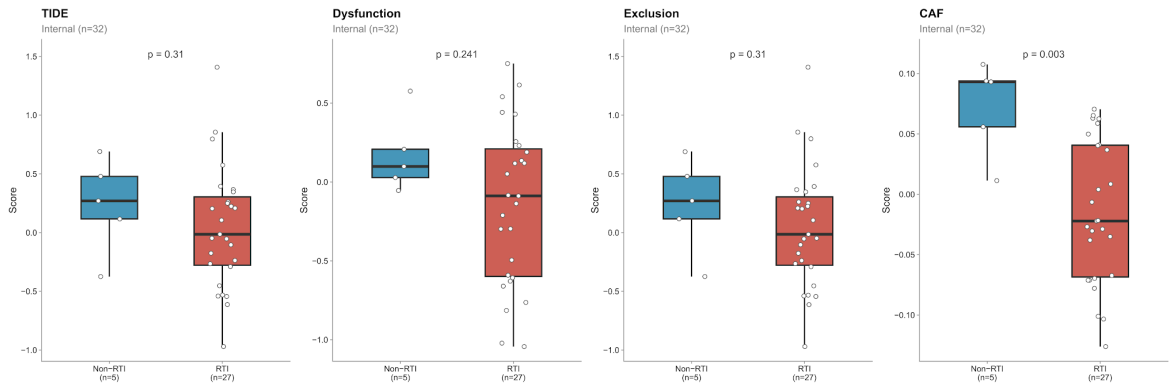


Figure 3.7 (A) TIDE, Dysfunction, Exclusion, and CAF scores per group, internal cohort (n=32). Wilcoxon rank-sum test p-values shown.

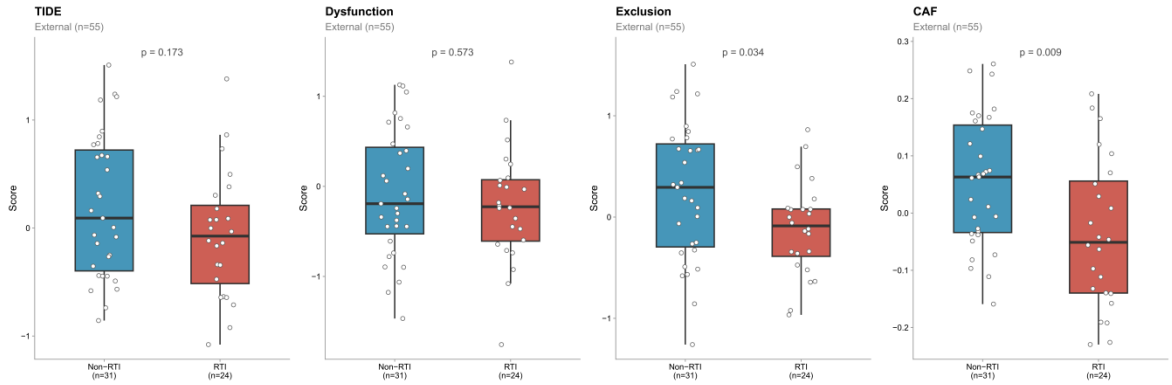


Figure 3.7 (B) TIDE, Dysfunction, Exclusion, and CAF scores per group, external cohort (n=55). Wilcoxon rank-sum test p-values shown.

3.8 Antigen Presentation Machinery Scoring

The activity of six immune gene signatures was quantified per sample using singscore, applied independently to each cohort.

In the internal discovery cohort, no statistically significant differences were observed between RTI and Non-RTI groups for any of the six signatures after FDR correction. T cell inflamed GEP showed a nominally higher score in RTI tumours (median RTI -0.309 vs Non-RTI -0.425; Wilcoxon rank-sum test, p=0.054) that did not survive FDR correction (FDR=0.322). All remaining signatures including Antigen Processing, MHC Class I, MHC

Class II, Immune Checkpoint, and TLS Signature showed no significant differences (all FDR > 0.32). Results are visualized as boxplots (Figure 3.8A).

In the external validation cohort, five of six signatures were significantly higher in RTI tumours after FDR correction. Immune Checkpoint score was the most significantly elevated in RTI (median RTI 0.271 vs Non-RTI 0.236; p=0.0001, FDR=0.0006). TLS Signature score was also significantly higher in RTI (median RTI 0.287 vs Non-RTI 0.155; p=0.0014, FDR=0.0032). T cell inflamed GEP was significantly higher in RTI (median RTI 0.237 vs Non-RTI 0.156; p=0.0016, FDR=0.0032). MHC Class I score was significantly higher in RTI (median RTI 0.475 vs Non-RTI 0.467; p=0.0046, FDR=0.0069). Antigen Processing score was significantly higher in RTI (median RTI 0.468 vs Non-RTI 0.460; p=0.012, FDR=0.015). MHC Class II showed no significant difference (FDR=0.581). Results are visualized as boxplots (Figure 3.8B). None of the significant findings in the external cohort were statistically replicated in the internal cohort, though the direction of effect was consistent across all signatures RTI tumours showed higher scores in both cohorts.

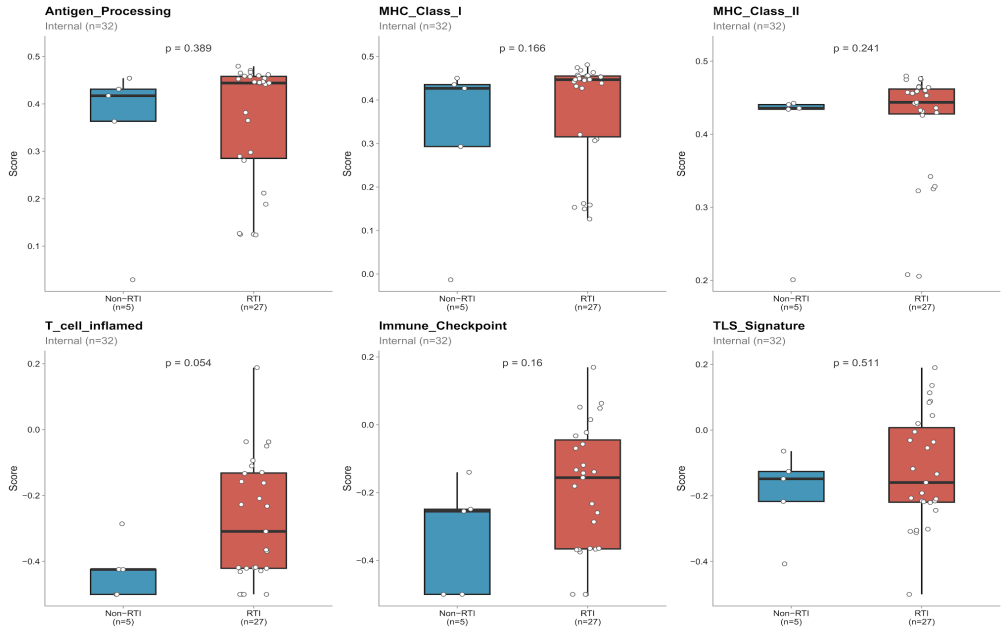


Figure 3.8(A) Singscore values for six immune signatures per group, internal cohort (n=32). Wilcoxon rank-sum test p-values shown.

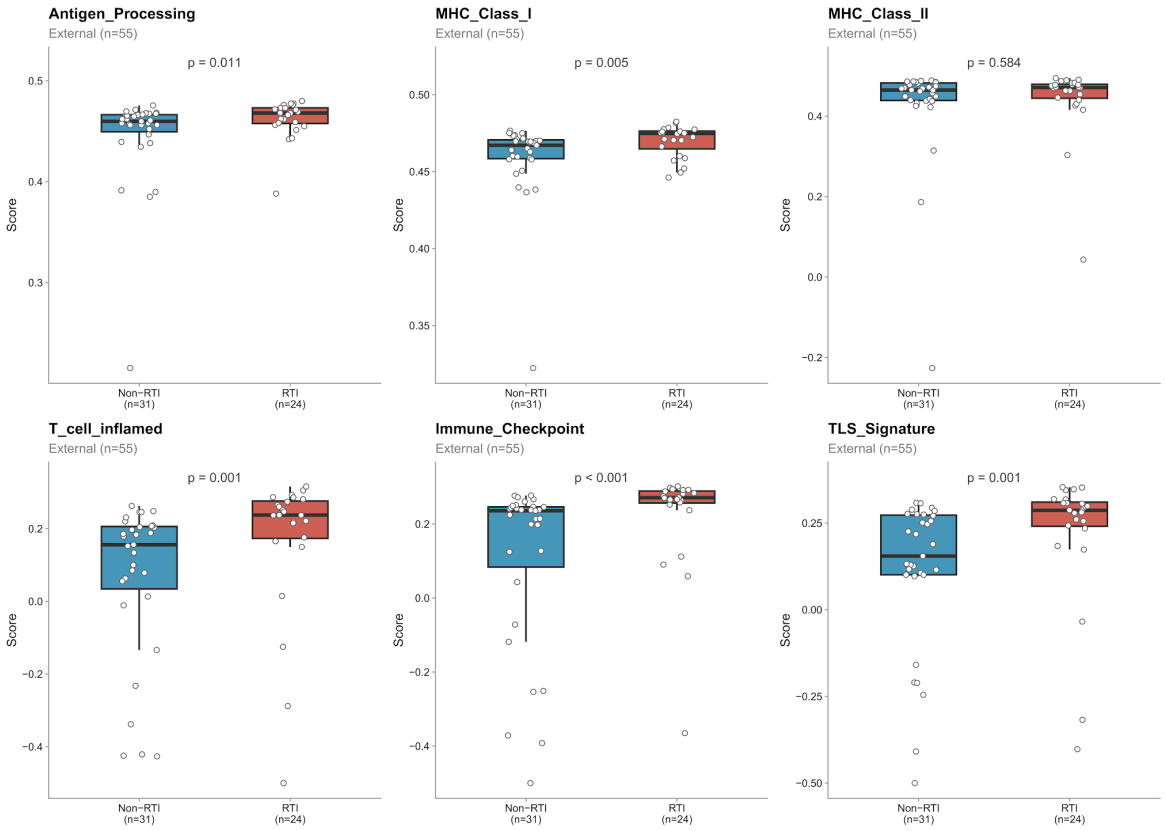


Figure 3.8 (B) Singscore values for six immune signatures per group, external cohort (n=55). Wilcoxon rank-sum test p-values shown.

3.9 Microsatellite Instability Prediction

Microsatellite instability probability was estimated from bulk RNA-seq data using MSIsensor-RNA, applied independently to each cohort. The algorithm outputs a continuous probability score between 0 and 1 per sample, where higher values indicate greater probability of MSI-high status.

In the internal discovery cohort, no significant difference in MSI probability score was observed between RTI and Non-RTI groups (median RTI 0.025 vs Non-RTI 0.043; Wilcoxon rank-sum test, p=0.468). Both groups showed low MSI probability scores overall. Results are visualized as a boxplot (Figure 3.9A).

In the external validation cohort, RTI tumours showed significantly higher MSI probability scores than Non-RTI tumours (median RTI 0.142 vs Non-RTI 0.018; Wilcoxon rank-sum test, p=0.029). Despite reaching statistical significance, the absolute MSI probability scores remained low in both groups, with no sample approaching the conventional MSI-high

threshold. Results are visualized as a boxplot (Figure 3.9B). This finding was not replicated in the internal cohort.

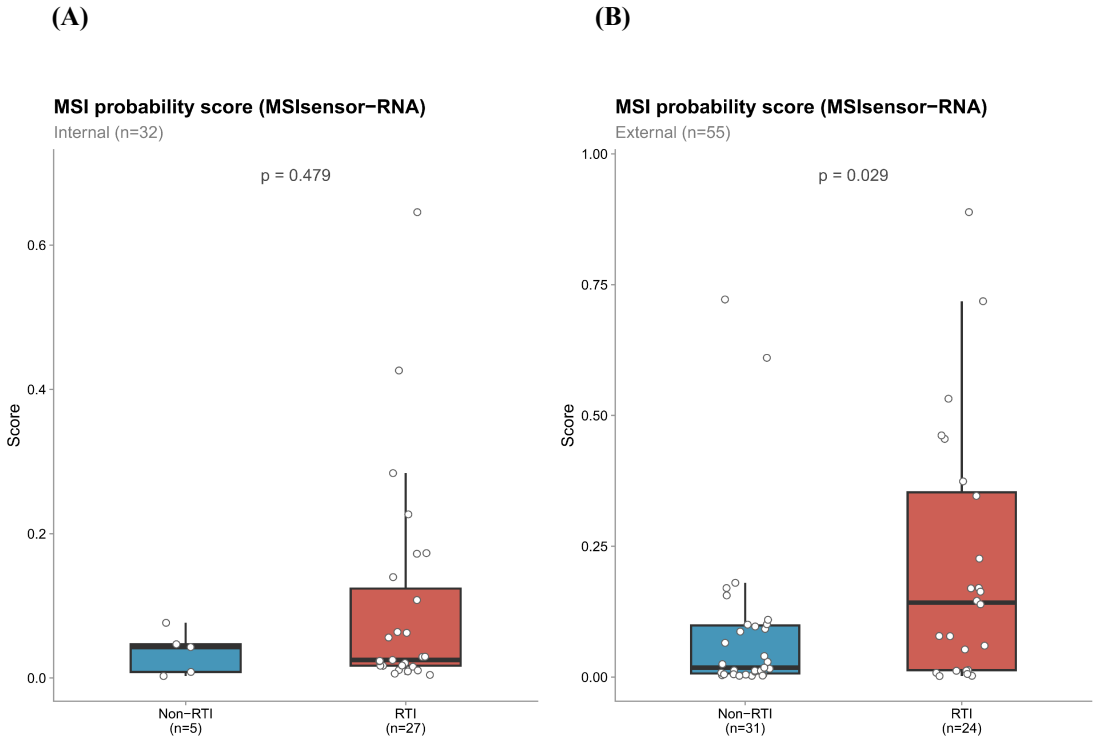


Figure 3.9 (A) MSI probability score per group, internal cohort (n=32). Wilcoxon rank-sum test p-value shown. **(B)** MSI probability score per group, external cohort (n=55). Wilcoxon rank-sum test p-value shown.

3.10 Immune Checkpoint Gene Expression

The expression of six immune checkpoint genes was quantified from log2(TPM+1) normalized matrices and compared between RTI and Non-RTI groups independently.

In the internal discovery cohort, CTLA4 showed nominally higher expression in RTI tumours (median RTI 4.104 vs Non-RTI 0.000; Wilcoxon rank-sum test, p=0.023) that did not survive FDR correction (FDR=0.136). No significant differences were observed for CD274, PDCD1, LAG3, HAVCR2, or TIGIT after FDR correction (all FDR > 0.59). Results are visualized as boxplots (Figure 3.10A).

In the external validation cohort, two checkpoint genes were significantly differentially expressed after FDR correction. CTLA4 was significantly higher in RTI tumours (median RTI 6.346 vs Non-RTI 2.431; Wilcoxon rank-sum test, p=0.0001, FDR=0.0003). TIGIT

was also significantly higher in RTI tumours (median RTI 0.890 vs Non-RTI 0.263; p<0.0001, FDR<0.0001). No significant differences were observed for CD274, PDCD1, LAG3, or HAVCR2 after FDR correction (all FDR > 0.28). Results are visualized as boxplots (Figure 3.10B). The finding of higher CTLA4 expression in RTI tumours showed a consistent direction in both cohorts, reaching statistical significance only in the external cohort. TIGIT was not replicated in the internal cohort.

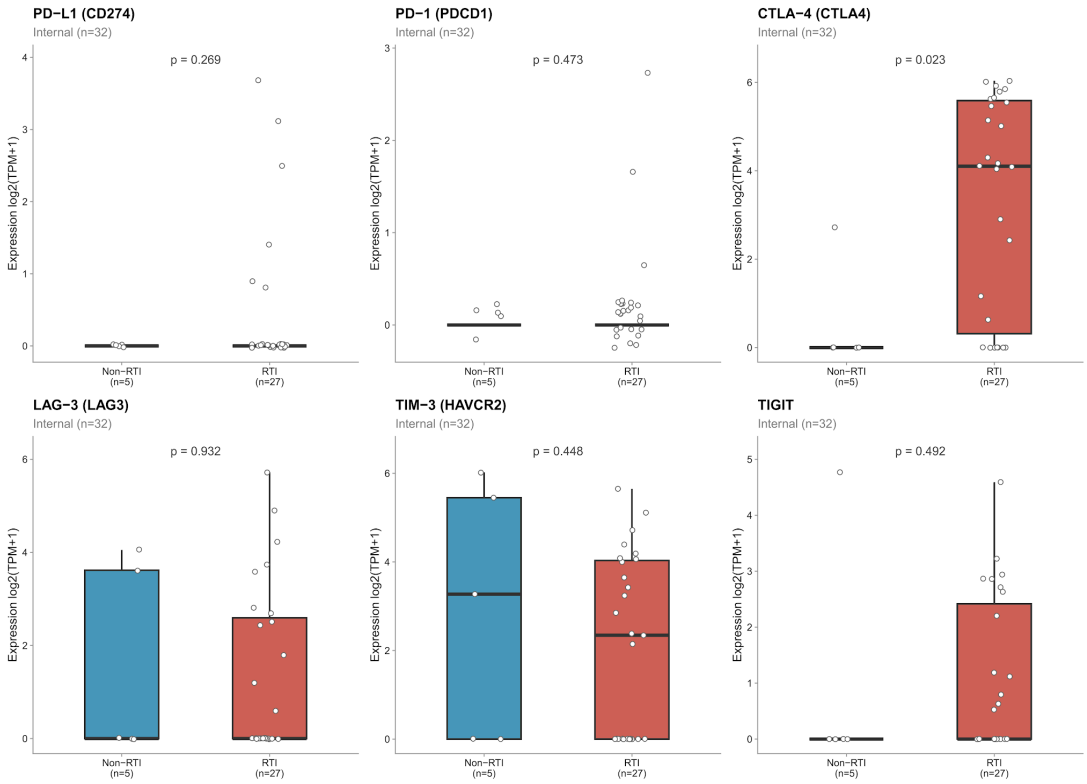


Figure 3.10 (A) Log2(TPM+1) expression of six checkpoint genes per group, internal cohort (n=32). Wilcoxon rank-sum test p-values shown.

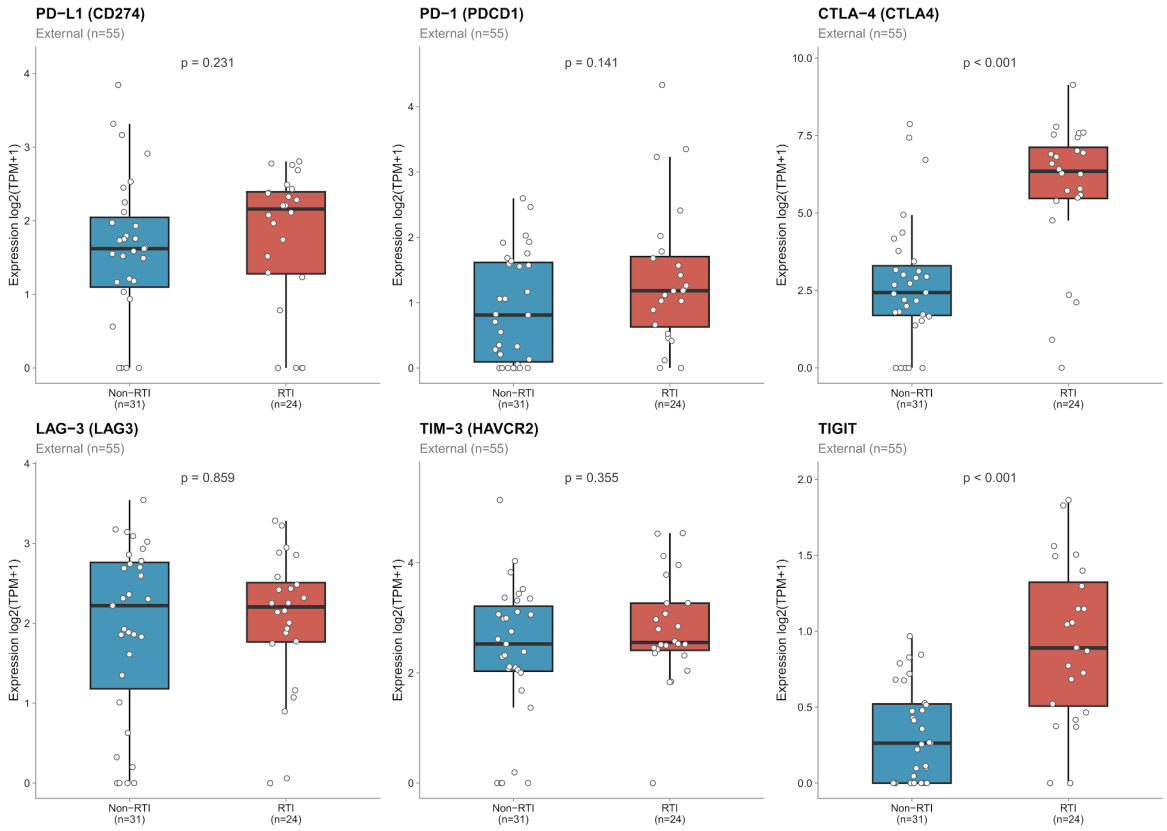


Figure 3.10 (B) Log2(TPM+1) expression of six checkpoint genes per group, external cohort (n=55). Wilcoxon rank-sum test p-values shown.

3.11 Drug Sensitivity Prediction

In the internal discovery cohort, no drug reached statistical significance after FDR correction in either direction. Among RTI-predicted sensitive drugs, Navitoclax (p=0.026, FDR=0.107) and TAF1 inhibitor (p=0.029, FDR=0.107) showed nominal enrichment in RTI tumours. Among Non-RTI-predicted sensitive drugs, JAK inhibitor (p=0.007, FDR=0.095), IGF1R inhibitor (p=0.010, FDR=0.095), NU7441 (p=0.022, FDR=0.107), AZD8186 (p=0.033, FDR=0.107), SB505124 (p=0.038, FDR=0.107), and JQ1 (p=0.043, FDR=0.107) showed nominal enrichment in Non-RTI tumours. None of these findings survived FDR correction. Results are visualized as a lollipop plot (Figure 3.11A).

In the external validation cohort, ten drugs were predicted to have significantly greater sensitivity in RTI tumours after FDR correction. The most significantly RTI-sensitive drugs included Gefitinib ($p<0.0001$, FDR=0.003), Erlotinib ($p<0.0001$, FDR=0.001), IRAK4 inhibitor (p=0.0001, FDR=0.004), Fludarabine (p=0.0001, FDR=0.004), Dihydrorotenone (p=0.0001, FDR=0.003), and Ulixertinib (p=0.0001, FDR=0.004). One drug reached significance for Non-RTI predicted sensitivity: Nutlin-3a (p=0.003, FDR=0.028). Results are visualized as a lollipop plot (Figure 3.11B). None of the significant external findings were replicated in the internal cohort.

Tamoxifen showed consistent predicted sensitivity in RTI tumours across both cohorts, reaching statistical significance in the external cohort (p=0.0007, FDR=0.013) with a concordant nominal trend in the internal cohort (p=0.108). JAK inhibitor showed consistent predicted sensitivity in Non-RTI tumours across both cohorts, though neither reached FDR correction in either dataset.

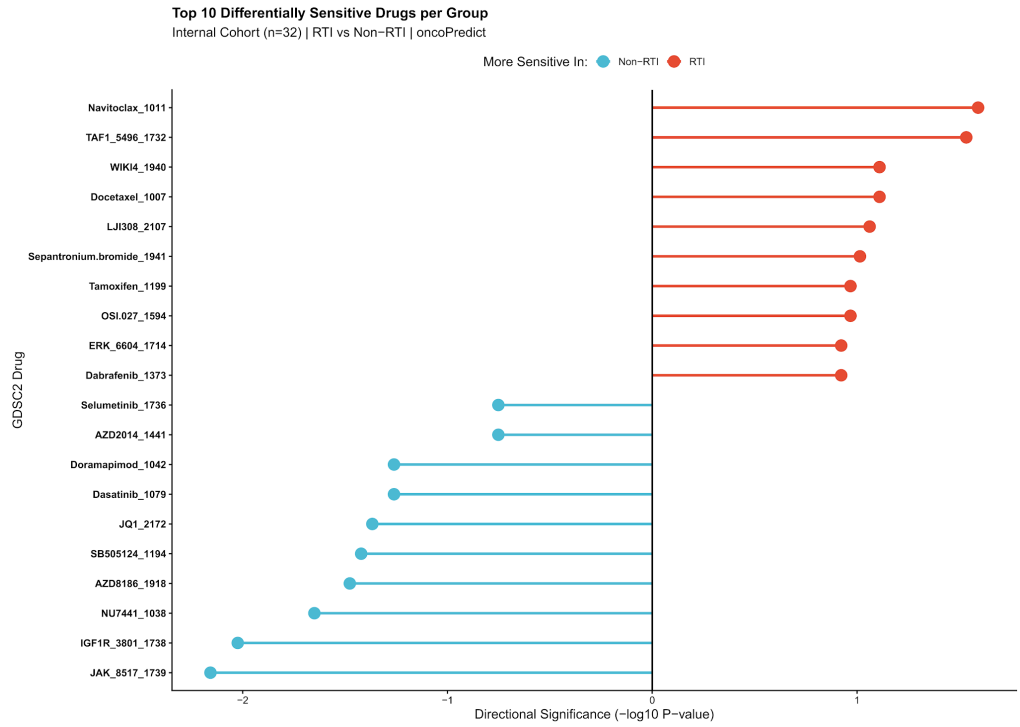


Figure 3.11 (A) Top 10 differentially sensitive drugs per direction, internal cohort (n=32). Directional significance expressed as signed -log10(p-value). No drugs survived FDR correction.

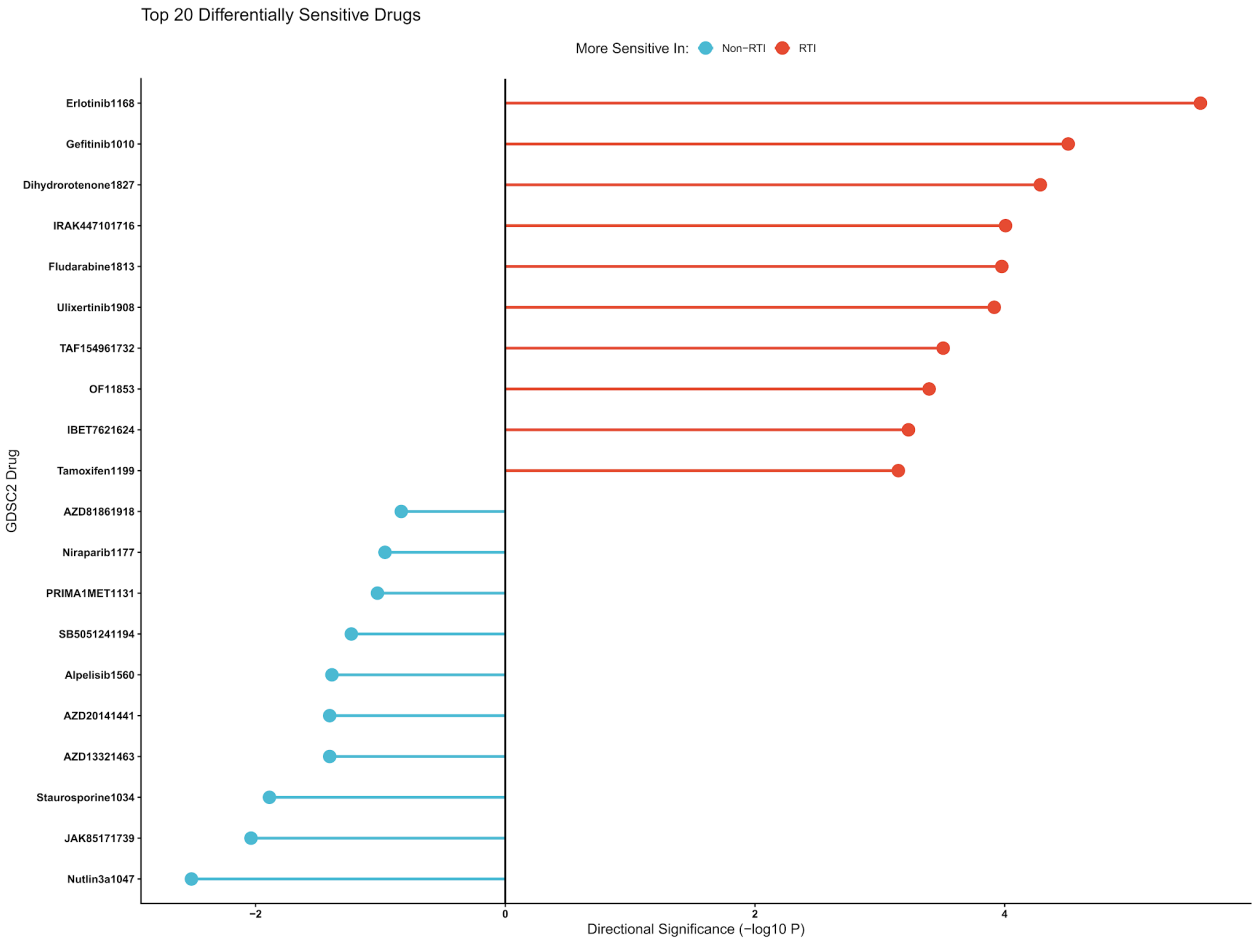


Figure 3.11 (B) Top 20 differentially sensitive drugs per direction, external cohort (n=55). Directional significance expressed as signed -log10(p-value).

CHAPTER IV: DISCUSSION

Breast angiosarcoma is a rare and aggressive vascular malignancy that arises either spontaneously or as a late complication of prior radiotherapy for breast cancer, and current knowledge of the disease is limited by the rarity and inconsistency of available data, confined largely to small retrospective case series and case reports. Radiation-induced (RTI) and non-radiation-induced (Non-RTI) subtypes differ substantially in their epidemiological profiles and clinical behaviour despite overlapping histological features, yet their molecular distinctions remain poorly characterized. This study aimed to comprehensively dissect the transcriptomic differences between RTI and Non-RTI breast angiosarcoma and to identify subtype-specific molecular features with potential therapeutic relevance. Using an internal discovery cohort of 32 patients and an independent external validation cohort of 55 patients, differential gene expression, pathway enrichment, tumour microenvironment deconvolution, and immune profiling analyses were performed independently in each cohort, with cross-cohort replication used to distinguish robust findings from cohort-specific signals. The main findings indicate that RTI breast angiosarcoma is characterized by an immune-hot transcriptional phenotype, marked by enrichment of MYC target, E2F target, mTORC1 signalling, and oxidative phosphorylation pathways, higher T cell and myeloid dendritic cell abundance, elevated antigen presentation machinery activity, and upregulation of immune checkpoint genes including *CTLA4* and *TIGIT*. In contrast, Non-RTI tumours were consistently associated with an immune-cold and stromal phenotype, evidenced by enrichment of the epithelial-mesenchymal transition pathway, elevated cancer-associated fibroblast activity, and higher endothelial cell abundance. These contrasting biological phenotypes were observed across multiple independent analytical frameworks and showed strong concordance between cohorts. Together, these findings support the hypothesis that RTI and Non-RTI breast angiosarcoma represent molecularly distinct disease entities, with implications for subtype-specific therapeutic strategies.

The significantly older age at diagnosis observed in RTI patients compared to Non-RTI patients in both cohorts reflects a fundamental biological distinction between the two subtypes. In the internal discovery cohort, RTI patients presented at a median age of 73.5

years compared to 37.2 years in Non-RTI patients, a difference replicated in the external validation cohort with median ages of 63.0 and 38.0 years respectively. RTI breast angiosarcoma arises as a late complication of radiotherapy for primary breast cancer, while Non-RTI is a de novo vascular malignancy predominantly affecting younger women. This pattern is well established in the literature, with registry-based analyses reporting a median age at diagnosis of approximately 70 to 74 years in secondary compared to 50 to 54 years in primary breast angiosarcoma (16)

The median latency from the recorded start date of prior radiotherapy to angiosarcoma diagnosis was 6.5 years in 22 of 24 RTI patients in the external cohort, consistent with published breast-specific series reporting latency periods of 5 to 10 years (14).

No statistically significant difference in disease-specific survival was observed in the internal cohort; however this reflects the severe underpowering of the analysis rather than true biological equivalence. With only one disease-specific death in the Non-RTI group, a meaningful survival comparison was not feasible. Published data consistently associate RTI breast angiosarcoma with worse outcomes compared to Non-RTI, partly attributable to older age and more advanced disease at presentation **(**16).

No grouping by centre of origin was observed in the internal discovery cohort, indicating the absence of a detectable batch effect across the three collection sites. A partial separation between RTI and Non-RTI samples was observed along the first principal component in the internal cohort, although the small Non-RTI group did not form a distinct cluster of its own, consistent with its limited sample size. A broader separation between RTI and Non-RTI samples was observed in the external validation cohort, where both groups formed more distinct clusters. Together, these exploratory analyses indicated that RTI and Non-RTI breast angiosarcoma display partially distinct transcriptomic profiles, providing a basis for the differential expression and pathway analyses that follow.

Differential gene expression analysis revealed a pronounced asymmetry between the two cohorts in both the number and directionality of significant genes, reflecting differences in cohort size and group balance rather than a fundamental biological inconsistency.

In the internal discovery cohort, 26 genes were differentially expressed between RTI and Non-RTI tumours, all of which were upregulated in RTI. The complete absence of Non-RTI-upregulated genes at the applied thresholds most likely reflects the severe group imbalance in this cohort, where only 5 Non-RTI samples were available. Among the RTI-upregulated genes, *CTC1* encodes the largest subunit of the CST complex, a conserved single-stranded DNA binding complex that plays central roles in telomere maintenance and genome-wide replication stress response (63). The CST complex regulates telomere elongation by limiting telomerase activity after G-strand synthesis, thereby preventing excessive telomere extension, and promotes the rescue of stalled replication forks at both telomeric and non-telomeric genomic loci (63). Upregulation of *CTC1* in RTI tumours may therefore reflect an adaptive response to radiation-induced replication stress, potentially stabilizing telomere length within a viable range and enhancing replication recovery under the genomic instability that characterizes radiation-exposed endothelial cells. These observations are consistent with the enrichment of DNA Repair and Reactive Oxygen Species pathways for the RTI group, which together suggest a coordinated transcriptional response to sustained genotoxic stress. *MRPL15* encodes a mitochondrial ribosomal protein of the large 39S subunit involved in mitochondrial translation and has been identified as upregulated and prognostically relevant in multiple cancer types, where it has been linked to cell cycle regulation, DNA replication, and mTORC1 signalling pathways (64). Its upregulation in RTI tumours is consistent with the oxidative phosphorylation and mTORC1 pathway enrichment.

In the external validation cohort, a cluster of epidermal differentiation complex genes appeared prominently among the highest RTI-upregulated genes on the volcano plot, including *LOR*, *FLG*, *SBSN*, and *KRTDAP*, which encode structural proteins of the cornified cell envelope with no known biological relevance to angiosarcoma. Their presence at high log2FC values reflects skin tissue contamination in a subset of external samples rather than true tumour biology. These 36 genes were therefore excluded from heatmap visualization, without affecting the differential expression results themselves. Among the most significantly RTI-upregulated genes after this exclusion, *FGFR3* encodes a receptor tyrosine kinase whose aberrant activation drives downstream proliferation and survival signalling cascades and has been documented across multiple cancer types (65).

TIMD4 encodes a phosphatidylserine receptor primarily expressed on antigen-presenting cells that plays bimodal roles in T-cell regulation, suppressing naive T-cell activation while promoting proliferation of already-activated T cells through AKT1 and ERK1/2 phosphorylation (66). Its upregulation in RTI tumours is consistent with the immune-activated phenotype of this subtype and the higher T cell and myeloid dendritic cell abundance observed in downstream analyses. Among the most significantly Non-RTI-upregulated genes, *HSPG2* encodes perlecan, a large heparan sulphate proteoglycan that is an integral component of basement membranes where it modulates extracellular matrix structure, cell adhesion, and growth factor signalling (67). Alterations in heparan sulphate composition can regulate EMT-related pathways including Wnt signalling, and the upregulation of HSPG2 in Non-RTI tumours may therefore contribute to the stromal and mesenchymal transcriptional program supported by the EMT pathway enrichment for this group.

Gene set enrichment analysis revealed a highly concordant transcriptional program between the two cohorts, with all 12 Hallmark pathways significant in the internal discovery cohort (10 RTI-enriched, 2 Non-RTI-enriched) also reaching significance in the external validation cohort with concordant direction, corresponding to a complete overlap for the internal findings as shown in the Venn diagram (Figure 3.4 D). The external cohort additionally identified 22 further RTI-enriched pathways not significant in the internal cohort, reflecting its larger and more balanced sample size, while no additional Non-RTI-enriched pathways were identified beyond the two shared with the internal cohort. This degree of concordance for the internal findings is particularly notable given the differences in cohort size, group balance, and data source, and substantially strengthens the biological interpretation of the individual pathway findings.

The most significantly RTI-activated pathways were MYC Targets V1 and V2, E2F Targets, mTORC1 Signalling, and Oxidative Phosphorylation, which together delineate a coordinated proliferative and metabolic transcriptional program. MYC is a master transcriptional regulator that promotes cell growth and proliferation by coordinating nutrient acquisition, DNA replication, and cell cycle entry, and its downstream targets include E2F transcription factors, creating a feed-forward loop that drives progression into

S phase and sustains DNA replication. The co-enrichment of MYC Targets and E2F Targets in RTI tumours in both cohorts therefore reflects the known convergence of these two oncogenic programs in actively proliferating tumour cells and is directly consistent with the established role of MYC amplification as a molecular hallmark of RTI breast angiosarcoma reported in up to 100% of secondary angiosarcomas (15). While prior pathway analysis of primary versus radiation-associated breast angiosarcoma identified TGF-β, Wnt, Hippo, and PI3K-Akt signalling using GO and KEGG enrichment on a limited set of 18 DEGs (35), the present study extends this work by applying Hallmark GSEA to a substantially larger and more balanced dataset, revealing a coordinated MYC-driven proliferative and metabolic program that was not previously characterized at this resolution and was validated in an independent cohort. The concurrent enrichment of mTORC1 Signalling and Oxidative Phosphorylation in RTI tumours is mechanistically coherent in the context of prior radiation exposure. Ionizing radiation has been shown to induce a shift from aerobic glycolysis toward mitochondrial oxidative phosphorylation through mTOR-mediated mechanisms, representing an adaptive metabolic response that supports tumour cell survival under conditions of chronic genotoxic stress (68). The enrichment of DNA Repair and Reactive Oxygen Species pathways in RTI tumours further supports this interpretation, suggesting that RTI tumour cells are transcriptionally primed to manage persistent radiation-induced oxidative damage while maintaining proliferative capacity.

In contrast, the two pathways consistently enriched in Non-RTI tumours across both cohorts were Epithelial Mesenchymal Transition (EMT) and Myogenesis. EMT is a transcriptional program through which cells acquire mesenchymal characteristics including enhanced migratory and invasive capacity, remodelling of cell-cell adhesion, and increased extracellular matrix production (30). Its enrichment in Non-RTI tumours is consistent with a stromal and mesenchymal transcriptional identity in this subtype and is further supported by the higher cancer-associated fibroblast activity and endothelial cell abundance observed in downstream deconvolution analyses. The co-enrichment of Myogenesis in Non-RTI tumours most likely reflects the transcriptional contribution of stromal myofibroblasts within the tumour microenvironment of this subtype, given that the Hallmark Myogenesis gene set captures broader muscle-lineage gene programs expressed by myofibroblasts in the

tumour stroma (69) and is consistent with the higher CAF scores and fibroblast abundance independently identified by TIDE and MCP-counter analyses in both cohorts.

Together, the pathway enrichment findings support a model in which RTI breast angiosarcoma is driven by MYC-dependent oncogenic proliferation and metabolic adaptation to genotoxic stress, while Non-RTI breast angiosarcoma is characterized by a stromal and mesenchymal transcriptional program that may underpin its distinct biological behaviour.

Tumour microenvironment composition was assessed using the ESTIMATE algorithm, which generates per-sample immune and stromal infiltration scores from bulk RNA-seq data. The interpretation of ESTIMATE results in this study requires an important technical caveat: absolute scores are not comparable across cohorts due to the different number of genes present in each expression matrix, as the score scale is influenced by matrix dimensions. All comparisons were therefore performed strictly within each cohort.

In the internal discovery cohort, no statistically significant differences were observed in Immune Score or Stromal Score between RTI and Non-RTI groups. The absence of significant ESTIMATE findings in the internal cohort most likely reflects the limited statistical power associated with only 5 Non-RTI samples, rather than a true absence of microenvironmental differences between the two groups. Notably, the direction of effect was consistent with the external cohort, with RTI tumours showing numerically higher Immune Score than Non-RTI tumours. While Stromal Score did not reach significance in either cohort, the directional trend was consistent with higher stromal abundance in Non-RTI tumours in both cohorts. This directional signal, although insufficient to constitute independent replication, is corroborated by more sensitive cell-type specific analyses including MCP-counter fibroblast scores, TIDE CAF scores, and GSEA EMT enrichment, which together provide convergent evidence for a stromal transcriptional identity in Non-RTI breast angiosarcoma. The convergence of these independent analytical frameworks strengthens the biological interpretation beyond what any single analysis could support alone.

In the external validation cohort, RTI tumours showed significantly higher Immune Score compared to Non-RTI tumours, indicating greater overall immune cell infiltration within the tumour microenvironment of radiation-induced tumours. No significant difference was observed in Stromal Score or Tumour Purity between the two groups. The finding of elevated immune infiltration in RTI breast angiosarcoma is consistent with emerging evidence from radiation-induced sarcomas more broadly, where higher immune cell abundance compared to primary sarcomas has been reported, including elevated levels of activated CD8+ T cells, dendritic cells, and regulatory T cells (38). This pattern has been attributed to the higher mutational burden and genomic instability that characterizes radiation-associated tumours, which may generate a larger pool of neoantigens capable of priming an adaptive immune response. The convergence of significantly higher Immune Score in the external cohort with the directionally consistent trend in the internal cohort, despite the latter not reaching statistical significance, supports the biological plausibility of a genuinely more immune-infiltrated microenvironment in RTI breast angiosarcoma.

Immune and stromal cell type abundances estimated by MCP-counter provided cell-type specific resolution of the tumour microenvironment differences suggested by ESTIMATE and revealed a pattern consistent with the immune-hot versus immune-cold distinction between RTI and Non-RTI breast angiosarcoma.

In the internal discovery cohort, no cell type reached statistical significance after FDR correction, most likely reflecting the insufficient statistical power of the 5 Non-RTI samples to detect cell-type specific differences. Cytotoxic lymphocytes showed a nominally higher score in RTI tumours and fibroblasts showed a nominally higher score in Non-RTI tumours, both consistent in direction with the external cohort findings, though neither survived multiple testing correction.

In the external validation cohort, T cells and myeloid dendritic cells were significantly more abundant in RTI tumours, while endothelial cells were significantly more abundant in Non-RTI tumours. The higher T cell abundance in RTI tumours is consistent with immunological profiling of angiosarcoma subgroups using multiplex immunohistochemistry, which demonstrated significantly higher CD3+, CD8+, CD4+, and

FoxP3+ T cell densities in secondary compared to primary angiosarcomas across a large cohort of 257 patients (70)**.** Critically, the clinical relevance of this immune infiltration was prospectively validated in the CEMangio phase II trial, which treated 18 patients with locally advanced or metastatic secondary angiosarcoma, including 12 with radiation-associated breast angiosarcoma, with the PD-1 inhibitor cemiplimab (19). High intratumoral CD3+, CD8+, CD4+, and FoxP3+ T cell density measured by multiplex immunohistochemistry was significantly associated with treatment response in that trial, directly mirroring the transcriptomic T cell enrichment observed in RTI tumours across both cohorts of the present study. Together, these findings suggest that the immune-hot microenvironment of RTI breast angiosarcoma identified here at the transcriptomic level has measurable clinical consequences for immunotherapy response.

The significantly higher myeloid dendritic cell abundance in RTI tumours is biologically coherent with the elevated T cell infiltration, as myeloid dendritic cells are the professional antigen presenting cells responsible for cross-presenting tumour-derived antigens to naive CD8+ T cells and initiating adaptive antitumor immune responses (71). Their concurrent enrichment with T cells in RTI tumours supports a model of coordinated immune activation in this subtype, where dendritic cell-mediated antigen presentation drives the T cell infiltration that characterizes the immune-hot phenotype.

In contrast, endothelial cells were significantly more abundant in Non-RTI tumours in the external cohort. Non-RTI breast angiosarcoma arises de novo from endothelial cells within the breast without prior radiation exposure, and the higher endothelial cell signal in bulk RNA-seq data of Non-RTI tumours is consistent with the intrinsic vascular biology of this subtype. The nominally higher fibroblast abundance in Non-RTI tumours in both cohorts, though not reaching FDR significance in either, is directionally consistent with the significantly higher CAF scores observed in TIDE analysis and with the EMT pathway enrichment identified in GSEA, further supporting a stromal and mesenchymal transcriptional identity in Non-RTI breast angiosarcoma, where CAF activity and stromal remodelling may represent key drivers of the immune-cold phenotype observed in this subtype.

The higher CAF score observed in Non-RTI tumours in both cohorts is consistent with a stromal transcriptional identity in this subtype, supported by convergent evidence from multiple independent analyses including higher fibroblast abundance in MCP-counter, EMT pathway enrichment, and the TIDE CAF score itself. The TIDE CAF score is a purely computational transcriptomic score derived from the expression signatures of cancer-associated fibroblasts, identified as one of three primary immunosuppressive cell types driving T cell exclusion alongside myeloid-derived suppressor cells and M2 tumour-associated macrophages, and validated across 43 solid tumour datasets (72). A higher CAF score therefore reflects greater transcriptional activity of CAF-associated gene programs within the tumour bulk, consistent with a microenvironment in which fibroblast-mediated stromal remodelling actively impedes cytotoxic T cell infiltration. It should be noted that the TIDE framework was originally developed and validated in melanoma cohorts treated with anti-PD1 and anti-CTLA4 (72), and its application to angiosarcoma represents an extrapolation that should be interpreted with appropriate caution.

Cancer-associated fibroblasts are increasingly recognized as key regulators of the tumour immune microenvironment across multiple cancer types, where their stromal remodelling activity and immunosuppressive signalling create physical and biochemical barriers to T cell infiltration (42)**.** In the context of sarcomas specifically, CAF activity has been linked to immune exclusion and the formation of immunosuppressive stromal niches that limit the efficacy of T cell-based immunotherapies (73). The replicated elevation of the CAF score in Non-RTI breast angiosarcoma across both cohorts suggests that analogous CAF activity may be a defining feature of the immune-cold phenotype of this subtype.

The overall TIDE score and T cell dysfunction score did not reach statistical significance in either cohort. The non-significant dysfunction score combined with the nominally elevated exclusion score and the significantly higher CAF score in Non-RTI tumours is consistent with an immune evasion mechanism operating primarily through T cell exclusion rather than T cell dysfunction in this subtype, a distinction with potential implications for the selection of therapeutic strategies aimed at overcoming immune resistance in Non-RTI breast angiosarcoma (72).

The singscore-based immune signature analysis provided additional transcriptional resolution to the elevated T cell infiltration observed in RTI tumours by MCP-counter, revealing that the immune-hot phenotype of RTI breast angiosarcoma extends beyond cell abundance to encompass coordinated upregulation of the molecular machinery required for antigen processing, presentation, and T cell activation.

The significantly higher Antigen Processing and MHC Class I scores in RTI tumors in the external cohort indicate that RTI tumor cells display a more active antigen presentation program, reflected by higher expression of the genes composing this signature. The concurrent elevation of the T cell inflamed signature score in RTI tumors further supports the interpretation that these tumors sustain an actively inflamed immune microenvironment rather than one characterized by passive immune infiltration alone. Together, these findings suggest that the higher T cell abundance identified by MCP-counter in RTI tumors reflects a functionally engaged antitumor immune response accompanying active antigen presentation, rather than non-specific immune cell accumulation.

The significantly higher TLS signature score in RTI tumours in the external cohort is particularly noteworthy from a clinical standpoint. Tertiary lymphoid structures are organized ectopic lymphoid aggregates that form within tumour tissue under conditions of sustained immune activation, and their presence has been associated with favourable prognosis and improved response to immune checkpoint inhibitors across multiple cancer types including sarcomas (74)**.** In angiosarcoma specifically, TLS formation has been shown to correlate with favourable prognosis in cutaneous angiosarcoma, where TLS-positive cases demonstrated distinct transcriptomic profiles by RNA sequencing and gene set enrichment analysis compared to TLS-negative cases (25). The elevated TLS signature in RTI tumours in the present study suggests that RTI breast angiosarcoma may be particularly prone to the formation of organized intratumoral immune structures, which could represent a substrate for effective immunotherapy response in this subtype.

The significantly higher Immune Checkpoint signature score in RTI tumours reflects elevated expression of inhibitory checkpoint molecules including CD274, PDCD1, CTLA4, LAG3, HAVCR2, and TIGIT at the gene set level. This finding should not be interpreted as

evidence of immune suppression in RTI tumours. When T cells become activated and begin mounting an antitumor immune response, they upregulate checkpoint molecules on their own surface as a physiological mechanism to prevent excessive immune activity and autoimmune damage. The elevated checkpoint signature in RTI tumours therefore most likely reflects the degree of T cell activation and engagement rather than primary immune dysfunction, an interpretation consistent with the non-significant T cell dysfunction score observed in TIDE analysis for both cohorts (72).

MHC Class II was the only signature that did not reach statistical significance in the external cohort. Unlike MHC Class I molecules, which are expressed constitutively on virtually all nucleated cells including tumour cells, MHC Class II expression is restricted to professional antigen presenting cells such as dendritic cells, B cells, and macrophages. In bulk RNA-seq data, the MHC Class II signal therefore reflects the overall abundance of antigen presenting cells in the tumour rather than tumour-intrinsic antigen presentation activity. The absence of a significant MHC Class II difference between RTI and Non-RTI tumours may reflect the fact that while myeloid dendritic cells were significantly more abundant in RTI tumours by MCP-counter, the total pool of MHC Class II-expressing cells which includes B cells and macrophages that were not significantly different between groups may have been comparable between the two subtypes at the bulk transcriptomic level.

In the internal discovery cohort, none of the six signatures reached statistical significance after FDR correction, though the direction of effect was consistent with the external cohort across all signatures, with RTI tumours showing higher scores in every case. This consistent directionality across both cohorts, despite the internal cohort lacking the statistical power to detect significant differences with only five Non-RTI samples, substantially strengthens the biological credibility of the external findings.

MSI probability scores derived from bulk RNA-seq data remained low in both RTI and Non-RTI tumours across both cohorts, with no sample approaching the conventional MSI-high threshold in either group. Although a nominally higher MSI probability score was observed in RTI tumours in the external cohort, this difference was not replicated in the

internal cohort and carries limited biological meaning given that both groups remained well within the MSI-low range.

The significantly higher *CTLA4* expression in RTI tumours, observed in a consistent direction across both cohorts and reaching statistical significance in the external cohort, represents one of the most clinically relevant findings of this study. *CTLA4* encodes cytotoxic T lymphocyte associated protein 4, an inhibitory receptor expressed on activated T cells and regulatory T cells that competes with the co-stimulatory receptor CD28 for binding to CD80 and CD86 on antigen presenting cells, thereby attenuating T cell activation (75). Its elevated expression in RTI tumours is consistent with the broader immune-hot phenotype of this subtype, *CTLA4* is upregulated on T cells as a consequence of activation rather than as a primary driver of immune suppression, and its transcriptomic elevation therefore reflects the degree of T cell engagement in the tumour microenvironment. This interpretation is supported by the CEMangio phase II trial, in which high intratumoral T cell density measured by multiplex immunohistochemistry predicted response to PD-1 blockade specifically in RTI breast angiosarcoma patients (19), suggesting that the immune-activated microenvironment reflected by *CTLA4* upregulation in RTI tumours may translate into clinical ICI sensitivity. The clinical rationale for targeting CTLA4 in angiosarcoma more broadly is supported by the DART trial (SWOG S1609, cohort 51), a prospective multicentre phase II study that treated 16 patients with metastatic or unresectable angiosarcoma of mixed anatomical origins with dual ipilimumab and nivolumab, targeting CTLA4 and PD-1 simultaneously, and reported an objective response rate of 25% (22). It should be noted however that three of the four confirmed responses occurred in patients with primary cutaneous scalp or face angiosarcoma, which are UV-signature tumours with high mutational burden biologically distinct from RTI breast angiosarcoma, and the trial did not stratify patients by radiation history or breast origin. The DART results therefore support the broader principle of CTLA4 and PD-1 co-targeting in angiosarcoma while underscoring the need for subtype-specific trial design.

TIGIT was also significantly higher in RTI tumours in the external cohort, though this finding was not replicated in the internal cohort. TIGIT is an inhibitory receptor expressed on activated T cells, regulatory T cells, and NK cells that suppresses antitumor immunity

by competing with the activating receptor CD226 for binding to shared ligands CD112 and CD155 (76). TIGIT expression increases on T cells following activation and is preferentially co-expressed with PD-1 on tumour infiltrating lymphocytes across multiple cancer types, making it a marker of T cell engagement in the tumour microenvironment rather than a standalone suppressive mechanism in isolation. Its elevated expression in RTI tumours is consistent with the broader pattern of immune activation observed across multiple analyses in this subtype and represents a potential co-inhibitory therapeutic target alongside CTLA4 in RTI breast angiosarcoma.

The findings of this study, taken together across multiple independent analytical frameworks, support a coherent biological model in which RTI and Non-RTI breast angiosarcoma represent molecularly distinct disease entities that differ fundamentally in their transcriptional programs, tumour microenvironment composition, and immune landscape.

RTI breast angiosarcoma is characterized by an immune-hot and proliferative phenotype whose biological logic can be traced to the consequences of prior radiation exposure. Ionizing radiation induces double-strand DNA breaks, chromosomal instability, and oxidative stress in endothelial cells within the irradiated field, creating a mutational landscape that drives MYC amplification, the defining molecular event of RTI angiosarcoma (15). MYC amplification activates a downstream transcriptional cascade encompassing E2F targets, mTORC1 signalling, and oxidative phosphorylation, reflecting a tumour that is both proliferatively reprogrammed and metabolically adapted to sustained genotoxic stress. The same radiation-induced genomic instability that drives this proliferative program also generates a higher burden of tumour-specific neoantigens, which activates the antigen presentation machinery, reflected in the higher Antigen Processing, MHC Class I, and T cell inflamed GEP scores in RTI tumours, and recruits myeloid dendritic cells that cross-present these neoantigens to naive CD8+ T cells, initiating an adaptive antitumor immune response. The result is an immune-infiltrated microenvironment with significantly higher T cell abundance, elevated TLS signature scores suggesting organized intratumoral immune structures, and upregulation of immune checkpoint molecules including CTLA4 and TIGIT as a physiological consequence of

sustained T cell activation rather than primary immune suppression. The convergence of GSEA pathway enrichment, ESTIMATE Immune Score, MCP-counter T cell and dendritic cell abundance, singscore immune signatures, and checkpoint gene expression across both cohorts into a single coherent immune-hot narrative represents the central and most robust finding of this study, with direct clinical implications supported by the CEMangio trial demonstrating that high intratumoral T cell density predicts response to PD-1 blockade specifically in RTI breast angiosarcoma (19).

Non-RTI breast angiosarcoma presents the biological mirror image. These tumours are defined by a stromal and mesenchymal transcriptional program, evidenced by consistent enrichment of EMT and Myogenesis pathways across both cohorts, higher CAF scores replicated in both cohorts by TIDE, nominally higher fibroblast abundance by MCP-counter, and a directionally consistent higher Stromal Score by ESTIMATE. The elevated CAF activity in Non-RTI tumours is the mechanistic link between the stromal transcriptional program and the immune-cold phenotype: myofibroblastic CAFs remodel the extracellular matrix and create physical and biochemical barriers that exclude cytotoxic T cells from the tumour parenchyma (72), consistent with the nominally elevated exclusion score and non-significant dysfunction score in TIDE analysis, which together point to T cell exclusion rather than T cell dysfunction as the primary mechanism of immune evasion in this subtype. The higher endothelial cell abundance in Non-RTI tumours by MCP-counter further reflects the intrinsic vascular biology of Non-RTI angiosarcoma arising from endothelial cells.

This model is schematically represented in Figure 4.13, which illustrates the proposed biological distinction between RTI and Non-RTI breast angiosarcoma across the dimensions of oncogenic driver, metabolic program, immune microenvironment composition, and stromal architecture.

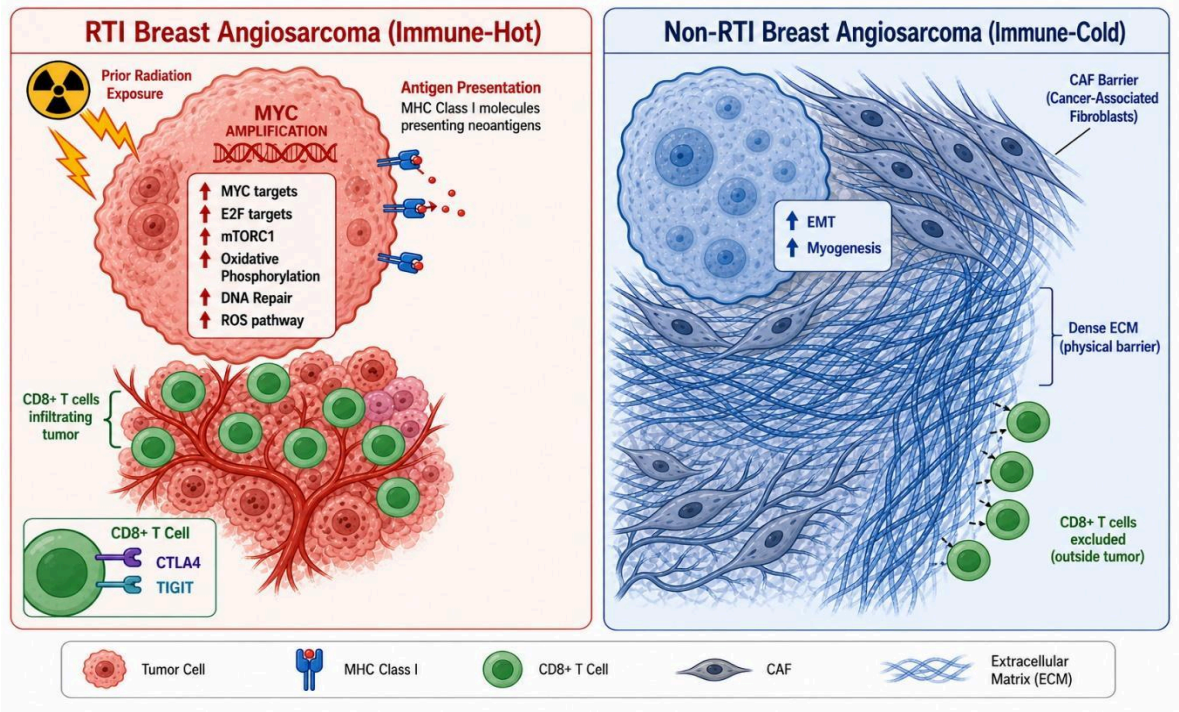


Figure 4.13 Proposed biological model contrasting the immune-hot, MYC-driven oncogenic and metabolic program of RTI breast angiosarcoma with the immune-cold, CAF-rich stromal program of Non-RTI breast angiosarcoma.

The transcriptomic findings of this study have direct biological and clinical implications for patient stratification and therapeutic decision-making in breast angiosarcoma, a disease for which subtype-specific treatment strategies have not been formally established (77).

The immune-hot phenotype of RTI breast angiosarcoma supports its prioritization as the subtype most likely to benefit from immune checkpoint blockade. The clinical validity of this transcriptomic profile is most directly supported by the CEMangio phase II trial, in which high intratumoral T cell density predicted response to PD-1 blockade with cemiplimab specifically in RTI breast angiosarcoma patients (19). The convergent evidence from ESTIMATE, MCP-counter, singscore, and checkpoint expression analyses across two independent cohorts strengthens the biological rationale for prospective trials that enrol RTI patients based on transcriptomic immune profiling rather than histological diagnosis alone.

The significantly higher TLS signature score in RTI tumours carries additional clinical relevance. TLS presence has been associated with improved prognosis and greater benefit

from immune checkpoint inhibition in soft tissue sarcomas (74) and in angiosarcoma specifically (25), suggesting that TLS scoring from bulk RNA-seq or histological assessment could serve as a stratification biomarker in future clinical trials enrolling RTI breast angiosarcoma patients.

The upregulation of *CTLA4* and *TIGIT* in RTI tumours further supports the potential utility of dual checkpoint blockade in this subtype. The DART trial demonstrated that dual ipilimumab and nivolumab produced an objective response rate of 25% in angiosarcoma broadly, though the responding patients were predominantly cutaneous UV-signature tumours rather than RTI breast angiosarcoma specifically (22). The mechanistic rationale for CTLA4 and TIGIT blockade in a tumour with high expression of both molecules and a highly activated T cell infiltrate applies directly to RTI breast angiosarcoma and warrants dedicated prospective evaluation.

For Non-RTI breast angiosarcoma, the immune-cold and CAF-rich stromal phenotype suggests that standard immune checkpoint blockade is unlikely to be effective without concurrent strategies to overcome the stromal barrier.

More broadly, RTI and Non-RTI breast angiosarcoma are currently treated as a single disease entity in most clinical settings despite their fundamentally distinct molecular profiles. Incorporating transcriptomic immune and stromal profiling into the clinical workup of breast angiosarcoma patients could improve patient stratification and guide more rational subtype-specific therapeutic strategies (77).

Study Limitations

As with any retrospective transcriptomic study of a rare cancer, certain limitations should be acknowledged when interpreting the findings presented here. Given the overall sample sizes of both cohorts, this study should be considered exploratory and hypothesis-generating rather than confirmatory, and its findings warrant validation in larger, prospectively collected cohorts before firm therapeutic conclusions can be drawn. The internal discovery cohort included only five Non-RTI patients, which limited the statistical power of analyses performed within this cohort and likely explains the absence of

FDR-significant findings across several analytical frameworks despite directionally consistent results with the external cohort.

This study is entirely computational and based on bulk RNA sequencing. Bulk RNA sequencing captures an averaged transcriptional signal across all cell types within a sample, and computational deconvolution tools such as MCP-counter estimate relative cell-type abundance from this averaged signal but cannot resolve cell-type-specific gene expression, distinguish transcriptionally similar subpopulations, or capture spatial organization within the tumour. While independent immunohistochemistry data from van Ravensteijn et al. (13) support higher T cell density in RTI relative to Non-RTI angiosarcoma at the protein level, this thesis does not include orthogonal experimental validation of its own findings, such as immunohistochemistry, flow cytometry, or spatial transcriptomics, which would be required to confirm cell-type proportions and spatial architecture within these specific cohorts.

RTI patients were significantly older than Non-RTI patients in both cohorts. As immunosenescence is known to alter immune cell composition independently of underlying disease biology, age may represent a partial confounder of the immune-hot phenotype observed in RTI tumours that this study cannot fully disentangle from radiation status alone.

Clinical outcome data were not comprehensively assessed in this study due to a limited number of disease-specific events in the internal cohort and the absence of survival data in the external cohort. Whether the transcriptomic differences identified between RTI and Non-RTI breast angiosarcoma translate into differences in clinical outcome therefore remains to be established.

Future Directions

The findings of this study suggest several directions for future research. At the clinical level, the most immediate priority is the design of prospective trials that stratify breast angiosarcoma patients by RTI and Non-RTI status at enrolment, rather than treating these two biologically distinct subtypes as a single disease entity. Such stratification would allow

subtype-specific therapeutic hypotheses to be tested rigorously and would prevent the dilution of treatment signals that arises when molecularly heterogeneous populations are analysed together.

Within the RTI subtype, the significantly higher *CTLA4* expression identified across both cohorts provides a transcriptomic rationale for evaluating dual checkpoint blockade targeting both PD-1 and CTLA4 specifically in RTI breast angiosarcoma, building on the promising activity of anti-PD-1 monotherapy demonstrated in the CEMangio trial as previously discussed, and warrants dedicated prospective evaluation. TIGIT represents an additional candidate co-inhibitory target in RTI breast angiosarcoma, given its significant upregulation in this subtype in the external cohort. Anti-TIGIT agents are currently in clinical development across multiple cancer types but have not yet received regulatory approval in any indication, and future studies should evaluate whether TIGIT expression in RTI breast angiosarcoma tumours can serve as a predictive biomarker for response to anti-TIGIT therapy.

For Non-RTI breast angiosarcoma, the replicated CAF-rich immune-cold phenotype provides a preclinical rationale for evaluating anti-stromal strategies aimed at reducing CAF activity and restoring cytotoxic T cell access to the tumour parenchyma, which could potentially sensitize Non-RTI tumours to subsequent immune checkpoint inhibition and warrants investigation in appropriate preclinical angiosarcoma models.

Finally, the transcriptomic characterization presented here represents one layer of a more complete molecular portrait of breast angiosarcoma. Integration of genomic, epigenomic, and proteomic data with the RNA-seq findings of this study would allow a more comprehensive understanding of the molecular drivers and therapeutic vulnerabilities of both subtypes and would provide the foundation for biomarker-driven precision oncology approaches in this rare disease.

Conclusion

This study provides an independently validated transcriptomic characterization of RTI and Non-RTI breast angiosarcoma across two cohorts, demonstrating that these two subtypes represent molecularly distinct disease entities despite sharing the same histological diagnosis. Through the convergent evidence of differential gene expression, pathway enrichment, tumour microenvironment deconvolution, immune profiling, and checkpoint expression analyses, RTI breast angiosarcoma was consistently identified as an immune-hot tumour characterized by MYC-driven proliferative reprogramming, active antigen presentation, T cell infiltration, and upregulation of immune checkpoint molecules including CTLA4 and TIGIT, suggesting CTLA4 and TIGIT blockade as rational therapeutic strategies for this subtype, while Non-RTI breast angiosarcoma was characterized by an immune-cold and stromal phenotype driven by CAF activity, EMT enrichment, and higher endothelial cell and fibroblast abundance. The strong cross-cohort concordance of these findings, replicated across independent discovery and validation cohorts, substantially strengthens their biological credibility and distinguishes them from cohort-specific artefacts.This study offers a relatively detailed characterization of the immune and stromal microenvironment of RTI and Non-RTI breast angiosarcoma, supported by independent validation across two cohorts and the findings presented here provide a molecular foundation for subtype-specific therapeutic strategies in a disease that has historically been approached as a single entity. It is hoped that these results will contribute to improving patient stratification and guiding more rational treatment decisions in breast angiosarcoma.

References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57–63. doi:10.1038/nrg2484

2. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 Dec;17(1):13. doi:10.1186/s13059-016-0881-8

3. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000 Aug 17;406(6797):747–52. doi:10.1038/35021093

4. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J Clin Oncol. 2009 Mar 10;27(8):1160–7. doi:10.1200/JCO.2008.18.1370

5. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012 Oct;490(7418):61–70. doi:10.1038/nature11412

6. Han Y, Shah A, Yao Y, Mutter RW, Xu-Welliver M. Clinical Differences Among Histological Categories of Sarcoma: Insights from 97,062 Patients. Cancers. 2025 May 20;17(10):1706. doi:10.3390/cancers17101706

7. Merry E, Thway K, Jones RL, Huang PH. Predictive and prognostic transcriptomic biomarkers in soft tissue sarcomas. Npj Precis Oncol. 2021 Mar 5;5(1):17. doi:10.1038/s41698-021-00157-4

8. Pennock ND, Jindal S, Horton W, Sun D, Narasimhan J, Carbone L, et al. RNA-seq from archival FFPE breast cancer samples: molecular pathway fidelity and novel discovery. BMC Med Genomics. 2019 Dec;12(1):195. doi:10.1186/s12920-019-0643-z

9. Benton A, Liu B, Gartenhaus LE, Hanna JA. Genomic landscape and preclinical models of angiosarcoma. Mol Oncol. 2025 Apr;19(4):965–83. doi:10.1002/1878-0261.13744

10. Gaballah AH, Jensen CT, Palmquist S, Pickhardt PJ, Duran A, Broering G, et al. Angiosarcoma: clinical and imaging features from head to toe. Br J Radiol. 2017 Jul 1;90(1075):20170039. doi:10.1259/bjr.20170039

11. Colas M, Gérazime A, Popescu D, Puzenat E, Chaigneau L, Woronoff AS, et al. Angiosarcoma: A population-based cancer registry descriptive study of 45 consecutive cases diagnosed between 1979 and 2016. Rare Tumors. 2020 Jan;12:2036361320979216. doi:10.1177/2036361320979216

12. Wagner MJ, Ravi V, Schaub SK, Kim EY, Sharib J, Mogal H, et al. Incidence and Presenting Characteristics of Angiosarcoma in the US, 2001-2020. JAMA Netw Open. 2024 Apr 12;7(4):e246235. doi:10.1001/jamanetworkopen.2024.6235

13. Van Ravensteijn SG, Versleijen-Jonkers YMH, Hillebrandt-Roeffen MHS, Weidema ME, Nederkoorn MJL, Bol KF, et al. Immunological and Genomic Analysis Reveals Clinically Relevant Distinctions between Angiosarcoma Subgroups. Cancers. 2022 Nov 30;14(23):5938. doi:10.3390/cancers14235938

14. Cozzi S, Finocchi Ghersi S, Tava F, Bardoscia L, Najafi M, Ruggieri MP, et al. Radiation-Associated Angiosarcoma of the Breast: The State of the Art of a Rare and Aggressive Disease. J Pers Med. 2024 Aug 14;14(8):859. doi:10.3390/jpm14080859

15. Guo T, Zhang L, Chang N, Singer S, Maki RG, Antonescu CR. Consistent *MYC* and *FLT4* gene amplification in radiation‐induced angiosarcoma but not in other radiation‐associated atypical vascular lesions. Genes Chromosomes Cancer. 2011 Jan;50(1):25–33. doi:10.1002/gcc.20827

16. Yin M, Wang W, Drabick JJ, Harold HA. Prognosis and treatment of non-metastatic primary and secondary breast angiosarcoma: a comparative study. BMC Cancer. 2017 Dec;17(1):295. doi:10.1186/s12885-017-3292-7

17. Yan Q, Fernandez RA, Elmi M, Gelfond J, Davies MG. Outcomes of Interventions for Angiosarcoma. Front Surg. 2022 Apr 11;9:819099. doi:10.3389/fsurg.2022.819099

18. Buta M, Santrac N, Zegarac M, Goran M, Jeftic N, Savkovic N, et al. Radiation-Induced Breast Angiosarcoma—A Single-Institution Experience. Diagnostics. 2024 Oct 18;14(20):2326. doi:10.3390/diagnostics14202326

19. van Ravensteijn SG, de Haan JJ, Gelderblom H, Nederkoorn MJL, Hillebrandt-Roeffen MHS, Gorris MAJ, et al. Cemiplimab in Locally Advanced or Metastatic Secondary Angiosarcomas (CEMangio): A Phase II Clinical Trial and Biomarker Analyses. Clin Cancer Res Off J Am Assoc Cancer Res. 2025 Sep

2;31(17):3678–91. doi:10.1158/1078-0432.CCR-25-0311 PubMed PMID: 40632032; PubMed Central PMCID: PMC12402795.

20. Italiano A, Cioffi A, Penel N, Levra MG, Delcambre C, Kalbacher E, et al. Comparison of doxorubicin and weekly paclitaxel efficacy in metastatic angiosarcomas. Cancer. 2012 Jul;118(13):3330–6. doi:10.1002/cncr.26599

21. van der Graaf WTA, Blay JY, Chawla SP, Kim DW, Bui-Nguyen B, Casali PG, et al. Pazopanib for metastatic soft-tissue sarcoma (PALETTE): a randomised, double-blind, placebo-controlled phase 3 trial. Lancet. 2012 May 19;379(9829):1879–86. doi:10.1016/S0140-6736(12)60651-5 PubMed PMID: 22595799.

22. Wagner MJ, Othus M, Patel SP, Ryan C, Sangal A, Powers B, et al. Multicenter phase II trial (SWOG S1609, cohort 51) of ipilimumab and nivolumab in metastatic or unresectable angiosarcoma: a substudy of dual anti-CTLA-4 and anti-PD-1 blockade in rare tumors (DART). J Immunother Cancer. 2021 Aug;9(8):e002990. doi:10.1136/jitc-2021-002990 PubMed PMID: 34380663; PubMed Central PMCID: PMC8330584.

23. Liu Q, Shi X, Tayier N, Ma L. Tumor irradiation induced immunogenic response: the impact of DNA damage induction and misrepair. Radiat Oncol. 2025 Aug 21;20(1):133. doi:10.1186/s13014-025-02711-x

24. Furtado LV, Bifulco C, Dolderer D, Hsiao SJ, Kipp BR, Lindeman NI, et al. Recommendations for Tumor Mutational Burden Assay Validation and Reporting. J Mol Diagn. 2024 Aug;26(8):653–68. doi:10.1016/j.jmoldx.2024.05.002

25. Magara T, Nakamura M, Nojiri Y, Yoshimitsu M, Kano S, Kato H, et al. Tumor immune microenvironment of cutaneous angiosarcoma with cancer testis antigens and the formation of tertiary lymphoid structures. Front Oncol. 2023 Apr 4;13:1106434. doi:10.3389/fonc.2023.1106434

26. Sahai E, Astsaturov I, Cukierman E, DeNardo DG, Egeblad M, Evans RM, et al. A framework for advancing our understanding of cancer-associated fibroblasts. Nat Rev Cancer. 2020 Mar;20(3):174–86. doi:10.1038/s41568-019-0238-1

27. Galon J, Bruni D. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. Nat Rev Drug Discov. 2019 Mar;18(3):197–218. doi:10.1038/s41573-018-0007-y

28. Wu B, Zhang B, Li B, Wu H, Jiang M. Cold and hot tumors: from molecular mechanisms to targeted therapy. Signal Transduct Target Ther. 2024 Oct 18;9(1):274. doi:10.1038/s41392-024-01979-x

29. Amer H, Flanagan KL, Kampan NC, Itsiopoulos C, Scott CL, Kartikasari AER, et al. Interleukin-6 Is a Crucial Factor in Shaping the Inflammatory Tumor Microenvironment in Ovarian Cancer and Determining Its Hot or Cold Nature with Diagnostic and Prognostic Utilities. Cancers. 2025 May 17;17(10):1691. doi:10.3390/cancers17101691

30. Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. J Clin Invest. 2009 Jun 1;119(6):1420–8. doi:10.1172/JCI39104

31. Taki M, Abiko K, Ukita M, Murakami R, Yamanoi K, Yamaguchi K, et al. Tumor Immune Microenvironment during Epithelial–Mesenchymal Transition. Clin Cancer Res. 2021 Sep 1;27(17):4669–79. doi:10.1158/1078-0432.CCR-20-4459

32. Wei Y, Yang X, Gao L, Xu Y, Yi C. Differences in potential key genes and pathways between primary and radiation-associated angiosarcoma of the breast. Transl Oncol. 2022 May;19:101385. doi:10.1016/j.tranon.2022.101385

33. He X, Xu C. Immune checkpoint signaling and cancer immunotherapy. Cell Res. 2020 Aug;30(8):660–9. doi:10.1038/s41422-020-0343-4

34. Lipson EJ, Drake CG. Ipilimumab: An Anti-CTLA-4 Antibody for Metastatic Melanoma. Clin Cancer Res. 2011 Nov 15;17(22):6958–62. doi:10.1158/1078-0432.CCR-11-1595

35. Painter CA, Jain E, Tomson BN, Dunphy M, Stoddard RE, Thomas BS, et al. The Angiosarcoma Project: enabling genomic and clinical discoveries in a rare cancer through patient-partnered research. Nat Med. 2020 Feb;26(2):181–7. doi:10.1038/s41591-019-0749-z

36. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047–8. doi:10.1093/bioinformatics/btw354 PubMed PMID: 27312411; PubMed Central PMCID: PMC5039924.

37. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer

Genomics Data. Cancer Discov. 2012 May 1;2(5):401–4. doi:10.1158/2159-8290.CD-12-0095

38. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci Signal. 2013 Apr 2;6(269). doi:10.1126/scisignal.2004088

39. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011 May 2;17(1):10. doi:10.14806/ej.17.1.200

40. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15–21. doi:10.1093/bioinformatics/bts635 PubMed PMID: 23104886; PubMed Central PMCID: PMC3530905.

41. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi:10.1186/s13059-014-0550-8 PubMed PMID: 25516281; PubMed Central PMCID: PMC4302049.

42. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40. doi:10.1093/bioinformatics/btp616 PubMed PMID: 19910308; PubMed Central PMCID: PMC2796818.

43. R core team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing. Available from: https://www.R-project.org/

44. Galili T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics. 2015 Nov 15;31(22):3718–20. doi:10.1093/bioinformatics/btv428 PubMed PMID: 26209431; PubMed Central PMCID: PMC4817050.

45. Ward JH. Hierarchical Grouping to Optimize an Objective Function. J Am Stat Assoc. 1963 Mar;58(301):236–44. doi:10.1080/01621459.1963.10500845

46. Wickham H. ggplot2 [Internet]. Cham: Springer International Publishing; 2016 [cited 2026 Apr 16]. (Use R!). Available from: http://link.springer.com/10.1007/978-3-319-24277-4 doi:10.1007/978-3-319-24277-4

47. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics. 2019 Jun 1;35(12):2084–92. doi:10.1093/bioinformatics/bty895 PubMed PMID: 30395178; PubMed Central PMCID: PMC6581436.

48. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016 Sep 15;32(18):2847–9. doi:10.1093/bioinformatics/btw313 PubMed PMID: 27207943.

49. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. Mol Cell Proteomics MCP. 2014 Feb;13(2):397–406. doi:10.1074/mcp.M113.035600 PubMed PMID: 24309898; PubMed Central PMCID: PMC3916642.

50. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innov Camb Mass. 2021 Aug 28;2(3):100141. doi:10.1016/j.xinn.2021.100141 PubMed PMID: 34557778; PubMed Central PMCID: PMC8454663.

51. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015 Dec 23;1(6):417–25. doi:10.1016/j.cels.2015.12.004 PubMed PMID: 26771021; PubMed Central PMCID: PMC4707969.

52. Carlson M. org.Hs.eg.db [Internet]. Bioconductor; 2017 [cited 2026 Apr 16]. Available from: https://bioconductor.org/packages/org.Hs.eg.db doi:10.18129/B9.BIOC.ORG.HS.EG.DB

53. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Stat Methodol. 1995 Jan 1;57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

54. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612. doi:10.1038/ncomms3612 PubMed PMID: 24113773; PubMed Central PMCID: PMC3826632.

55. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016 Oct 20;17(1):218. doi:10.1186/s13059-016-1070-5 PubMed PMID: 27765066; PubMed Central PMCID: PMC5073889.

56. Dong F, Tao C, Wu J, Su Y, Wang Y, Wang Y, et al. Author Correction: Detection of cervical lymph node metastasis from oral cavity cancer using a non-radiating, noninvasive digital infrared thermal imaging system. Sci Rep. 2018 Jul 10;8(1):10624. doi:10.1038/s41598-018-28632-2 PubMed PMID: 29991743; PubMed Central PMCID: PMC6039482.

57. Foroutan M, Bhuva DD, Lyu R, Horan K, Cursons J, Davis MJ. Single sample scoring of molecular phenotypes. BMC Bioinformatics. 2018 Nov 6;19(1):404. doi:10.1186/s12859-018-2435-4 PubMed PMID: 30400809; PubMed Central PMCID: PMC6219008.

58. Ayers M, Lunceford J, Nebozhyn M, Murphy E, Loboda A, Kaufman DR, et al. IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. J Clin Invest. 2017 Aug 1;127(8):2930–40. doi:10.1172/JCI91190 PubMed PMID: 28650338; PubMed Central PMCID: PMC5531419.

59. Wu TD, Madireddi S, de Almeida PE, Banchereau R, Chen YJJ, Chitre AS, et al. Peripheral T cell expansion predicts tumour infiltration and clinical response. Nature. 2020 Mar;579(7798):274–8. doi:10.1038/s41586-020-2056-8 PubMed PMID: 32103181.

60. Jia P, Yang X, Yang X, Wang T, Xu Y, Ye K. MSIsensor-RNA: Microsatellite Instability Detection for Bulk and Single-cell Gene Expression Data. Liu S, editor. Genomics Proteomics Bioinformatics. 2024 Sep 13;22(3):qzae004. doi:10.1093/gpbjnl/qzae004

61. Anderson AC, Joller N, Kuchroo VK. Lag-3, Tim-3, and TIGIT: Co-inhibitory Receptors with Specialized Functions in Immune Regulation. Immunity. 2016 May 17;44(5):989–1004. doi:10.1016/j.immuni.2016.05.001 PubMed PMID: 27192565; PubMed Central PMCID: PMC4942846.

62. Maeser D, Gruener RF, Huang RS. oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data. Brief

Bioinform. 2021 Nov 5;22(6):bbab260. doi:10.1093/bib/bbab260 PubMed PMID: 34260682; PubMed Central PMCID: PMC8574972.

63. Stewart JA. Emerging roles of CST in maintaining genome stability and human disease. Front Biosci. 2018;23(8):1564–86. doi:10.2741/4661

64. Xu H, Zou R, Li F, Liu J, Luan N, Wang S, et al. MRPL15 is a novel prognostic biomarker and therapeutic target for epithelial ovarian cancer. Cancer Med. 2021 Jun;10(11):3655–73. doi:10.1002/cam4.3907

65. Helsten T, Schwaederle M, Kurzrock R. Fibroblast growth factor receptor signaling in hereditary and neoplastic disease: biologic and clinical implications. Cancer Metastasis Rev. 2015 Sep;34(3):479–96. doi:10.1007/s10555-015-9579-8

66. Liu W, Xu L, Liang X, Liu X, Zhao Y, Ma C, et al. Tim-4 in Health and Disease: Friend or Foe? Front Immunol. 2020 Apr 2;11:537. doi:10.3389/fimmu.2020.00537

67. Elgundi Z, Papanicolaou M, Major G, Cox TR, Melrose J, Whitelock JM, et al. Cancer Metastasis: The Role of the Extracellular Matrix and the Heparan Sulfate Proteoglycan Perlecan. Front Oncol. 2020 Jan 17;9:1482. doi:10.3389/fonc.2019.01482

68. Lu CL, Qin L, Liu HC, Candas D, Fan M, Li JJ. Tumor Cells Switch to Mitochondrial Oxidative Phosphorylation under Radiation via mTOR-Mediated Hexokinase II Inhibition - A Warburg-Reversing Effect. Lee YJ, editor. PLOS ONE. 2015 Mar 25;10(3):e0121046. doi:10.1371/journal.pone.0121046

69. Chida K. Gastric cancer with enhanced myogenesis is associated with less cell proliferation, enriched epithelial-to-mesenchymal transition and angiogenesis, and poor clinical outcomes. Am J Cancer Res. 2024;14(1):355–67. doi:10.62347/NCIM3072

70. Van Ravensteijn SG, Versleijen-Jonkers YMH, Hillebrandt-Roeffen MHS, Weidema ME, Nederkoorn MJL, Bol KF, et al. Immunological and Genomic Analysis Reveals Clinically Relevant Distinctions between Angiosarcoma Subgroups. Cancers. 2022 Nov 30;14(23):5938. doi:10.3390/cancers14235938

71. Del Prete A, Salvi V, Soriani A, Laffranchi M, Sozio F, Bosisio D, et al. Dendritic cell subsets in cancer immunity and tumor antigen sensing. Cell Mol Immunol. 2023 Mar 22;20(5):432–47. doi:10.1038/s41423-023-00990-6

72. Jiang P, Gu S, Pan D, Fu J, Sahu A, Hu X, et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. Nat Med. 2018 Oct;24(10):1550–8. doi:10.1038/s41591-018-0136-1

73. Badran O, Cohen I, Bar-Sela G. Cancer-Associated Fibroblasts in Solid Tumors and Sarcomas: Heterogeneity, Function, and Therapeutic Implications. Cells. 2025 Sep 7;14(17):1398. doi:10.3390/cells14171398

74. Wang XX, Liu YP, Lu Y, Wu LH, Ren JY, Ji H, et al. Identifying specific TLS-associated genes as potential biomarkers for predicting prognosis and evaluating the efficacy of immunotherapy in soft tissue sarcoma. Front Immunol. 2024 Apr 24;15:1372692. doi:10.3389/fimmu.2024.1372692

75. Buchbinder EI, Desai A. CTLA-4 and PD-1 Pathways: Similarities, Differences, and Implications of Their Inhibition. Am J Clin Oncol. 2016 Feb;39(1):98–106. doi:10.1097/COC.0000000000000239

76. Annese T, Tamma R, Ribatti D. Update in TIGIT Immune-Checkpoint Role in Cancer. Front Oncol. 2022 May 17;12:871085. doi:10.3389/fonc.2022.871085

77. Chen TWW, Burns J, Jones RL, Huang PH. Optimal Clinical Management and the Molecular Biology of Angiosarcomas. Cancers. 2020 Nov 10;12(11):3321. doi:10.3390/cancers12113321

Appendix

This appendix contains the complete analysis pipeline developed for the transcriptomic comparison of RTI and Non-RTI breast angiosarcoma across two independent cohorts. The pipeline is implemented primarily in R, with two analytical steps requiring Python: tumour immune dysfunction and exclusion analysis using the tidepy package (TIDE) and microsatellite instability prediction using MSIsensor-RNA. All scripts are organised sequentially, reflecting the order of execution from raw data preprocessing through to downstream immune and clinical analyses. The pipeline is structured across seven analytical modules: preprocessing, differential gene expression analysis (DESeq2), exploratory analysis, visualization, gene set enrichment analysis (GSEA), downstream immune profiling (ESTIMATE, MCP-counter, TIDE, CIBERSORT, singscore, MSIsensor-RNA, immune checkpoint expression, and oncoPredict), and clinical analysis. A central configuration file (config.R) and a shared utility functions file (utils.R) are sourced by all R scripts to ensure consistency in parameters, colour schemes, and statistical testing across cohorts.

The complete analysis pipeline, including all scripts used for data processing, differential expression, and downstream analyses, is available at:
https://github.com/AmirpashaM7/Angiosarcoma

A.1 Repository Structure

The following directory tree shows the complete organisation of the analysis repository. All scripts are numbered in the order they should be executed.

```
angiosarcoma-rti-vs-nonrti/
├── README.md
├── config/
│   ├── config.R
│   └── utils.R
├── envs/
│   └── session_info.txt
└── scripts/
    ├── 01_preprocessing/
    │   ├── 01_load_internal_counts.R
    │   └── 02_load_external_tpm.R
    ├── 02_deseq2/
    │   ├── 01_deseq2_internal.R
    │   └── 02_deseq2_external.R
    ├── 03_exploratory/
    │   ├── 01_pca.R
    │   └── 02_hierarchical_clustering.R
    ├── 04_visualization/
```

```
    │   ├── 01_volcano.R
    │   └── 02_heatmap.R
    ├── 05_gsea/
    │   ├── 01_gsea.R
    │   └── 02_cross_cohort_concordance.R
    ├── 06_downstream/
    │   ├── 01_estimate.R
    │   ├── 02_mcp_counter.R
    │   ├── 03_tide.R
    │   ├── 04_cibersort.R
    │   ├── 05_apm_singscore.R
    │   ├── 06_msi.R
    │   ├── 07_checkpoints.R
    │   └── 08_oncopredict.R
    └── 07_clinical/
        └── 01_clinical.R
```

A.2 Configuration

config/config.R

Central configuration file: paths, parameters, group labels, and colour palette for all downstream scripts.

```
# ============================================================================
# config.R — Central configuration for the angiosarcoma RTI vs Non-RTI pipeline
# Edit BASE_DIR and DATA_DIR to match your environment before running any script.
# ============================================================================

# ── Working directory ────────────────────────────────────────────────────────
BASE_DIR   <- "~/angiosarcoma"          # server working directory
OUTPUT_DIR <- file.path(BASE_DIR, "output")
DATA_DIR   <- file.path(BASE_DIR, "data")

# ── Internal cohort data files ───────────────────────────────────────────────
INTERNAL_COUNTS_FILE   <- file.path(DATA_DIR, "internal_counts_clean.csv")
INTERNAL_METADATA_FILE <- file.path(DATA_DIR, "internal_metadata_clean.csv")
INTERNAL_TPM_LOG_FILE  <- file.path(DATA_DIR, "internal_tpm_log.csv")
DDS_FILE               <- file.path(DATA_DIR, "dds.rds")

# ── External cohort data files (cBioPortal / The Angiosarcoma Project) ─────
EXTERNAL_TPM_FILE          <- file.path(DATA_DIR, "data_mrna_seq_tpm.txt")
EXTERNAL_COUNTS_FILE       <- file.path(DATA_DIR, "data_mrna_seq_read_counts.txt")
EXTERNAL_CLINICAL_PATIENT  <- file.path(DATA_DIR, "data_clinical_patient.txt")
EXTERNAL_CLINICAL_SAMPLE   <- file.path(DATA_DIR, "data_clinical_sample.txt")
EXTERNAL_CLINICAL_55PT_FILE <- file.path(DATA_DIR,
"Final_Clinical_Data_QC_55_Patients.txt")

# ── GDSC2 training data for oncoPredict ──────────────────────────────────────
GDSC2_EXPR_FILE <- file.path(DATA_DIR, "GDSC2_Expr_RMA_Log.rds")
GDSC2_RES_FILE  <- file.path(DATA_DIR, "GDSC2_Res.rds")

# ── CIBERSORT files ──────────────────────────────────────────────────────────
CIBERSORT_R_FILE <- file.path(DATA_DIR, "CIBERSORT.R")
LM22_FILE        <- file.path(DATA_DIR, "LM22.txt")

# ── Internal clinical data (UDI-mapped) ──────────────────────────────────────
INTERNAL_CLINICAL_FILE <- file.path(DATA_DIR, "colData_32pts_full_clinical.csv")

# ── MSIsensor-RNA model ──────────────────────────────────────────────────────
MSI_MODEL_PATH <- file.path(BASE_DIR, "msisensor-rna/model/TCGA.MSIPopular.model.pkl")

# ── Analysis parameters ──────────────────────────────────────────────────────
PADJ_THRESHOLD  <- 0.05
LOG2FC_THRESHOLD <- 1.0
MIN_COUNT_FILTER <- 10        # minimum total counts per gene for DESeq2 pre-filter
MIN_LIBRARY_SIZE <- 1e6       # minimum library size for QC (external cohort)

# ── Group labels (as stored in metadata) ─────────────────────────────────────
# Internal: Sporadic_vs_Radiotherapy_induced column values
INTERNAL_RTI_LABEL    <- "RTi"
```

```
INTERNAL_NONRTI_LABEL <- "Sporadic"

# External: MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA column values
EXTERNAL_RTI_LABEL    <- "Yes"
EXTERNAL_NONRTI_LABEL <- "No"

# ── Plot colors ──────────────────────────────────────────────────────────
COL_RTI    <- "#C0392B"
COL_NONRTI <- "#1A7FA8"
GROUP_COLS <- c("Non-RTI" = COL_NONRTI, "RTI" = COL_RTI)

# Create output directory if it does not exist
if (!dir.exists(OUTPUT_DIR)) dir.create(OUTPUT_DIR, recursive = TRUE)
```

config/utils.R

Shared utility functions: boxplot helpers, Wilcoxon summary table, ICI bar chart, used across all downstream analyses.

```
# ==============================================================================
# utils.R — Shared plotting and statistics functions for the entire pipeline
#
# Source this file at the top of every downstream script:
#   source("config/config.R")
#   source("config/utils.R")
# ==============================================================================

# ── two-group boxplot (continuous score) ─────────────────────────────────
# Used by: 01_estimate, 02_mcp_counter, 03_tide, 05_signatures
boxplot_2grp <- function(df, y_col, subtitle) {
  rti     <- df[[y_col]][df$Group == "RTI"]
  non_rti <- df[[y_col]][df$Group == "Non-RTI"]
  p_val   <- wilcox.test(rti, non_rti)$p.value
  p_lab   <- ifelse(p_val < 0.001, "p < 0.001", paste0("p = ", round(p_val, 3)))
  n_lmap  <- setNames(
    paste0(c("Non-RTI", "RTI"), "\n(n=",
           c(sum(df$Group == "Non-RTI"), sum(df$Group == "RTI")), ")"),
    c("Non-RTI", "RTI"))
  ggplot(df, aes(x = Group, y = .data[[y_col]], fill = Group)) +
    geom_boxplot(outlier.shape = NA, alpha = 0.8, width = 0.5, linewidth = 0.7) +
    geom_jitter(width = 0.12, size = 1.8, alpha = 0.5, color = "black") +
    scale_fill_manual(values = GROUP_COLS) +
    scale_x_discrete(labels = n_lmap) +
    annotate("text", x = 1.5, y = max(df[[y_col]], na.rm = TRUE) * 1.08,
             label = p_lab, size = 4, fontface = "bold") +
    theme_classic(base_size = 11) +
    labs(title = y_col, subtitle = subtitle, x = NULL, y = "Score") +
    theme(legend.position = "none",
          plot.title = element_text(face = "bold"),
          axis.text  = element_text(color = "black", face = "bold"))
}

# ── small boxplot for many-panel layouts (CIBERSORT 22 cells, MCP 10 cells) ──
# Used by: 02_mcp_counter, 04_cibersort
boxplot_cell <- function(df, cell, subtitle) {
  rti     <- df[[cell]][df$Group == "RTI"]
  non_rti <- df[[cell]][df$Group == "Non-RTI"]
  p_val   <- wilcox.test(rti, non_rti)$p.value
  p_lab   <- ifelse(p_val < 0.001, "p < 0.001", paste0("p = ", round(p_val, 3)))
  n_lmap  <- setNames(
    paste0(c("Non-RTI", "RTI"), "\n(n=",
           c(sum(df$Group == "Non-RTI"), sum(df$Group == "RTI")), ")"),
    c("Non-RTI", "RTI"))
  ggplot(df, aes(x = Group, y = .data[[cell]], fill = Group)) +
    geom_boxplot(outlier.shape = NA, alpha = 0.8, width = 0.5, linewidth = 0.6) +
    geom_jitter(width = 0.12, size = 1.2, alpha = 0.5, color = "black") +
    scale_fill_manual(values = GROUP_COLS) +
    scale_x_discrete(labels = n_lmap) +
    annotate("text", x = 1.5, y = max(df[[cell]], na.rm = TRUE) * 1.1,
             label = p_lab, size = 3, fontface = "bold") +
    theme_classic(base_size = 9) +
```

```
    labs(title = cell, subtitle = subtitle, x = NULL, y = "Fraction") +
    theme(legend.position = "none",
          plot.title = element_text(face = "bold", size = 8),
          axis.text  = element_text(color = "black"))
}

# ── ICI responder bar chart (Fisher exact test on binary Responder column) ─────
# Used by: 03_tide
# Fisher exact test is used here — not Wilcoxon — because Responder is a
# binary yes/no outcome, not a continuous score. Fisher tests whether the
# proportion of predicted ICI responders differs between RTI and Non-RTI.
ici_bar <- function(df, subtitle) {
  rti_resp     <- sum(df$Responder[df$Group == "RTI"]     == TRUE,  na.rm = TRUE)
  rti_nonresp  <- sum(df$Responder[df$Group == "RTI"]     == FALSE, na.rm = TRUE)
  nrti_resp    <- sum(df$Responder[df$Group == "Non-RTI"] == TRUE,  na.rm = TRUE)
  nrti_nonresp <- sum(df$Responder[df$Group == "Non-RTI"] == FALSE, na.rm = TRUE)
  mat <- matrix(c(rti_resp, rti_nonresp, nrti_resp, nrti_nonresp), nrow = 2)
  ft  <- fisher.test(mat)
  p_lab <- ifelse(ft$p.value < 0.001, "Fisher p < 0.001",
                  paste0("Fisher p = ", round(ft$p.value, 3)))
  prop_df <- df %>%
    group_by(Group) %>%
    summarise(n = n(), Responders = sum(Responder == TRUE, na.rm = TRUE),
              Pct = Responders / n * 100, .groups = "drop") %>%
    mutate(Label = paste0(Group, "\n(n=", n, ")"),
           Label = factor(Label, levels = Label[order(Group)]))
  ggplot(prop_df, aes(x = Label, y = Pct, fill = Group)) +
    geom_col(width = 0.5, alpha = 0.9) +
    geom_text(aes(label = paste0(round(Pct, 1), "%")),
              vjust = -0.6, size = 4.5, fontface = "bold") +
    scale_fill_manual(values = GROUP_COLS) +
    annotate("text", x = 1.5, y = max(prop_df$Pct) + 10,
             label = p_lab, size = 4, fontface = "bold") +
    scale_y_continuous(limits = c(0, max(prop_df$Pct) + 18), expand = c(0, 0)) +
    theme_classic(base_size = 12) +
    labs(title = "Predicted ICI responders", subtitle = subtitle,
         x = NULL, y = "Responders (%)") +
    theme(legend.position = "none",
          plot.title = element_text(face = "bold"),
          axis.text  = element_text(color = "black", face = "bold"))
}

# ── Wilcoxon summary table (continuous scores) ──────────────────────────────
# Used by: 01_estimate, 02_mcp_counter, 04_cibersort, 05_signatures
# score_cols: character vector of column names to test
# tool:       string label e.g. "ESTIMATE", "MCPcounter", "CIBERSORT"
# cohort:     string label e.g. "Internal", "External"
wilcox_tbl <- function(df, score_cols, tool, cohort) {
  rows <- lapply(score_cols, function(col) {
    if (!col %in% colnames(df)) return(NULL)
    rti     <- na.omit(df[[col]][df$Group == "RTI"])
    non_rti <- na.omit(df[[col]][df$Group == "Non-RTI"])
    if (length(rti) < 3 || length(non_rti) < 3) return(NULL)
    w <- wilcox.test(rti, non_rti, exact = FALSE)
    data.frame(
      Analysis    = col,
      Tool        = tool,
      Cohort      = cohort,
      n_RTI       = length(rti),
      n_NonRTI    = length(non_rti),
      P_Value     = round(w$p.value, 4),
      Direction   = ifelse(mean(rti) > mean(non_rti),
                           "Higher in RTI", "Higher in Non-RTI"),
      RTI_Mean    = round(mean(rti), 4),
      NonRTI_Mean = round(mean(non_rti), 4),
      Significant = ifelse(w$p.value < 0.05, "Yes",
                           ifelse(w$p.value < 0.1, "Trend", "No"))
    )
  })
  tbl <- bind_rows(Filter(Negate(is.null), rows))
  tbl$FDR <- round(p.adjust(tbl$P_Value, method = "BH"), 4)
```

```
  tbl
}

# ── TIDE Wilcoxon table (continuous scores + Fisher ICI row) ──────────────────
# Used by: 03_tide only
# Extends wilcox_tbl with a Fisher exact test row for the binary Responder col
wilcox_tbl_tide <- function(df, cohort) {
  tbl <- wilcox_tbl(df, c("TIDE", "Dysfunction", "Exclusion", "CAF"),
                    "TIDE", cohort)
  if ("Responder" %in% colnames(df)) {
    rti_resp     <- sum(df$Responder[df$Group == "RTI"]     == TRUE,  na.rm = TRUE)
    rti_nonresp  <- sum(df$Responder[df$Group == "RTI"]     == FALSE, na.rm = TRUE)
    nrti_resp    <- sum(df$Responder[df$Group == "Non-RTI"] == TRUE,  na.rm = TRUE)
    nrti_nonresp <- sum(df$Responder[df$Group == "Non-RTI"] == FALSE, na.rm = TRUE)
    mat <- matrix(c(rti_resp, rti_nonresp, nrti_resp, nrti_nonresp), nrow = 2)
    ft  <- fisher.test(mat)
    ici_row <- data.frame(
      Analysis    = "ICI Responder Rate (%)",
      Tool        = "TIDE/Fisher",
      Cohort      = cohort,
      n_RTI       = rti_resp + rti_nonresp,
      n_NonRTI    = nrti_resp + nrti_nonresp,
      P_Value     = round(ft$p.value, 4),
      Direction   = ifelse(rti_resp / (rti_resp + rti_nonresp) >
                           nrti_resp / (nrti_resp + nrti_nonresp),
                           "Higher in RTI", "Higher in Non-RTI"),
      RTI_Mean    = round(rti_resp / (rti_resp + rti_nonresp) * 100, 1),
      NonRTI_Mean = round(nrti_resp / (nrti_resp + nrti_nonresp) * 100, 1),
      Significant = ifelse(ft$p.value < 0.05, "Yes",
                           ifelse(ft$p.value < 0.1, "Trend", "No")),
      FDR         = NA
    )
    tbl <- bind_rows(tbl, ici_row)
    tbl$FDR <- round(p.adjust(tbl$P_Value, method = "BH"), 4)
  }
  tbl
}
```

A.3 Preprocessing

scripts/01_preprocessing/01_load_internal_counts.R

Load raw count matrix for internal cohort, align with metadata, apply pre-filtering, and build DESeq2 object.

```
# ==============================================================================
# 01_load_internal_counts.R
# Load the internal raw count matrix, align with metadata, build DESeq2 object.
# Output: dds.rds saved to DATA_DIR
# ==============================================================================

source("config/config.R")

library(DESeq2)
library(dplyr)

# ── 1. Load counts ────────────────────────────────────────────────────────────
message("Loading internal count matrix...")
counts <- read.csv(INTERNAL_COUNTS_FILE, row.names = 1, check.names = FALSE)
message("  Dimensions: ", nrow(counts), " genes x ", ncol(counts), " samples")

# ── 2. Load metadata ──────────────────────────────────────────────────────────
message("Loading internal metadata...")
meta <- read.csv(INTERNAL_METADATA_FILE, row.names = 1)
message("  Groups: ", paste(table(meta$Sporadic_vs_Radiotherapy_induced), collapse = ",
"))

# ── 3. Align samples ──────────────────────────────────────────────────────────
common_samples <- intersect(colnames(counts), rownames(meta))
if (length(common_samples) == 0) stop("No matching sample IDs between counts and
metadata.")
counts <- counts[, common_samples]
meta   <- meta[common_samples, , drop = FALSE]

# ── 4. Pre-filter low-count genes ─────────────────────────────────────────────
keep <- rowSums(counts) >= MIN_COUNT_FILTER
counts <- counts[keep, ]
message("  Genes after pre-filter: ", nrow(counts))

# ── 5. Build DESeq2 object ────────────────────────────────────────────────────
meta$Sporadic_vs_Radiotherapy_induced <- factor(
  meta$Sporadic_vs_Radiotherapy_induced,
  levels = c(INTERNAL_NONRTI_LABEL, INTERNAL_RTI_LABEL)
)

dds <- DESeqDataSetFromMatrix(
  countData = round(as.matrix(counts)),
  colData   = meta,
  design    = ~ Sporadic_vs_Radiotherapy_induced
)

# ── 6. Estimate size factors (poscounts for zero-heavy RNA-seq data) ──────────
dds <- estimateSizeFactors(dds, type = "poscounts")

# ── 7. Save ───────────────────────────────────────────────────────────────────
saveRDS(dds, DDS_FILE)
message("Saved DESeq2 object: ", DDS_FILE)
```

scripts/01_preprocessing/02_load_external_tpm.R

Load external TPM data from The Angiosarcoma Project (cBioPortal), filter for breast angiosarcoma, apply QC, and export log2 TPM matrices.

```
# ==============================================================================
# 02_load_external_tpm.R
# Load the cBioPortal TPM expression file (The Angiosarcoma Project, April 2025)
# and filter for breast angiosarcoma samples (RTI vs Non-RTI).
# Outputs:
#   - Final_Clinical_Data_QC_55_Patients.txt  (after library-size QC)
```

```
#   - matrix_rti / matrix_non_rti             (log2 TPM matrices for downstream use)
# ============================================================================

source("config/config.R")

library(readr)
library(dplyr)
library(tibble)

# ── 1. Load clinical files ──────────────────────────────────────────────
message("Loading external clinical data...")
patient_data <- read_tsv(EXTERNAL_CLINICAL_PATIENT, comment = "#", show_col_types =
FALSE)
sample_data  <- read_tsv(EXTERNAL_CLINICAL_SAMPLE,  comment = "#", show_col_types =
FALSE)

# ── 2. Filter for breast angiosarcoma ──────────────────────────────────
breast_patients <- patient_data %>%
  filter(grepl("Breast", MRD_PRIMARY_SITE, ignore.case = TRUE))

rti_ids     <- breast_patients %>%
  filter(MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA == EXTERNAL_RTI_LABEL) %>%
  pull(PATIENT_ID)
non_rti_ids <- breast_patients %>%
  filter(MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA == EXTERNAL_NONRTI_LABEL) %>%
  pull(PATIENT_ID)

# Map patient IDs to sample IDs
rti_samples     <- sample_data %>% filter(PATIENT_ID %in% rti_ids)     %>%
pull(SAMPLE_ID)
non_rti_samples <- sample_data %>% filter(PATIENT_ID %in% non_rti_ids) %>%
pull(SAMPLE_ID)

# Keep one sample per patient (first T-sample)
clinical_filtered <- sample_data %>%
  filter(PATIENT_ID %in% c(rti_ids, non_rti_ids)) %>%
  group_by(PATIENT_ID) %>%
  slice_head(n = 1) %>%
  ungroup() %>%
  mutate(Group = ifelse(PATIENT_ID %in% rti_ids, "RTI", "Non-RTI"),
         MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA = ifelse(PATIENT_ID %in% rti_ids,
                                                        EXTERNAL_RTI_LABEL,
                                                        EXTERNAL_NONRTI_LABEL))

message("  Breast samples — RTI: ", sum(clinical_filtered$Group == "RTI"),
        "  Non-RTI: ", sum(clinical_filtered$Group == "Non-RTI"))

# ── 3. Load expression data ─────────────────────────────────────────────
message("Loading TPM expression data (this may take a minute)...")
expression_data <- read_tsv(EXTERNAL_TPM_FILE, show_col_types = FALSE)

# ── 4. Library-size QC ──────────────────────────────────────────────────
valid_samples <- clinical_filtered$SAMPLE_ID
valid_cols    <- intersect(colnames(expression_data), valid_samples)

expr_num <- expression_data %>%
  select(all_of(valid_cols)) %>%
  as.matrix()

lib_sizes <- colSums(expr_num, na.rm = TRUE)
good_cols <- names(lib_sizes[lib_sizes >= MIN_LIBRARY_SIZE])
message("  Samples passing QC (≥", MIN_LIBRARY_SIZE, " reads): ", length(good_cols))

clinical_qc <- clinical_filtered %>% filter(SAMPLE_ID %in% good_cols)

# ── 5. Build clean expression matrices (log2 TPM) ──────────────────────
clean_expression_data <- function(df, group_ids) {
  df %>%
    select(Hugo_Symbol, all_of(group_ids)) %>%
    group_by(Hugo_Symbol) %>%
    summarise(across(where(is.numeric), mean), .groups = "drop") %>%
```

```
    column_to_rownames("Hugo_Symbol") %>%
    as.matrix() %>%
    { log2(. + 1) }
}

rti_cols     <- intersect(colnames(expression_data),
                          clinical_qc$SAMPLE_ID[clinical_qc$Group == "RTI"])
non_rti_cols <- intersect(colnames(expression_data),
                          clinical_qc$SAMPLE_ID[clinical_qc$Group == "Non-RTI"])

matrix_rti     <- clean_expression_data(expression_data, rti_cols)
matrix_non_rti <- clean_expression_data(expression_data, non_rti_cols)

message("  RTI matrix:     ", nrow(matrix_rti), " genes x ", ncol(matrix_rti), "
samples")
message("  Non-RTI matrix: ", nrow(matrix_non_rti), " genes x ", ncol(matrix_non_rti),
" samples")

# ── 6. Save ──────────────────────────────────────────────────────────────
write.table(clinical_qc, EXTERNAL_CLINICAL_55PT_FILE,
            sep = "\t", row.names = FALSE, quote = FALSE)

saveRDS(matrix_rti,     file.path(DATA_DIR, "matrix_rti.rds"))
saveRDS(matrix_non_rti, file.path(DATA_DIR, "matrix_non_rti.rds"))
saveRDS(clinical_qc,    file.path(DATA_DIR, "external_clinical_qc.rds"))

message("Saved: ", EXTERNAL_CLINICAL_55PT_FILE)
message("Saved: matrix_rti.rds, matrix_non_rti.rds, external_clinical_qc.rds")
```

A.4 Differential Expression Analysis

scripts/02_deseq2/01_deseq2_internal.R

DESeq2 differential expression for internal cohort. Produces raw Wald stat results for GSEA and apeglm-shrunken results for visualization.

```
# ==============================================================================
# 01_deseq2_internal.R
# DESeq2 differential expression — internal cohort (RTi vs Sporadic)
#
# Saves TWO result tables:
#   DEG_internal_raw.rds      → raw Wald stat  → used by GSEA
#   DEG_internal_shrunken.rds → apeglm LFC     → used by volcano + heatmap
#
# CSV outputs (shrunken, for reporting):
#   DEG_Internal_All.csv
#   DEG_Internal_Upregulated_RTI.csv
# ==============================================================================

source("config/config.R")

library(DESeq2)
library(AnnotationDbi)
library(org.Hs.eg.db)
library(dplyr)
library(tibble)

# ── 1. Load dds ──────────────────────────────────────────────────────────
message("Loading DESeq2 object...")
dds <- readRDS(DDS_FILE)
message("  ", nrow(dds), " genes x ", ncol(dds), " samples")

dds$Sporadic_vs_Radiotherapy_induced <- factor(
  dds$Sporadic_vs_Radiotherapy_induced,
  levels = c(INTERNAL_NONRTI_LABEL, INTERNAL_RTI_LABEL)
  #           "Sporadic" = reference, "RTi" = numerator
  #           so LFC > 0 means higher in RTI
)

# ── 2. Run DESeq2 ────────────────────────────────────────────────────────
```

```
message("Running DESeq2...")
dds <- DESeq(dds)
coef_name <- grep(INTERNAL_RTI_LABEL, resultsNames(dds), value = TRUE)
message("  Using coef: ", coef_name)

# ── 3. RAW results — Wald stat for GSEA ranking ──────────────────────────
# Do NOT use lfcShrink here.
# GSEA ranks genes by the stat column.
# lfcShrink changes the stat, so using it would make the internal
# and external rankings incomparable.
message("Extracting raw Wald stat results (for GSEA)...")

res_raw <- results(
  dds,
  contrast = c("Sporadic_vs_Radiotherapy_induced",
               INTERNAL_RTI_LABEL,
               INTERNAL_NONRTI_LABEL)
) %>%
  as.data.frame() %>%
  rownames_to_column("Gene") %>%
  filter(!is.na(stat))   # stat is what GSEA needs

res_raw$Symbol <- mapIds(org.Hs.eg.db,
                         keys      = res_raw$Gene,
                         column    = "SYMBOL",
                         keytype   = "ENSEMBL",
                         multiVals = "first")

saveRDS(res_raw, file.path(DATA_DIR, "DEG_internal_raw.rds"))
message("  Saved DEG_internal_raw.rds  (", nrow(res_raw), " genes)")
message("  -> 01_gsea.R will use the stat column from this file")

# ── 4. SHRUNKEN results — apeglm, for volcano plot + heatmap ─────────────
# apeglm shrinks noisy fold changes for low-count genes toward zero,
# giving a cleaner visual. Only use this for plots, not for GSEA.
message("Applying apeglm LFC shrinkage (for volcano + heatmap)...")

res_shrink <- lfcShrink(dds, coef = coef_name, type = "apeglm") %>%
  as.data.frame() %>%
  rownames_to_column("Gene") %>%
  filter(!is.na(padj))

res_shrink$Symbol <- mapIds(org.Hs.eg.db,
                            keys      = res_shrink$Gene,
                            column    = "SYMBOL",
                            keytype   = "ENSEMBL",
                            multiVals = "first")

res_shrink <- res_shrink %>%
  mutate(Expression = case_when(
    padj < PADJ_THRESHOLD & log2FoldChange >  LOG2FC_THRESHOLD ~ "Upregulated in RTI",
    padj < PADJ_THRESHOLD & log2FoldChange < -LOG2FC_THRESHOLD ~ "Upregulated in
Non-RTI",
    TRUE ~ "Not Significant"
  ))

saveRDS(res_shrink, file.path(DATA_DIR, "DEG_internal_shrunken.rds"))
message("  Saved DEG_internal_shrunken.rds")
message("  -> 01_volcano.R and 02_heatmap.R will use this file")

# ── 5. Save dds for downstream scripts ──────────────────────────────────
saveRDS(dds, file.path(DATA_DIR, "dds_internal_run.rds"))

# ── 6. CSV outputs for reporting ─────────────────────────────────────────
upregulated_rti <- res_shrink %>%
  filter(Expression == "Upregulated in RTI") %>%
  arrange(padj)

message("  Upregulated in RTI: ", nrow(upregulated_rti),
        " genes (padj<", PADJ_THRESHOLD, ", log2FC>", LOG2FC_THRESHOLD, ")")
```

```
write.csv(res_shrink,
          file.path(OUTPUT_DIR, "DEG_Internal_All.csv"),
          row.names = FALSE)
write.csv(upregulated_rti,
          file.path(OUTPUT_DIR, "DEG_Internal_Upregulated_RTI.csv"),
          row.names = FALSE)

message("Done: 01_deseq2_internal.R")
```

scripts/02_deseq2/02_deseq2_external. R

DESeq2 differential expression for external cohort (55 QC-passed patients from The Angiosarcoma Project).

```
# ==============================================================================
# 02_deseq2_external.R
# DESeq2 differential expression — external cohort (55 QC-passed patients)
# Input: data_mrna_seq_read_counts.txt (raw counts from cBioPortal)
#
# Saves TWO result tables:
#   DEG_external_raw.rds      → raw Wald stat  → used by GSEA
#   DEG_external_shrunken.rds → apeglm LFC     → used by volcano + heatmap
#
# CSV outputs (shrunken, for reporting):
#   DEG_External_All.csv
#   DEG_External_Upregulated_RTI.csv
#
# NOTE: The factor levels are set as c("No", "Yes") — "No" (Non-RTI) is the
# reference (denominator), "Yes" (RTI) is the numerator, so:
#   LFC > 0 = higher in RTI
#   LFC < 0 = higher in Non-RTI
# ==============================================================================

source("config/config.R")

library(DESeq2)
library(dplyr)
library(tibble)

# ── 1. Load clinical QC data ──────────────────────────────────────────────────
message("Loading 55-patient clinical data...")
clin_qc <- readRDS(file.path(DATA_DIR, "external_clinical_qc.rds"))

clin_qc$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA <- factor(
  clin_qc$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA,
  levels = c(EXTERNAL_NONRTI_LABEL, EXTERNAL_RTI_LABEL)
  #           "No" = reference,     "Yes" = numerator
)

message("  RTI: ",    sum(clin_qc$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA ==
EXTERNAL_RTI_LABEL),
        "  Non-RTI: ", sum(clin_qc$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA ==
EXTERNAL_NONRTI_LABEL))

# ── 2. Load and align count matrix ────────────────────────────────────────────
message("Loading count matrix: ", EXTERNAL_COUNTS_FILE)
count_raw <- read.delim(EXTERNAL_COUNTS_FILE, check.names = FALSE)

# First column = Hugo_Symbol, second = Entrez_Gene_Id — drop both
gene_names <- make.unique(as.character(count_raw[, 1]))
rownames(count_raw) <- gene_names
count_data <- count_raw[, -c(1, 2)]

# Keep only the 55 QC-passed samples, in the same order as clin_qc
valid_cols   <- intersect(clin_qc$SAMPLE_ID, colnames(count_data))
count_subset <- count_data[, valid_cols]
clin_subset  <- clin_qc[match(valid_cols, clin_qc$SAMPLE_ID), ]
rownames(clin_subset) <- clin_subset$SAMPLE_ID

message("  Samples matched: ", length(valid_cols))
```

```
# ── 3. Pre-filter low-count genes ──────────────────────────────────────
keep <- rowSums(count_subset) >= MIN_COUNT_FILTER
count_subset <- count_subset[keep, ]
message("  Genes after pre-filter (rowSums >= ", MIN_COUNT_FILTER, "): ",
nrow(count_subset))

# ── 4. Build DESeq2 object ─────────────────────────────────────────────
dds_ext <- DESeqDataSetFromMatrix(
  countData = round(as.matrix(count_subset)),
  colData   = clin_subset,
  design    = ~ MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA
)
dds_ext <- estimateSizeFactors(dds_ext, type = "poscounts")

# ── 5. Run DESeq2 ──────────────────────────────────────────────────────
message("Running DESeq2 (external)...")
dds_ext <- DESeq(dds_ext)
coef_name <- grep(EXTERNAL_RTI_LABEL, resultsNames(dds_ext), value = TRUE)
message("  Using coef: ", coef_name)

# ── 6. RAW results — Wald stat for GSEA ranking ────────────────────────
# Same approach as internal: use raw stat for GSEA so both cohorts
# are ranked by the same metric (raw Wald stat, not apeglm stat).
message("Extracting raw Wald stat results (for GSEA)...")

res_raw <- results(
  dds_ext,
  contrast = c("MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA",
               EXTERNAL_RTI_LABEL,
               EXTERNAL_NONRTI_LABEL)
) %>%
  as.data.frame() %>%
  rownames_to_column("Gene") %>%
  filter(!is.na(stat))

saveRDS(res_raw, file.path(DATA_DIR, "DEG_external_raw.rds"))
message("  Saved DEG_external_raw.rds  (", nrow(res_raw), " genes)")
message("  -> 01_gsea.R will use the stat column from this file")

# ── 7. SHRUNKEN results — apeglm, for volcano + heatmap ────────────────
message("Applying apeglm LFC shrinkage (for volcano + heatmap)...")

res_shrink <- lfcShrink(dds_ext, coef = coef_name, type = "apeglm") %>%
  as.data.frame() %>%
  rownames_to_column("Gene") %>%
  filter(!is.na(padj)) %>%
  mutate(Expression = case_when(
    padj < PADJ_THRESHOLD & log2FoldChange >  LOG2FC_THRESHOLD ~ "Upregulated in RTI",
    padj < PADJ_THRESHOLD & log2FoldChange < -LOG2FC_THRESHOLD ~ "Upregulated in
Non-RTI",
    TRUE ~ "Not Significant"
  ))

saveRDS(res_shrink, file.path(DATA_DIR, "DEG_external_shrunken.rds"))
message("  Saved DEG_external_shrunken.rds")
message("  -> 01_volcano.R and 02_heatmap.R will use this file")

# ── 8. Save dds for downstream ─────────────────────────────────────────
saveRDS(dds_ext, file.path(DATA_DIR, "dds_ext_run.rds"))

# ── 9. CSV outputs for reporting ───────────────────────────────────────
upregulated_rti <- res_shrink %>%
  filter(Expression == "Upregulated in RTI") %>%
  arrange(padj)

message("  Upregulated in RTI: ", nrow(upregulated_rti),
        " genes (padj<", PADJ_THRESHOLD, ", log2FC>", LOG2FC_THRESHOLD, ")")

write.csv(res_shrink,
          file.path(OUTPUT_DIR, "DEG_External_All.csv"),
          row.names = FALSE)
```

```
write.csv(upregulated_rti,
          file.path(OUTPUT_DIR, "DEG_External_Upregulated_RTI.csv"),
          row.names = FALSE)

message("Done: 02_deseq2_external.R")
```

A.5 Exploratory Analysis

scripts/03_exploratory/01_pca. R

Principal component analysis on VST-normalised expression for internal and external cohorts.

```
# ============================================================================
# 01_pca.R
# PCA on VST-normalized counts for internal and external cohorts.
# Outputs: PCA_Internal.pdf, PCA_External.pdf
# ============================================================================

source("config/config.R")

library(DESeq2)
library(ggplot2)
library(patchwork)
library(dplyr)

# ── Helper: build PCA plot ──────────────────────────────────────────
make_pca_plot <- function(vsd_mat, group_vec, pc_x = 1, pc_y = 2, title = "") {
  pca_res    <- prcomp(t(vsd_mat))
  percentVar <- round(100 * pca_res$sdev^2 / sum(pca_res$sdev^2), 1)
  pca_data   <- data.frame(pca_res$x, Group = group_vec)

  ggplot(pca_data, aes(x = .data[[paste0("PC", pc_x)]],
                       y = .data[[paste0("PC", pc_y)]],
                       color = Group)) +
    geom_point(size = 3, alpha = 0.85) +
    scale_color_manual(values = GROUP_COLS) +
    theme_classic(base_size = 12) +
    labs(
      title = title,
      x = paste0("PC", pc_x, " (", percentVar[pc_x], "%)"),
      y = paste0("PC", pc_y, " (", percentVar[pc_y], "%)")
    ) +
    theme(legend.position = "top", legend.title = element_blank())
}

# ═══════════════════════════════════════════════════════════════════════
# INTERNAL COHORT
# ═══════════════════════════════════════════════════════════════════════
message("Running PCA — internal cohort...")

dds <- readRDS(DDS_FILE)
dds <- estimateSizeFactors(dds, type = "poscounts")
vsd <- vst(dds, blind = FALSE)
vsd_mat <- assay(vsd)

group_int <- ifelse(dds$Sporadic_vs_Radiotherapy_induced == INTERNAL_RTI_LABEL,
                    "RTI", "Non-RTI")

p1 <- make_pca_plot(vsd_mat, group_int, 1, 2, "PCA: PC1 vs PC2 — Internal")
p2 <- make_pca_plot(vsd_mat, group_int, 2, 3, "PCA: PC2 vs PC3 — Internal")

ggsave(file.path(OUTPUT_DIR, "PCA_Internal.pdf"),
       p1 + p2, width = 14, height = 6)
message("  Saved: PCA_Internal.pdf")

# ═══════════════════════════════════════════════════════════════════════
# EXTERNAL COHORT (55 patients)
# ═══════════════════════════════════════════════════════════════════════
message("Running PCA — external cohort...")
```

```
dds_ext <- readRDS(file.path(DATA_DIR, "dds_ext_deseq_run.rds"))
vsd_ext <- vst(dds_ext, blind = FALSE)
vsd_mat_ext <- assay(vsd_ext)

group_ext <- ifelse(dds_ext$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA ==
EXTERNAL_RTI_LABEL,
                    "RTI", "Non-RTI")

p3 <- make_pca_plot(vsd_mat_ext, group_ext, 1, 2, "PCA: PC1 vs PC2 — External")
p4 <- make_pca_plot(vsd_mat_ext, group_ext, 2, 3, "PCA: PC2 vs PC3 — External")

ggsave(file.path(OUTPUT_DIR, "PCA_External.pdf"),
       p3 + p4, width = 14, height = 6)
message("  Saved: PCA_External.pdf")

# Save VST matrices for downstream use
saveRDS(vsd_mat,     file.path(DATA_DIR, "vsd_mat_internal.rds"))
saveRDS(vsd_mat_ext, file.path(DATA_DIR, "vsd_mat_external.rds"))
```

scripts/03_exploratory/02_hierarchical_clustering.R

Hierarchical clustering (Ward D2) on VST-normalised expression with coloured dendrograms.

```
# ==============================================================================
# 02_hierarchical_clustering.R
# Hierarchical clustering on VST-normalized expression for both cohorts.
# Outputs: Dendrogram_Internal.pdf, Dendrogram_External.pdf
# ==============================================================================

source("config/config.R")

library(DESeq2)
library(dendextend)
library(ggplot2)

# ── Helper: colored dendrogram ────────────────────────────────────────────────
save_dendrogram <- function(vsd_mat, group_vec, out_file, title) {
  sample_dists <- dist(t(vsd_mat))
  hc    <- hclust(sample_dists, method = "ward.D2")
  dend  <- as.dendrogram(hc)

  label_colors <- ifelse(group_vec[labels(dend)] == "RTI", COL_RTI, COL_NONRTI)
  labels_colors(dend) <- label_colors

  pdf(out_file, width = 18, height = 8)
  plot(dend, main = title, ylab = "Distance")
  legend("topright",
         legend = c("RTI", "Non-RTI"),
         fill   = c(COL_RTI, COL_NONRTI),
         border = "black",
         box.lty = 1, cex = 1.2)
  dev.off()
  message("  Saved: ", out_file)
}

# ── Internal ──────────────────────────────────────────────────────────────────
message("Hierarchical clustering — internal cohort...")

vsd_mat <- readRDS(file.path(DATA_DIR, "vsd_mat_internal.rds"))
dds     <- readRDS(DDS_FILE)
group_int <- setNames(
  ifelse(dds$Sporadic_vs_Radiotherapy_induced == INTERNAL_RTI_LABEL, "RTI", "Non-RTI"),
  colnames(vsd_mat)
)

save_dendrogram(
  vsd_mat, group_int,
  file.path(OUTPUT_DIR, "Dendrogram_Internal.pdf"),
  "Hierarchical Clustering — Internal Cohort (n=32)"
```

```
)

# ── External ──────────────────────────────────────────────────────────
message("Hierarchical clustering — external cohort...")

vsd_mat_ext <- readRDS(file.path(DATA_DIR, "vsd_mat_external.rds"))
clin_qc     <- readRDS(file.path(DATA_DIR, "external_clinical_qc.rds"))
group_ext <- setNames(
  ifelse(clin_qc$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA == EXTERNAL_RTI_LABEL, "RTI",
"Non-RTI"),
  clin_qc$SAMPLE_ID
)
# Align to colnames of vsd_mat_ext
group_ext <- group_ext[colnames(vsd_mat_ext)]

save_dendrogram(
  vsd_mat_ext, group_ext,
  file.path(OUTPUT_DIR, "Dendrogram_External.pdf"),
  "Hierarchical Clustering — External Cohort (n=55)"
)
```

A.6 Visualization

scripts/04_visualization/01_volcano.R

Volcano plots of apeglm-shrunken DESeq2 results for internal and external cohorts.

```
# ============================================================================
# 01_volcano.R
# Volcano plots for internal and external DESeq2 results.
# Labels the top 20 most significant genes per cohort.
# Outputs: Volcano_Internal.pdf, Volcano_External.pdf
# ============================================================================

source("config/config.R")

library(ggplot2)
library(ggrepel)
library(dplyr)

# ── Helper ────────────────────────────────────────────────────────────
make_volcano <- function(res_df, gene_col, title, out_file, n_labels = 20) {

  # Colour categories
  res_df <- res_df %>%
    mutate(Gene_Type = case_when(
      padj < PADJ_THRESHOLD & log2FoldChange >  LOG2FC_THRESHOLD ~ "Upregulated in
RTI",
      padj < PADJ_THRESHOLD & log2FoldChange < -LOG2FC_THRESHOLD ~ "Upregulated in
Non-RTI",
      TRUE ~ "Not Significant"
    ))

  # Top genes for labelling
  sig_genes <- res_df %>%
    filter(Gene_Type != "Not Significant") %>%
    arrange(padj)

  top_genes <- head(sig_genes[[gene_col]], n_labels)
  res_df$Label <- ifelse(res_df[[gene_col]] %in% top_genes, res_df[[gene_col]], NA)

  p <- ggplot(res_df, aes(x = log2FoldChange, y = -log10(padj), color = Gene_Type)) +
    geom_point(alpha = 0.8, size = 1.5) +
    geom_text_repel(aes(label = Label), size = 3.5, color = "black",
                    max.overlaps = 30, box.padding = 0.5, na.rm = TRUE) +
    scale_color_manual(values = c(
      "Not Significant"      = "grey80",
      "Upregulated in RTI"   = COL_RTI,
      "Upregulated in Non-RTI" = COL_NONRTI
    )) +
    geom_vline(xintercept = c(-LOG2FC_THRESHOLD, LOG2FC_THRESHOLD),
               linetype = "dashed", color = "black") +
```

```
    geom_hline(yintercept = -log10(PADJ_THRESHOLD),
               linetype = "dashed", color = "black") +
    theme_classic(base_size = 12) +
    labs(title = title,
         x = "Log2 Fold Change (RTI vs Non-RTI)",
         y = expression(-log[10](padj)),
         color = NULL) +
    theme(legend.position = "bottom")

  ggsave(out_file, plot = p, width = 8, height = 6)
  message("  Saved: ", out_file)
  invisible(p)
}

# ── Internal ──────────────────────────────────────────────────────────────
message("Generating internal volcano plot...")
res_int <- read.csv(file.path(OUTPUT_DIR, "DEG_Internal_All.csv"))
res_int$Symbol[is.na(res_int$Symbol)] <- res_int$Gene[is.na(res_int$Symbol)]

make_volcano(
  res_int, "Symbol",
  "Volcano Plot — Internal Cohort (RTI vs Non-RTI, apeglm-shrunken)",
  file.path(OUTPUT_DIR, "Volcano_Internal.pdf")
)

# ── External ──────────────────────────────────────────────────────────────
message("Generating external volcano plot...")
res_ext <- read.csv(file.path(OUTPUT_DIR, "DEG_External_All.csv"))

make_volcano(
  res_ext, "Gene",
  "Volcano Plot — External Cohort (RTI vs Non-RTI, 55 patients)",
  file.path(OUTPUT_DIR, "Volcano_External.pdf")
)
```

scripts/04_visualization/02_heatmap. R

Heatmaps of top 50 differentially expressed genes using ComplexHeatmap with VST row Z-scores.

```
# ==============================================================================
# 02_heatmap.R
# Heatmaps of top 50 RTI-upregulated and top 50 Non-RTI-upregulated genes
# for internal and external cohorts, using VST row Z-scores.
# Uses ComplexHeatmap for publication figures.
# Outputs: Heatmap_Internal_Top50.pdf, Heatmap_External_Top50.pdf
# ==============================================================================

source("config/config.R")

library(DESeq2)
library(ComplexHeatmap)
library(circlize)
library(dplyr)
library(tibble)
library(org.Hs.eg.db)
library(AnnotationDbi)

# ── Helper ────────────────────────────────────────────────────────────────
draw_heatmap <- function(mat, annotation_df, top_genes, title, out_file,
                         col_order = NULL) {

  # Row Z-score
  z_mat <- t(scale(t(mat[top_genes, ])))

  # Column annotation
  col_ann <- HeatmapAnnotation(
    Group = annotation_df$Group,
    col   = list(Group = c("RTI" = COL_RTI, "Non-RTI" = COL_NONRTI)),
    show_legend    = TRUE,
    show_annotation_name = FALSE
```

```
  )

  # Colour scale
  col_fun <- colorRamp2(c(-2, 0, 2), c("#1A7FA8", "white", "#C0392B"))

  ht <- Heatmap(
    z_mat,
    name              = "Row Z-score",
    col               = col_fun,
    top_annotation    = col_ann,
    show_column_names = FALSE,
    row_names_gp      = gpar(fontsize = 7),
    cluster_rows      = FALSE,
    cluster_columns   = TRUE,
    column_split      = annotation_df$Group,
    row_title         = NULL,
    column_title      = title,
    column_title_gp   = gpar(fontsize = 14, fontface = "bold"),
    heatmap_legend_param = list(title_gp = gpar(fontsize = 10))
  )

  pdf(out_file, width = 14, height = 12)
  draw(ht)
  dev.off()
  message("  Saved: ", out_file)
}

# ============================================================================
# INTERNAL COHORT
# ============================================================================
message("Building internal heatmap...")

vsd_mat <- readRDS(file.path(DATA_DIR, "vsd_mat_internal.rds"))
dds     <- readRDS(DDS_FILE)

res_int <- read.csv(file.path(OUTPUT_DIR, "DEG_Internal_All.csv"))
res_int$Symbol[is.na(res_int$Symbol)] <- res_int$Gene[is.na(res_int$Symbol)]

sig_int <- res_int %>% filter(padj < PADJ_THRESHOLD)

top25_rti_int     <- sig_int %>% filter(log2FoldChange > 0) %>%
  arrange(desc(log2FoldChange)) %>% slice_head(n = 25) %>% pull(Gene)
top25_nonrti_int  <- sig_int %>% filter(log2FoldChange < 0) %>%
  arrange(log2FoldChange) %>% slice_head(n = 25) %>% pull(Gene)
top50_int <- c(top25_rti_int, top25_nonrti_int)

# Annotate gene symbols for row labels
sym_map <- setNames(res_int$Symbol, res_int$Gene)
rownames(vsd_mat)[rownames(vsd_mat) %in% names(sym_map)] <-
  sym_map[rownames(vsd_mat)[rownames(vsd_mat) %in% names(sym_map)]]

top50_int_sym <- sym_map[top50_int]
top50_int_sym[is.na(top50_int_sym)] <- top50_int[is.na(top50_int_sym)]

ann_int <- data.frame(
  Group = ifelse(dds$Sporadic_vs_Radiotherapy_induced == INTERNAL_RTI_LABEL,
                 "RTI", "Non-RTI"),
  row.names = colnames(vsd_mat)
)

# Keep only genes present in vsd_mat after symbol renaming
top50_present <- top50_int_sym[top50_int_sym %in% rownames(vsd_mat)]

draw_heatmap(
  vsd_mat, ann_int, top50_present,
  "Top 50 DEGs: RTI vs Non-RTI (Internal Cohort)",
  file.path(OUTPUT_DIR, "Heatmap_Internal_Top50.pdf")
)

# ============================================================================
# EXTERNAL COHORT
```

```
# ============================================================================
message("Building external heatmap...")

vsd_mat_ext <- readRDS(file.path(DATA_DIR, "vsd_mat_external.rds"))
clin_qc     <- readRDS(file.path(DATA_DIR, "external_clinical_qc.rds"))

res_ext <- read.csv(file.path(OUTPUT_DIR, "DEG_External_All.csv"))
sig_ext <- res_ext %>% filter(padj < PADJ_THRESHOLD)

top25_rti_ext    <- sig_ext %>% filter(log2FoldChange > 0) %>%
  arrange(desc(log2FoldChange)) %>% slice_head(n = 25) %>% pull(Gene)
top25_nonrti_ext <- sig_ext %>% filter(log2FoldChange < 0) %>%
  arrange(log2FoldChange) %>% slice_head(n = 25) %>% pull(Gene)
top50_ext <- c(top25_rti_ext, top25_nonrti_ext)
top50_ext <- top50_ext[top50_ext %in% rownames(vsd_mat_ext)]

ann_ext <- data.frame(
  Group = ifelse(clin_qc$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA == EXTERNAL_RTI_LABEL,
                 "RTI", "Non-RTI"),
  row.names = clin_qc$SAMPLE_ID
)[colnames(vsd_mat_ext), , drop = FALSE]

draw_heatmap(
  vsd_mat_ext, ann_ext, top50_ext,
  "Top 50 DEGs: RTI vs Non-RTI (External Cohort, n=55)",
  file.path(OUTPUT_DIR, "Heatmap_External_Top50.pdf")
)
```

A.7 Gene Set Enrichment Analysis (GSEA)

scripts/05_gsea/01_gsea.R

GSEA with MSigDB Hallmark gene sets using raw Wald statistics from DESeq2 for both cohorts.

```
# ============================================================================
# 01_gsea.R — GSEA with Hallmark gene sets
#
# Uses raw Wald stat from DESeq2 results (not apeglm) for both cohorts,
# so the gene ranking metric is identical and the cross-cohort NES
# comparison is fair.
#
# Inputs:
#   DEG_internal_raw.rds   (from 01_deseq2_internal.R)
#   DEG_external_raw.rds   (from 02_deseq2_external.R)
#
# Outputs per cohort:
#   GSEA_Hallmark_Internal.csv / GSEA_Hallmark_External.csv
#   GSEA_Ridgeplot_Internal.pdf / GSEA_Ridgeplot_External.pdf
#   GSEA_Dotplot_Internal.pdf   / GSEA_Dotplot_External.pdf
# ============================================================================

source("config/config.R")

library(clusterProfiler)
library(enrichplot)
library(msigdbr)
library(org.Hs.eg.db)
library(AnnotationDbi)
library(dplyr)
library(tibble)
library(ggplot2)

# Hallmark gene sets (symbols — works for both cohorts)
hallmark_t2g <- msigdbr(species = "Homo sapiens", category = "H") %>%
  dplyr::select(gs_name, gene_symbol)

# ── helper: build ranked gene list from raw DESeq2 results ──────────────────
# internal: Gene column = Ensembl ID → convert to symbol first
# external: Gene column = Hugo symbol already
```

```
build_ranked_list <- function(res_df, gene_col, id_type = "SYMBOL") {

  if (id_type == "ENSEMBL") {
    res_df$Symbol <- mapIds(org.Hs.eg.db,
                            keys      = res_df[[gene_col]],
                            column    = "SYMBOL",
                            keytype   = "ENSEMBL",
                            multiVals = "first")
    res_df <- res_df %>% filter(!is.na(Symbol), Symbol != "")
    gene_col <- "Symbol"
  }

  res_df <- res_df %>%
    filter(!is.na(stat)) %>%
    arrange(desc(stat))

  gene_list <- setNames(res_df$stat, res_df[[gene_col]])
  gene_list[!duplicated(names(gene_list))]
}

# ── helper: run GSEA and save outputs ──────────────────────────────────────
run_and_save <- function(gene_list, label) {
  message("  Running GSEA: ", label, "  (", length(gene_list), " genes)")

  set.seed(42)
  gsea_res <- GSEA(
    geneList     = gene_list,
    TERM2GENE    = hallmark_t2g,
    pvalueCutoff = 1,        # keep all pathways; filter by padj when needed
    minGSSize    = 15,
    maxGSSize    = 500,
    verbose      = FALSE
  )

  gsea_df <- as.data.frame(gsea_res) %>% arrange(pval)
  n_sig   <- sum(gsea_df$p.adjust < 0.05, na.rm = TRUE)
  message("  Significant pathways (padj<0.05): ", n_sig)

  # Ridgeplot — top 10
  p_ridge <- ridgeplot(gsea_res, showCategory = 10) +
    labs(title    = paste("GSEA Ridgeplot —", label),
         subtitle = "Hallmark gene sets | ranked by raw Wald statistic",
         x        = "Enrichment distribution") +
    theme_classic(base_size = 12) +
    theme(plot.title = element_text(face = "bold"))

  ggsave(file.path(OUTPUT_DIR, paste0("GSEA_Ridgeplot_", label, ".pdf")),
         p_ridge, width = 12, height = 8)

  # Dotplot — top 10 activated + top 10 suppressed (padj < 0.25)
  act  <- gsea_df %>% filter(NES > 0, p.adjust < 0.25) %>%
    arrange(p.adjust) %>% slice_head(n = 10) %>% pull(ID)
  supp <- gsea_df %>% filter(NES < 0, p.adjust < 0.25) %>%
    arrange(p.adjust) %>% slice_head(n = 10) %>% pull(ID)

  if (length(c(act, supp)) > 0) {
    p_dot <- dotplot(gsea_res, showCategory = c(act, supp), split = ".sign") +
      facet_grid(~ .sign) +
      labs(title    = paste("GSEA Dotplot —", label),
           subtitle = "Activated vs suppressed in RTI") +
      theme_classic(base_size = 11) +
      theme(plot.title = element_text(face = "bold"))

    ggsave(file.path(OUTPUT_DIR, paste0("GSEA_Dotplot_", label, ".pdf")),
           p_dot, width = 14, height = 10)
  }

  out_csv <- file.path(OUTPUT_DIR, paste0("GSEA_Hallmark_", label, ".csv"))
  write.csv(gsea_df, out_csv, row.names = FALSE)
  message("  Saved: GSEA_Hallmark_", label, ".csv")
```

```
  invisible(gsea_res)
}

# ============================================================================
# INTERNAL — Ensembl IDs → convert to symbol
# ============================================================================
message("=== GSEA: Internal cohort ===")

res_int      <- readRDS(file.path(DATA_DIR, "DEG_internal_raw.rds"))
gene_list_int <- build_ranked_list(res_int, "Gene", id_type = "ENSEMBL")
gsea_int      <- run_and_save(gene_list_int, "Internal")

# ============================================================================
# EXTERNAL — already Hugo symbols
# ============================================================================
message("=== GSEA: External cohort ===")

res_ext      <- readRDS(file.path(DATA_DIR, "DEG_external_raw.rds"))
gene_list_ext <- build_ranked_list(res_ext, "Gene", id_type = "SYMBOL")
gsea_ext      <- run_and_save(gene_list_ext, "External")

message("01_gsea.R complete.")
```

scripts/05_gsea/02_cross_cohort_concordance.R

Cross-cohort NES concordance analysis: Spearman correlation and identification of replicated Hallmark pathways.

```
# ============================================================================
# 02_cross_cohort_concordance.R — GSEA cross-cohort comparison
#
# Compares Hallmark NES scores between internal and external cohorts:
#   - Spearman correlation of NES values
#   - Identifies replicated pathways (same direction + significant in both)
#   - NES scatter plot
#
# Run AFTER 01_gsea.R
#
# Inputs:
#   GSEA_Hallmark_Internal.csv
#   GSEA_Hallmark_External.csv
#
# Outputs:
#   GSEA_CrossCohort.csv
#   GSEA_CrossCohort_NES_scatter.pdf
# ============================================================================

source("config/config.R")

library(dplyr)
library(ggplot2)
library(ggrepel)

int_file <- file.path(OUTPUT_DIR, "GSEA_Hallmark_Internal.csv")
ext_file <- file.path(OUTPUT_DIR, "GSEA_Hallmark_External.csv")

if (!file.exists(int_file) || !file.exists(ext_file)) {
  stop("Run 01_gsea.R first to generate both GSEA result files.")
}

# ── 1. Load and harmonise column names ──────────────────────────────────────
# Internal uses clusterProfiler column names (ID, p.adjust)
# External uses the same because run_and_save() uses the same GSEA() call
message("Loading GSEA results...")

gsea_int <- read.csv(int_file) %>%
  dplyr::select(pathway = ID, NES_int = NES, padj_int = p.adjust)

gsea_ext <- read.csv(ext_file) %>%
  dplyr::select(pathway = ID, NES_ext = NES, padj_ext = p.adjust)
```

```
# ── 2. Merge and annotate ──────────────────────────────────────────
cross <- inner_join(gsea_int, gsea_ext, by = "pathway") %>%
  mutate(
    same_direction = sign(NES_int) == sign(NES_ext),
    replicated     = same_direction & padj_int < 0.05 & padj_ext < 0.05,
    partial_rep    = same_direction & padj_ext < 0.05 & padj_int < 0.25,
    label          = gsub("HALLMARK_", "", pathway)   # clean name for plot
  ) %>%
  arrange(desc(replicated), padj_ext)

# ── 3. Print summary ──────────────────────────────────────────────
nes_cor <- cor.test(cross$NES_int, cross$NES_ext, method = "spearman")

message("Pathways compared: ", nrow(cross))
message("Same direction:    ", sum(cross$same_direction), "/", nrow(cross))
message("Both padj<0.05:    ", sum(cross$replicated))
message("NES Spearman rho = ", round(nes_cor$estimate, 3),
        "   p = ", signif(nes_cor$p.value, 3))

message("\nReplicated pathways:")
cross %>%
  filter(replicated) %>%
  dplyr::select(label, NES_int, padj_int, NES_ext, padj_ext) %>%
  print()

# ── 4. NES scatter plot ────────────────────────────────────────────
cor_label <- paste0("Spearman rho = ", round(nes_cor$estimate, 2),
                    "\np = ",           signif(nes_cor$p.value, 2))

p_scatter <- ggplot(cross, aes(x = NES_int, y = NES_ext)) +
  # reference line
  geom_abline(slope = 1, intercept = 0,
              linetype = "dashed", color = "grey60", linewidth = 0.5) +
  geom_hline(yintercept = 0, color = "grey80", linewidth = 0.3) +
  geom_vline(xintercept = 0, color = "grey80", linewidth = 0.3) +
  # all pathways
  geom_point(aes(color = replicated, size = replicated), alpha = 0.8) +
  # label replicated ones
  geom_text_repel(
    data    = filter(cross, replicated),
    aes(label = label),
    size    = 3, color = "black", max.overlaps = 20, box.padding = 0.4
  ) +
  scale_color_manual(values = c("FALSE" = "grey70", "TRUE" = COL_RTI)) +
  scale_size_manual(values  = c("FALSE" = 1.5,      "TRUE" = 3)) +
  annotate("text",
           x     = min(cross$NES_int, na.rm = TRUE) + 0.05,
           y     = max(cross$NES_ext, na.rm = TRUE),
           label = cor_label,
           hjust = 0, vjust = 1, size = 4) +
  theme_classic(base_size = 12) +
  labs(title    = "GSEA cross-cohort NES concordance",
       subtitle = "Hallmark gene sets | raw Wald stat ranking | highlighted: padj<0.05
in both",
       x        = "NES — Internal cohort",
       y        = "NES — External cohort") +
  theme(legend.position = "none",
        plot.title    = element_text(face = "bold"),
        plot.subtitle = element_text(color = "grey40"))

ggsave(file.path(OUTPUT_DIR, "GSEA_CrossCohort_NES_scatter.pdf"),
       p_scatter, width = 7, height = 7)

# ── 5. Save table ──────────────────────────────────────────────────
write.csv(cross %>% dplyr::select(-label),
          file.path(OUTPUT_DIR, "GSEA_CrossCohort.csv"),
          row.names = FALSE)

message("Saved: GSEA_CrossCohort.csv")
message("Saved: GSEA_CrossCohort_NES_scatter.pdf")
message("Done: 02_cross_cohort_concordance.R")
```

A.8 Downstream Analyses

01_estimate. R

ESTIMATE: tumour purity, stromal score, and immune score estimation for internal and external cohorts.

```
# ==============================================================================
# 01_estimate.R — ESTIMATE: stromal score, immune score, tumor purity
#
# Runs on:
#   A. Internal cohort (32 patients, log2 TPM+1, Ensembl IDs → Hugo via org.Hs.eg.db)
#   B. External cohort (55 QC patients, raw TPM → log2 TPM+1, Hugo symbols)
#
# Outputs:
#   ESTIMATE_Internal.csv / ESTIMATE_Internal_panels.pdf
#   ESTIMATE_External.csv / ESTIMATE_External_panels.pdf
#   ESTIMATE_Internal_wilcoxon.csv / ESTIMATE_External_wilcoxon.csv
# ==============================================================================

source("config/config.R")
source("config/utils.R")

library(estimate)
library(AnnotationDbi)
library(org.Hs.eg.db)
library(dplyr)
library(tidyr)
library(tibble)
library(readr)
library(ggplot2)
library(patchwork)

SCORE_COLS <- c("StromalScore", "ImmuneScore", "TumorPurity")

# ── write a GCT file manually (bypasses filterCommonGenes error) ──────────────
write_gct <- function(mat, gct_path) {
  f <- file(gct_path, "w")
  writeLines("#1.2", f)
  writeLines(paste(nrow(mat), ncol(mat), sep = "\t"), f)
  writeLines(paste(c("NAME", "Description", colnames(mat)), collapse = "\t"), f)
  out <- cbind(NAME = rownames(mat), Description = "na", as.data.frame(mat))
  write.table(out, file = f, sep = "\t", quote = FALSE,
              row.names = FALSE, col.names = FALSE)
  close(f)
}

# ── parse ESTIMATE GCT output → tidy dataframe ────────────────────────────────
parse_estimate_gct <- function(gct_path) {
  read_tsv(gct_path, skip = 2, show_col_types = FALSE) %>%
    dplyr::select(-Description) %>%
    pivot_longer(-NAME, names_to = "Sample_ID", values_to = "Value") %>%
    pivot_wider(names_from = NAME, values_from = Value) %>%
    mutate(
      TumorPurity = cos(0.6049872018 + 0.0001467884 * ESTIMATEScore),
      TumorPurity = pmin(pmax(TumorPurity, 0), 1)
    )
}

# ==============================================================================
# A. INTERNAL
# ==============================================================================
message("--- ESTIMATE: internal ---")

tpm_int  <- read.csv(INTERNAL_TPM_LOG_FILE, row.names = 1, check.names = FALSE)
meta_int <- read.csv(INTERNAL_METADATA_FILE, row.names = 1)

# Ensembl → Hugo via org.Hs.eg.db (offline, no internet needed)
message("  Translating Ensembl IDs to Hugo symbols...")
syms_int <- mapIds(org.Hs.eg.db,
                   keys      = rownames(tpm_int),
```

```
                  column    = "SYMBOL",
                  keytype   = "ENSEMBL",
                  multiVals = "first")

mat_int <- as.data.frame(as.matrix(tpm_int)) %>%
  mutate(SYM = syms_int) %>%
  filter(!is.na(SYM), SYM != "") %>%
  group_by(SYM) %>%
  summarise(across(everything(), max), .groups = "drop") %>%
  column_to_rownames("SYM") %>%
  as.matrix()

# Write GCT and run ESTIMATE
gct_in_int  <- file.path(BASE_DIR, "estimate_int_input.gct")
gct_out_int <- file.path(BASE_DIR, "estimate_int_scores.gct")
write_gct(mat_int, gct_in_int)
estimateScore(input.ds = gct_in_int, output.ds = gct_out_int, platform = "illumina")

# Parse results and attach group labels
shared_int <- intersect(colnames(mat_int), rownames(meta_int))
grp_int    <- setNames(
  ifelse(as.character(meta_int[shared_int, "Sporadic_vs_Radiotherapy_induced"]) ==
INTERNAL_RTI_LABEL,
         "RTI", "Non-RTI"),
  shared_int)

scores_int <- parse_estimate_gct(gct_out_int) %>%
  mutate(Group = factor(grp_int[Sample_ID], levels = c("Non-RTI", "RTI")))

write.csv(scores_int, file.path(OUTPUT_DIR, "ESTIMATE_Internal.csv"), row.names =
FALSE)

p_int <- wrap_plots(
  lapply(SCORE_COLS, \(col) boxplot_2grp(scores_int, col, "Internal (n=32)")),
  ncol = 3)
ggsave(file.path(OUTPUT_DIR, "ESTIMATE_Internal_panels.pdf"), p_int, width = 14, height
= 6)

write.csv(wilcox_tbl(scores_int, SCORE_COLS, "ESTIMATE", "Internal"),
          file.path(OUTPUT_DIR, "ESTIMATE_Internal_wilcoxon.csv"), row.names = FALSE)
message("  -> ESTIMATE_Internal_panels.pdf")

# ============================================================================
# B. EXTERNAL
# ============================================================================
message("--- ESTIMATE: external ---")

clin_ext <- read.delim(EXTERNAL_CLINICAL_55PT_FILE, stringsAsFactors = FALSE)
tpm_raw  <- read.delim(EXTERNAL_TPM_FILE, stringsAsFactors = FALSE)

# Subset to 55 QC patients, aggregate duplicate symbols, apply log2(TPM+1)
valid_cols <- intersect(colnames(tpm_raw), clin_ext$Sample_ID)

mat_ext <- tpm_raw %>%
  dplyr::select(Hugo_Symbol, all_of(valid_cols)) %>%
  group_by(Hugo_Symbol) %>%
  summarise(across(where(is.numeric), mean), .groups = "drop") %>%
  column_to_rownames("Hugo_Symbol") %>%
  as.matrix()

mat_ext <- log2(mat_ext + 1)

# Write GCT and run ESTIMATE on the combined 55-patient matrix
gct_in_ext  <- file.path(BASE_DIR, "estimate_ext_input.gct")
gct_out_ext <- file.path(BASE_DIR, "estimate_ext_scores.gct")
write_gct(mat_ext, gct_in_ext)
estimateScore(input.ds = gct_in_ext, output.ds = gct_out_ext, platform = "illumina")

# Parse results and attach group labels
grp_ext <- setNames(
  ifelse(clin_ext$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA[
```

```
          match(valid_cols, clin_ext$Sample_ID)] == EXTERNAL_RTI_LABEL,
        "RTI", "Non-RTI"),
  valid_cols)

scores_ext <- parse_estimate_gct(gct_out_ext) %>%
  mutate(Group = factor(grp_ext[Sample_ID], levels = c("Non-RTI", "RTI")))

write.csv(scores_ext, file.path(OUTPUT_DIR, "ESTIMATE_External.csv"), row.names =
FALSE)

p_ext <- wrap_plots(
  lapply(SCORE_COLS, \(col) boxplot_2grp(scores_ext, col, "External (n=55)")),
  ncol = 3)
ggsave(file.path(OUTPUT_DIR, "ESTIMATE_External_panels.pdf"), p_ext, width = 14, height
= 6)

write.csv(wilcox_tbl(scores_ext, SCORE_COLS, "ESTIMATE", "External"),
          file.path(OUTPUT_DIR, "ESTIMATE_External_wilcoxon.csv"), row.names = FALSE)
message("  -> ESTIMATE_External_panels.pdf")

message("01_estimate.R complete.")
```

02_mcp_counter.R

MCP-counter: immune and stromal cell type abundance estimation across ten cell populations.

```
# ================================================================================
# 02_mcp_counter.R — MCP-counter: immune and stromal cell type abundance
#
# Runs on:
#   A. Internal cohort (32 patients, log2 TPM+1, Ensembl IDs → Hugo via org.Hs.eg.db)
#   B. External cohort (55 QC patients, raw TPM → log2 TPM+1, Hugo symbols)
#
# Outputs:
#   MCPcounter_Internal.csv / MCPcounter_Internal_panels.pdf
#   MCPcounter_External.csv / MCPcounter_External_panels.pdf
#   MCPcounter_Internal_wilcoxon.csv / MCPcounter_External_wilcoxon.csv
# ================================================================================

source("config/config.R")
source("config/utils.R")

library(MCPcounter)
library(AnnotationDbi)
library(org.Hs.eg.db)
library(dplyr)
library(tibble)
library(ggplot2)
library(patchwork)

CELL_TYPES <- c(
  "T cells", "CD8 T cells", "Cytotoxic lymphocytes",
  "B lineage", "NK cells", "Monocytic lineage",
  "Myeloid dendritic cells", "Neutrophils",
  "Endothelial cells", "Fibroblasts"
)

# ================================================================================
# A. INTERNAL
# ================================================================================
message("--- MCP-counter: internal ---")

# INTERNAL_TPM_LOG_FILE is already log2(TPM+1) — convert Ensembl → Hugo
tpm_int <- read.csv(INTERNAL_TPM_LOG_FILE, row.names = 1, check.names = FALSE)
mat_int <- as.matrix(tpm_int)

syms_int <- mapIds(org.Hs.eg.db,
                   keys      = rownames(mat_int),
                   column    = "SYMBOL",
                   keytype   = "ENSEMBL",
```

```
                   multiVals = "first")

mat_int_sym <- as.data.frame(mat_int) %>%
  mutate(SYM = syms_int) %>%
  filter(!is.na(SYM), SYM != "") %>%
  group_by(SYM) %>%
  summarise(across(everything(), mean), .groups = "drop") %>%
  column_to_rownames("SYM") %>%
  as.matrix()

meta_int <- read.csv(INTERNAL_METADATA_FILE, row.names = 1)
shared_int <- intersect(colnames(mat_int_sym), rownames(meta_int))

grp_int <- setNames(
  ifelse(as.character(meta_int[shared_int, "Sporadic_vs_Radiotherapy_induced"]) ==
INTERNAL_RTI_LABEL,
         "RTI", "Non-RTI"),
  shared_int)

mcp_int <- as.data.frame(
  t(MCPcounter.estimate(mat_int_sym[, shared_int], featuresType = "HUGO_symbols"))) %>%
  rownames_to_column("Sample_ID") %>%
  mutate(Group = factor(grp_int[Sample_ID], levels = c("Non-RTI", "RTI")))

write.csv(mcp_int, file.path(OUTPUT_DIR, "MCPcounter_Internal.csv"), row.names = FALSE)

present_int <- intersect(CELL_TYPES, colnames(mcp_int))
p_int <- wrap_plots(
  lapply(present_int, \(ct) boxplot_cell(mcp_int, ct, "Internal (n=32)")),
  ncol = 5)
ggsave(file.path(OUTPUT_DIR, "MCPcounter_Internal_panels.pdf"), p_int, width = 20,
height = 10)

write.csv(wilcox_tbl(mcp_int, present_int, "MCPcounter", "Internal"),
          file.path(OUTPUT_DIR, "MCPcounter_Internal_wilcoxon.csv"), row.names = FALSE)
message("  -> MCPcounter_Internal_panels.pdf")

# ============================================================================
# B. EXTERNAL
# ============================================================================
message("--- MCP-counter: external ---")

clin_ext <- read.delim(EXTERNAL_CLINICAL_55PT_FILE, stringsAsFactors = FALSE)
tpm_raw  <- read.delim(EXTERNAL_TPM_FILE, stringsAsFactors = FALSE)

valid_cols <- intersect(colnames(tpm_raw), clin_ext$Sample_ID)

mat_ext <- tpm_raw %>%
  dplyr::select(Hugo_Symbol, all_of(valid_cols)) %>%
  group_by(Hugo_Symbol) %>%
  summarise(across(where(is.numeric), mean), .groups = "drop") %>%
  column_to_rownames("Hugo_Symbol") %>%
  as.matrix()

mat_ext <- log2(mat_ext + 1)

grp_ext <- setNames(
  ifelse(clin_ext$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA[
           match(valid_cols, clin_ext$Sample_ID)] == EXTERNAL_RTI_LABEL,
         "RTI", "Non-RTI"),
  valid_cols)

mcp_ext <- as.data.frame(
  t(MCPcounter.estimate(mat_ext, featuresType = "HUGO_symbols"))) %>%
  rownames_to_column("Sample_ID") %>%
  mutate(Group = factor(grp_ext[Sample_ID], levels = c("Non-RTI", "RTI")))

write.csv(mcp_ext, file.path(OUTPUT_DIR, "MCPcounter_External.csv"), row.names = FALSE)

present_ext <- intersect(CELL_TYPES, colnames(mcp_ext))
p_ext <- wrap_plots(
```

```
  lapply(present_ext, \(ct) boxplot_cell(mcp_ext, ct, "External (n=55)")),
  ncol = 5)
ggsave(file.path(OUTPUT_DIR, "MCPcounter_External_panels.pdf"), p_ext, width = 20,
height = 10)

write.csv(wilcox_tbl(mcp_ext, present_ext, "MCPcounter", "External"),
          file.path(OUTPUT_DIR, "MCPcounter_External_wilcoxon.csv"), row.names = FALSE)
message("  -> MCPcounter_External_panels.pdf")

message("02_mcp_counter.R complete.")
```

03_tide.R

TIDE: tumour immune dysfunction and exclusion analysis, including CAF score and ICI response prediction.

```
# ==============================================================================
# 03_tide.R — TIDE: immune evasion and ICI response prediction
#
# TIDE requires the downloaded tidepy package (the online TIDE portal is
# unavailable). This script runs in three steps:
#
#   STEP 1 (R)      — prepare mean-centred log2(TPM+1) input files
#   STEP 2 (Python) — run tidepy locally (commands printed by this script)
#   STEP 3 (R)      — load TIDE output, make boxplots + ICI responder bar chart
#
# Re-run this script after completing Step 2 — it will detect the result
# files and proceed directly to Step 3.
#
# Runs on:
#   A. Internal cohort (32 patients, Ensembl IDs → Hugo via org.Hs.eg.db)
#   B. External cohort (55 QC patients, Hugo symbols)
#
# Outputs:
#   TIDE_Internal.csv / TIDE_Internal_panels.pdf / TIDE_Internal_ICI_bar.pdf
#   TIDE_External.csv / TIDE_External_panels.pdf / TIDE_External_ICI_bar.pdf
#   TIDE_Internal_wilcoxon.csv / TIDE_External_wilcoxon.csv
# ==============================================================================

source("config/config.R")
source("config/utils.R")

library(AnnotationDbi)
library(org.Hs.eg.db)
library(dplyr)
library(tibble)
library(readr)
library(ggplot2)
library(patchwork)

TIDE_SCORE_COLS <- c("TIDE", "Dysfunction", "Exclusion", "CAF")

# ── mean-centre a log2 matrix (required by TIDE) ──────────────────────────────
centre_for_tide <- function(mat_log) {
  mat_log - rowMeans(mat_log)
}

# ==============================================================================
# STEP 1 — PREPARE INPUT FILES
# ==============================================================================
message("--- TIDE Step 1: preparing input files ---")

# ── Internal ──────────────────────────────────────────────────────────────────
tpm_int <- read.csv(INTERNAL_TPM_LOG_FILE, row.names = 1, check.names = FALSE)
mat_int <- as.matrix(tpm_int)

syms_int <- mapIds(org.Hs.eg.db,
                   keys      = rownames(mat_int),
                   column    = "SYMBOL",
                   keytype   = "ENSEMBL",
                   multiVals = "first")
```

```
mat_int_sym <- as.data.frame(mat_int) %>%
  mutate(SYM = syms_int) %>%
  filter(!is.na(SYM), SYM != "") %>%
  group_by(SYM) %>%
  summarise(across(everything(), mean), .groups = "drop") %>%
  column_to_rownames("SYM") %>%
  as.matrix()

tide_in_int  <- as.data.frame(centre_for_tide(mat_int_sym)) %>%
  rownames_to_column("Gene")
tide_input_int <- file.path(BASE_DIR, "TIDE_Input_Internal.txt")
write_tsv(tide_in_int, tide_input_int)
message("  Saved: TIDE_Input_Internal.txt")

# ── External ──────────────────────────────────────────────────────────
clin_ext <- read.delim(EXTERNAL_CLINICAL_55PT_FILE, stringsAsFactors = FALSE)
tpm_raw  <- read.delim(EXTERNAL_TPM_FILE, stringsAsFactors = FALSE)

valid_cols <- intersect(colnames(tpm_raw), clin_ext$Sample_ID)

mat_ext <- tpm_raw %>%
  dplyr::select(Hugo_Symbol, all_of(valid_cols)) %>%
  group_by(Hugo_Symbol) %>%
  summarise(across(where(is.numeric), mean), .groups = "drop") %>%
  column_to_rownames("Hugo_Symbol") %>%
  as.matrix()

mat_ext <- log2(mat_ext + 1)

tide_in_ext <- as.data.frame(centre_for_tide(mat_ext)) %>%
  rownames_to_column("Gene")
tide_input_ext <- file.path(BASE_DIR, "TIDE_Input_External.txt")
write_tsv(tide_in_ext, tide_input_ext)
message("  Saved: TIDE_Input_External.txt")

# ── Python commands to run locally ────────────────────────────────────
#
# The online TIDE portal is unavailable. Run tidepy locally from the terminal.
# Make sure tidepy is installed: pip install tidepy
#
tide_result_int <- file.path(BASE_DIR, "TIDE_Results_Internal.csv")
tide_result_ext <- file.path(BASE_DIR, "TIDE_Results_External.csv")

message("")
message("  *** STEP 2: run these two Python commands in your terminal ***")
message("")
message("  python3 -c \"")
message("  from tidepy.pred import TIDE; import pandas as pd")
message("  df = pd.read_csv('", tide_input_int, "', sep='\\t', index_col=0)")
message("  TIDE(df, cancer='Other', pretreat=False).to_csv('", tide_result_int, "')\"")
message("")
message("  python3 -c \"")
message("  from tidepy.pred import TIDE; import pandas as pd")
message("  df = pd.read_csv('", tide_input_ext, "', sep='\\t', index_col=0)")
message("  TIDE(df, cancer='Other', pretreat=False).to_csv('", tide_result_ext, "')\"")
message("")
message("  Then re-run this script — Step 3 will run automatically.")

# ═══════════════════════════════════════════════════════════════════════
# STEP 3 — LOAD RESULTS AND PLOT (runs only when Python output exists)
# ═══════════════════════════════════════════════════════════════════════

# ── Internal ──────────────────────────────────────────────────────────
if (!file.exists(tide_result_int)) {
  message("  Internal: TIDE_Results_Internal.csv not found — run Python step first.")
} else {
  message("--- TIDE Step 3: plotting internal ---")

  meta_int <- read.csv(INTERNAL_METADATA_FILE, row.names = 1)
  shared_int <- intersect(colnames(mat_int_sym), rownames(meta_int))
```

```
  grp_int <- setNames(
    ifelse(as.character(meta_int[shared_int, "Sporadic_vs_Radiotherapy_induced"]) ==
INTERNAL_RTI_LABEL,
           "RTI", "Non-RTI"),
    shared_int)

  tide_int <- read_csv(tide_result_int, show_col_types = FALSE) %>%
    rename(Sample_ID = 1) %>%
    mutate(Group = factor(grp_int[Sample_ID], levels = c("Non-RTI", "RTI"))) %>%
    filter(!is.na(Group))

  write.csv(tide_int, file.path(OUTPUT_DIR, "TIDE_Internal.csv"), row.names = FALSE)

  present_int <- intersect(TIDE_SCORE_COLS, colnames(tide_int))
  p_int <- wrap_plots(
    lapply(present_int, \(col) boxplot_2grp(tide_int, col, "Internal (n=32)")),
    ncol = 4)
  ggsave(file.path(OUTPUT_DIR, "TIDE_Internal_panels.pdf"), p_int, width = 16, height =
6)
  message("  -> TIDE_Internal_panels.pdf")

  if ("Responder" %in% colnames(tide_int)) {
    tide_int$Responder <- as.logical(tide_int$Responder)
    p_ici_int <- ici_bar(tide_int, "Internal (n=32)")
    ggsave(file.path(OUTPUT_DIR, "TIDE_Internal_ICI_bar.pdf"), p_ici_int, width = 6,
height = 7)
    message("  -> TIDE_Internal_ICI_bar.pdf")
  }

  write.csv(wilcox_tbl_tide(tide_int, "Internal"),
            file.path(OUTPUT_DIR, "TIDE_Internal_wilcoxon.csv"), row.names = FALSE)
}

# ── External ──────────────────────────────────────────────────────────
if (!file.exists(tide_result_ext)) {
  message("  External: TIDE_Results_External.csv not found — run Python step first.")
} else {
  message("--- TIDE Step 3: plotting external ---")

  grp_ext <- setNames(
    ifelse(clin_ext$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA[
             match(valid_cols, clin_ext$Sample_ID)] == EXTERNAL_RTI_LABEL,
           "RTI", "Non-RTI"),
    valid_cols)

  tide_ext <- read_csv(tide_result_ext, show_col_types = FALSE) %>%
    rename(Sample_ID = 1) %>%
    mutate(Group = factor(grp_ext[Sample_ID], levels = c("Non-RTI", "RTI"))) %>%
    filter(!is.na(Group))

  write.csv(tide_ext, file.path(OUTPUT_DIR, "TIDE_External.csv"), row.names = FALSE)

  present_ext <- intersect(TIDE_SCORE_COLS, colnames(tide_ext))
  p_ext <- wrap_plots(
    lapply(present_ext, \(col) boxplot_2grp(tide_ext, col, "External (n=55)")),
    ncol = 4)
  ggsave(file.path(OUTPUT_DIR, "TIDE_External_panels.pdf"), p_ext, width = 16, height =
6)
  message("  -> TIDE_External_panels.pdf")

  if ("Responder" %in% colnames(tide_ext)) {
    tide_ext$Responder <- as.logical(tide_ext$Responder)
    p_ici_ext <- ici_bar(tide_ext, "External (n=55)")
    ggsave(file.path(OUTPUT_DIR, "TIDE_External_ICI_bar.pdf"), p_ici_ext, width = 6,
height = 7)
    message("  -> TIDE_External_ICI_bar.pdf")
  }

  write.csv(wilcox_tbl_tide(tide_ext, "External"),
            file.path(OUTPUT_DIR, "TIDE_External_wilcoxon.csv"), row.names = FALSE)
}
```

```
message("03_tide.R complete.")
```

04_cibersort. R

CIBERSORT: immune cell fraction estimation using the LM22 reference matrix (22 cell types).

```
# ==============================================================================
# 04_cibersort.R — CIBERSORT: immune cell fractions (LM22, 22 cell types)
#
# Requires CIBERSORT.R and LM22.txt in DATA_DIR.
# Request from: https://cibersortx.stanford.edu/
#
# Runs on:
#   A. Internal cohort (32 patients, log2 TPM+1 → linear, Ensembl → Hugo)
#   B. External cohort (55 QC patients, raw TPM → linear)
#
# Outputs:
#   CIBERSORT_Internal.csv / CIBERSORT_Internal_panels.pdf
#   CIBERSORT_External.csv / CIBERSORT_External_panels.pdf
#   CIBERSORT_Internal_wilcoxon.csv / CIBERSORT_External_wilcoxon.csv
# ==============================================================================

source("config/config.R")
source("config/utils.R")

library(AnnotationDbi)
library(org.Hs.eg.db)
library(dplyr)
library(tibble)
library(ggplot2)
library(patchwork)

# ── check required CIBERSORT files ────────────────────────────────────────────
if (!file.exists(CIBERSORT_R_FILE) || !file.exists(LM22_FILE)) {
  stop("CIBERSORT.R or LM22.txt not found in DATA_DIR.\n",
       "Request from: https://cibersortx.stanford.edu/\n",
       "Expected:\n  ", CIBERSORT_R_FILE, "\n  ", LM22_FILE)
}

# Disable parallel processing — prevents crashes on HPC environments
mclapply <- function(..., mc.cores) { lapply(...) }
source(CIBERSORT_R_FILE)

# Read LM22 and rewrite with clean line endings (required by CIBERSORT)
lm22_data  <- read.table(LM22_FILE, header = TRUE, row.names = 1,
                         sep = "\t", check.names = FALSE)
lm22_fixed <- file.path(BASE_DIR, "LM22_Fixed.txt")
write.table(lm22_data, lm22_fixed, sep = "\t", quote = FALSE, col.names = NA)

CELL_TYPES_LM22 <- c(
  "B cells naive", "B cells memory", "Plasma cells",
  "T cells CD8", "T cells CD4 naive",
  "T cells CD4 memory resting", "T cells CD4 memory activated",
  "T cells follicular helper", "T cells regulatory (Tregs)",
  "T cells gamma delta", "NK cells resting", "NK cells activated",
  "Monocytes", "Macrophages M0", "Macrophages M1", "Macrophages M2",
  "Dendritic cells resting", "Dendritic cells activated",
  "Mast cells resting", "Mast cells activated",
  "Eosinophils", "Neutrophils"
)

# ==============================================================================
# A. INTERNAL
# ==============================================================================
message("--- CIBERSORT: internal ---")

tpm_int <- read.csv(INTERNAL_TPM_LOG_FILE, row.names = 1, check.names = FALSE)
mat_int <- as.matrix(tpm_int)
```

```
syms_int <- mapIds(org.Hs.eg.db,
                   keys      = rownames(mat_int),
                   column    = "SYMBOL",
                   keytype   = "ENSEMBL",
                   multiVals = "first")

mat_int_sym <- as.data.frame(mat_int) %>%
  mutate(SYM = syms_int) %>%
  filter(!is.na(SYM), SYM != "") %>%
  group_by(SYM) %>%
  summarise(across(everything(), mean), .groups = "drop") %>%
  column_to_rownames("SYM") %>%
  as.matrix()

# CIBERSORT requires linear TPM — un-log the log2(TPM+1) matrix
mat_int_lin <- 2^mat_int_sym - 1
mat_int_lin[!is.finite(mat_int_lin)] <- 0

common_int  <- intersect(rownames(mat_int_lin), rownames(lm22_data))
mat_int_lin <- mat_int_lin[common_int, ]
mat_int_lin <- mat_int_lin[, colSums(mat_int_lin) > 0, drop = FALSE]

cib_in_int <- file.path(BASE_DIR, "CIBERSORT_Input_Internal.txt")
write.table(mat_int_lin, cib_in_int, sep = "\t", quote = FALSE, col.names = NA)

message("  Running CIBERSORT internal (perm=100)...")
res_int <- CIBERSORT(lm22_fixed, cib_in_int, perm = 100, QN = FALSE)

meta_int <- read.csv(INTERNAL_METADATA_FILE, row.names = 1)
shared_int <- intersect(rownames(res_int), rownames(meta_int))
grp_int <- setNames(
  ifelse(as.character(meta_int[shared_int, "Sporadic_vs_Radiotherapy_induced"]) ==
INTERNAL_RTI_LABEL,
         "RTI", "Non-RTI"),
  shared_int)

cib_int <- as.data.frame(res_int) %>%
  rownames_to_column("Sample_ID") %>%
  mutate(Group = factor(grp_int[Sample_ID], levels = c("Non-RTI", "RTI")))

write.csv(cib_int, file.path(OUTPUT_DIR, "CIBERSORT_Internal.csv"), row.names = FALSE)

present_int <- intersect(CELL_TYPES_LM22, colnames(cib_int))
p_int <- wrap_plots(
  lapply(present_int, \(ct) boxplot_cell(cib_int, ct, "Internal (n=32)")),
  ncol = 6)
ggsave(file.path(OUTPUT_DIR, "CIBERSORT_Internal_panels.pdf"),
       p_int, width = 22, height = 16)

write.csv(wilcox_tbl(cib_int, present_int, "CIBERSORT", "Internal"),
          file.path(OUTPUT_DIR, "CIBERSORT_Internal_wilcoxon.csv"), row.names = FALSE)
message("  -> CIBERSORT_Internal_panels.pdf")

# ============================================================================
# B. EXTERNAL
# ============================================================================
message("--- CIBERSORT: external ---")

clin_ext <- read.delim(EXTERNAL_CLINICAL_55PT_FILE, stringsAsFactors = FALSE)
tpm_raw  <- read.delim(EXTERNAL_TPM_FILE, stringsAsFactors = FALSE)

valid_cols <- intersect(colnames(tpm_raw), clin_ext$Sample_ID)

mat_ext <- tpm_raw %>%
  dplyr::select(Hugo_Symbol, all_of(valid_cols)) %>%
  group_by(Hugo_Symbol) %>%
  summarise(across(where(is.numeric), mean), .groups = "drop") %>%
  column_to_rownames("Hugo_Symbol") %>%
  as.matrix()

# External TPM file is raw linear — use directly without un-logging
```

```
common_ext <- intersect(rownames(mat_ext), rownames(lm22_data))
mat_ext    <- mat_ext[common_ext, ]
mat_ext    <- mat_ext[, colSums(mat_ext) > 0, drop = FALSE]

cib_in_ext <- file.path(BASE_DIR, "CIBERSORT_Input_External.txt")
write.table(mat_ext, cib_in_ext, sep = "\t", quote = FALSE, col.names = NA)

message("  Running CIBERSORT external (perm=100)...")
res_ext <- CIBERSORT(lm22_fixed, cib_in_ext, perm = 100, QN = FALSE)

grp_ext <- setNames(
  ifelse(clin_ext$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA[
           match(valid_cols, clin_ext$Sample_ID)] == EXTERNAL_RTI_LABEL,
         "RTI", "Non-RTI"),
  valid_cols)

cib_ext <- as.data.frame(res_ext) %>%
  rownames_to_column("Sample_ID") %>%
  mutate(Group = factor(grp_ext[Sample_ID], levels = c("Non-RTI", "RTI")))

write.csv(cib_ext, file.path(OUTPUT_DIR, "CIBERSORT_External.csv"), row.names = FALSE)

present_ext <- intersect(CELL_TYPES_LM22, colnames(cib_ext))
p_ext <- wrap_plots(
  lapply(present_ext, \(ct) boxplot_cell(cib_ext, ct, "External (n=55)")),
  ncol = 6)
ggsave(file.path(OUTPUT_DIR, "CIBERSORT_External_panels.pdf"),
       p_ext, width = 22, height = 16)

write.csv(wilcox_tbl(cib_ext, present_ext, "CIBERSORT", "External"),
          file.path(OUTPUT_DIR, "CIBERSORT_External_wilcoxon.csv"), row.names = FALSE)
message("  -> CIBERSORT_External_panels.pdf")

message("04_cibersort.R complete.")
```

05_apm_singscore. R

singscore: single-sample gene set scoring for antigen presentation, MHC class I/II, T cell-inflamed GEP, immune checkpoint, and TLS signatures.

```
# ============================================================================
# 05_apm_singscore.R — APM singscore: antigen presentation machinery signatures
#
# singscore (Foroutan et al. 2018) is a rank-based single-sample gene set
# scoring method. It ranks each gene within each sample independently and
# computes an enrichment score for a given gene set without requiring a
# reference dataset. It is used here to quantify activity of antigen
# presentation and immune-related gene programmes in each patient.
#
# Signatures scored:
#   - Antigen_Processing  (TAP1, TAP2, TAPBP, PSMB8/9, CALR, CANX, B2M, HLA-A/B/C)
#   - MHC_Class_I         (HLA-A/B/C, B2M, TAP1/2)
#   - MHC_Class_II        (HLA-DR/DQ/DP, CD74)
#   - T_cell_inflamed     (CD8A/B, GZMB, PRF1, IFNG, CXCLs, IDO1, checkpoints)
#   - Immune_Checkpoint   (CD274, PDCD1, CTLA4, LAG3, HAVCR2, TIGIT, PDCD1LG2)
#   - TLS_Signature       (CCL19/21, CXCL13, SELL, CCR7, LAMP3, BCL6, CXCR5, CD79B,
MS4A1)
#
# Runs on:
#   A. Internal cohort (32 patients, log2 TPM+1, Ensembl IDs → Hugo via org.Hs.eg.db)
#   B. External cohort (55 QC patients, raw TPM → log2 TPM+1, Hugo symbols)
#
# Outputs:
#   APM_Internal.csv / APM_Internal_panels.pdf / APM_Internal_wilcoxon.csv
#   APM_External.csv / APM_External_panels.pdf / APM_External_wilcoxon.csv
# ============================================================================

source("config/config.R")
source("config/utils.R")

library(singscore)
```

```
library(GSEABase)
library(AnnotationDbi)
library(org.Hs.eg.db)
library(dplyr)
library(tibble)
library(ggplot2)
library(patchwork)

# ── APM gene sets ──────────────────────────────────────────────────────
APM_SETS <- list(
  Antigen_Processing = GeneSet(
    c("TAP1","TAP2","TAPBP","PSMB8","PSMB9","CALR","CANX","B2M",
      "HLA-A","HLA-B","HLA-C"),
    setName = "Antigen_Processing"),
  MHC_Class_I = GeneSet(
    c("HLA-A","HLA-B","HLA-C","B2M","TAP1","TAP2"),
    setName = "MHC_Class_I"),
  MHC_Class_II = GeneSet(
    c("HLA-DRA","HLA-DRB1","HLA-DQA1","HLA-DQB1",
      "HLA-DPA1","HLA-DPB1","CD74"),
    setName = "MHC_Class_II"),
  T_cell_inflamed = GeneSet(
    c("CD8A","CD8B","GZMB","PRF1","IFNG","CXCL9","CXCL10",
      "CXCL11","IDO1","TIGIT","CD274","PDCD1LG2"),
    setName = "T_cell_inflamed"),
  Immune_Checkpoint = GeneSet(
    c("CD274","PDCD1","CTLA4","LAG3","HAVCR2","TIGIT","PDCD1LG2"),
    setName = "Immune_Checkpoint"),
  TLS_Signature = GeneSet(
    c("CCL19","CCL21","CXCL13","SELL","CCR7","LAMP3",
      "BCL6","CXCR5","CD79B","MS4A1"),
    setName = "TLS_Signature")
)

SIG_COLS <- names(APM_SETS)

# ── run singscore on a Hugo-symbol matrix ──────────────────────────────
run_apm <- function(mat, grp, cohort_label) {
  ranked <- rankGenes(mat)
  scores <- data.frame(Sample_ID = colnames(ranked))
  for (nm in SIG_COLS) {
    scores[[nm]] <- simpleScore(ranked, upSet = APM_SETS[[nm]])$TotalScore
  }
  scores %>%
    mutate(Group = factor(grp[Sample_ID], levels = c("Non-RTI", "RTI")))
}

# ═══════════════════════════════════════════════════════════════════════
# A. INTERNAL
# ═══════════════════════════════════════════════════════════════════════
message("--- APM singscore: internal ---")

tpm_int  <- read.csv(INTERNAL_TPM_LOG_FILE, row.names = 1, check.names = FALSE)
meta_int <- read.csv(INTERNAL_METADATA_FILE, row.names = 1)
mat_int  <- as.matrix(tpm_int)

# Ensembl → Hugo via org.Hs.eg.db (offline)
syms_int <- mapIds(org.Hs.eg.db,
                   keys      = rownames(mat_int),
                   column    = "SYMBOL",
                   keytype   = "ENSEMBL",
                   multiVals = "first")

mat_int_sym <- as.data.frame(mat_int) %>%
  mutate(SYM = syms_int) %>%
  filter(!is.na(SYM), SYM != "") %>%
  group_by(SYM) %>%
  summarise(across(everything(), mean), .groups = "drop") %>%
  column_to_rownames("SYM") %>%
  as.matrix()
```

```
shared_int <- intersect(colnames(mat_int_sym), rownames(meta_int))
grp_int <- setNames(
  ifelse(as.character(meta_int[shared_int, "Sporadic_vs_Radiotherapy_induced"]) ==
INTERNAL_RTI_LABEL,
         "RTI", "Non-RTI"),
  shared_int)

apm_int <- run_apm(mat_int_sym[, shared_int], grp_int, "Internal")

write.csv(apm_int, file.path(OUTPUT_DIR, "APM_Internal.csv"), row.names = FALSE)

p_int <- wrap_plots(
  lapply(SIG_COLS, \(nm) boxplot_2grp(apm_int, nm, "Internal (n=32)")),
  ncol = 3)
ggsave(file.path(OUTPUT_DIR, "APM_Internal_panels.pdf"), p_int, width = 14, height =
10)

write.csv(wilcox_tbl(apm_int, SIG_COLS, "singscore", "Internal"),
          file.path(OUTPUT_DIR, "APM_Internal_wilcoxon.csv"), row.names = FALSE)
message("  -> APM_Internal_panels.pdf")

# ============================================================================
# B. EXTERNAL
# ============================================================================
message("--- APM singscore: external ---")

clin_ext <- read.delim(EXTERNAL_CLINICAL_55PT_FILE, stringsAsFactors = FALSE)
tpm_raw  <- read.delim(EXTERNAL_TPM_FILE, stringsAsFactors = FALSE)

valid_cols <- intersect(colnames(tpm_raw), clin_ext$Sample_ID)

mat_ext <- tpm_raw %>%
  dplyr::select(Hugo_Symbol, all_of(valid_cols)) %>%
  group_by(Hugo_Symbol) %>%
  summarise(across(where(is.numeric), mean), .groups = "drop") %>%
  column_to_rownames("Hugo_Symbol") %>%
  as.matrix()

mat_ext <- log2(mat_ext + 1)

grp_ext <- setNames(
  ifelse(clin_ext$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA[
           match(valid_cols, clin_ext$Sample_ID)] == EXTERNAL_RTI_LABEL,
         "RTI", "Non-RTI"),
  valid_cols)

apm_ext <- run_apm(mat_ext, grp_ext, "External")

write.csv(apm_ext, file.path(OUTPUT_DIR, "APM_External.csv"), row.names = FALSE)

p_ext <- wrap_plots(
  lapply(SIG_COLS, \(nm) boxplot_2grp(apm_ext, nm, "External (n=55)")),
  ncol = 3)
ggsave(file.path(OUTPUT_DIR, "APM_External_panels.pdf"), p_ext, width = 14, height =
10)

write.csv(wilcox_tbl(apm_ext, SIG_COLS, "singscore", "External"),
          file.path(OUTPUT_DIR, "APM_External_wilcoxon.csv"), row.names = FALSE)
message("  -> APM_External_panels.pdf")

message("05_apm_singscore.R complete.")
```

06_msi. R

MSIsensor-RNA: microsatellite instability prediction from bulk RNA-seq expression data.

```
# ============================================================================
# 06_msi.R — MSIsensor-RNA: microsatellite instability prediction from RNA-seq
#
# MSIsensor-RNA predicts microsatellite instability (MSI) status from bulk
# RNA-seq data using a machine learning model trained on TCGA data. MSI occurs
# when the DNA mismatch repair (MMR) system is defective, causing accumulation
# of mutations at short repetitive DNA sequences (microsatellites). Tumours
```

```
# with high MSI (MSI-H) are associated with higher tumour mutational burden
# and better response to immune checkpoint inhibitors. MSIsensor-RNA provides
# a continuous probability score (0-1) of MSI-H status per sample, allowing
# comparison between groups without requiring DNA sequencing.
#
# Tool: MSIsensor-RNA (Sun et al. 2021) — https://github.com/xjtu-omics/msisensor-rna
# Model: TCGA.MSIPopular.model.pkl (must be in BASE_DIR/msisensor-rna/model/)
#
# This script runs in three steps:
#   STEP 1 (R)      — prepare normalised input files (log2 + z-score)
#   STEP 2 (Python) — run MSIsensor-RNA locally (commands printed by this script)
#   STEP 3 (R)      — load results, make boxplots + Wilcoxon table
#
# Re-run this script after completing Step 2 — it will detect the result
# files and proceed directly to Step 3.
#
# Runs on:
#   A. Internal cohort (32 patients, log2 TPM+1, Ensembl IDs → Hugo via org.Hs.eg.db)
#   B. External cohort (55 QC patients, raw TPM → log2 TPM+1, Hugo symbols)
#
# Outputs:
#   MSI_Internal.csv / MSI_Internal_plot.pdf / MSI_Internal_wilcoxon.csv
#   MSI_External.csv / MSI_External_plot.pdf / MSI_External_wilcoxon.csv
# ==============================================================================

source("config/config.R")
source("config/utils.R")

library(AnnotationDbi)
library(org.Hs.eg.db)
library(dplyr)
library(tibble)
library(readr)
library(ggplot2)

# ── genes required by the MSIsensor-RNA model ─────────────────────────────────
# These genes must exist in the input matrix. If absent from the expression
# data they are added as zeros so the model does not crash.
MSI_REQUIRED_GENES <- c(
  "TMEM176A","DUSP4","CA4","DMD","FAM3B","ABLIM2","LSAMP","SPIN3","HPDL",
  "SEMA5A","UNC5CL","GPR153","SPINK1","RNLS","CH25H","PPARGC1A","BMP7",
  "SLC39A5","CFTR","PLA2G4A","SLC9A3","NEBL","CHL1","TNFRSF18","ABCC6",
  "HUNK","TMC7","FXYD1","ABAT","TCF21","UGT2A3","GNLY","ZFP3","SEMA6D",
  "COL9A3","RAB26","NAALADL1","GZMH","SULT1B1","AOAH","ULBP2","ANKRD29",
  "DEPDC7","HSPA4L","EPM2AIP1","QPCT","SLC22A3","SPAG16","GNG4","MATN3",
  "ZNF300","CHN2","QPRT","C3orf14","NKD1","WNT4","TNNC2","CARD11","PIPOX",
  "IL34","SERP2","KLHL29","SLC2A12","MLH1","ST6GALNAC2","PARVG","HMGCS2",
  "PAPLN","CEBPA","SLC19A3","EDNRB","BCL11A","JSRP1","SOSTDC1","CAMK1D",
  "MLF1","ZSWIM3","F10","SYTL5","SCARA5","CYP2B6","CTTNBP2","LPL","KCNH2",
  "ZSCAN12","FBXO16","ABCB1","C7orf31","HAAO","WNT5B","GSPT2","CEL","AGT",
  "TNFSF9","DPEP1","ARMCX4","TDGF1","TTC30A","SLCO3A1","MUC20","MAPK12",
  "ERP27","MSX1","NFASC","DTNA","PID1","APOLD1","CAB39L","CA8"
)

# ── prepare and normalise matrix for MSIsensor-RNA ────────────────────────────
# MSIsensor-RNA expects: samples as rows, genes as columns, log2 + z-score
prepare_msi_input <- function(mat_log, tag) {
  # z-score per gene across samples
  mat_z <- t(scale(t(mat_log)))
  mat_z[!is.finite(mat_z)] <- 0

  # transpose: samples as rows
  df_t <- as.data.frame(t(mat_z))

  # add any missing model genes as zeros
  missing <- setdiff(MSI_REQUIRED_GENES, colnames(df_t))
  if (length(missing) > 0) {
    df_t[missing] <- 0.0
    message("  Added ", length(missing), " missing model genes as zeros")
  }
```

```
  df_t <- rownames_to_column(df_t, "SampleID")
  out_path <- file.path(BASE_DIR, paste0("MSI_Input_", tag, ".csv"))
  write_csv(df_t, out_path)
  message("  Saved: MSI_Input_", tag, ".csv")
  out_path
}

# ============================================================================
# STEP 1 — PREPARE INPUT FILES
# ============================================================================
message("--- MSI Step 1: preparing input files ---")

# ── Internal ──────────────────────────────────────────────────────────────
tpm_int  <- read.csv(INTERNAL_TPM_LOG_FILE, row.names = 1, check.names = FALSE)
meta_int <- read.csv(INTERNAL_METADATA_FILE, row.names = 1)
mat_int  <- as.matrix(tpm_int)

syms_int <- mapIds(org.Hs.eg.db,
                   keys      = rownames(mat_int),
                   column    = "SYMBOL",
                   keytype   = "ENSEMBL",
                   multiVals = "first")

mat_int_sym <- as.data.frame(mat_int) %>%
  mutate(SYM = syms_int) %>%
  filter(!is.na(SYM), SYM != "") %>%
  group_by(SYM) %>%
  summarise(across(everything(), mean), .groups = "drop") %>%
  column_to_rownames("SYM") %>%
  as.matrix()

msi_input_int <- prepare_msi_input(mat_int_sym, "Internal")

# ── External ──────────────────────────────────────────────────────────────
clin_ext <- read.delim(EXTERNAL_CLINICAL_55PT_FILE, stringsAsFactors = FALSE)
tpm_raw  <- read.delim(EXTERNAL_TPM_FILE, stringsAsFactors = FALSE)

valid_cols <- intersect(colnames(tpm_raw), clin_ext$Sample_ID)

mat_ext <- tpm_raw %>%
  dplyr::select(Hugo_Symbol, all_of(valid_cols)) %>%
  group_by(Hugo_Symbol) %>%
  summarise(across(where(is.numeric), mean), .groups = "drop") %>%
  column_to_rownames("Hugo_Symbol") %>%
  as.matrix()

mat_ext <- log2(mat_ext + 1)

msi_input_ext <- prepare_msi_input(mat_ext, "External")

# ── Python commands to run locally ────────────────────────────────────────
msi_result_int <- file.path(BASE_DIR, "MSI_Results_Internal.txt")
msi_result_ext <- file.path(BASE_DIR, "MSI_Results_External.txt")

message("")
message("  *** STEP 2: run these two commands in your terminal ***")
message("")
message("  echo 'Yes' | msisensor-rna detection \\")
message("    -m ", MSI_MODEL_PATH, " \\")
message("    -i ", msi_input_int,  " \\")
message("    -o ", msi_result_int)
message("")
message("  echo 'Yes' | msisensor-rna detection \\")
message("    -m ", MSI_MODEL_PATH, " \\")
message("    -i ", msi_input_ext,  " \\")
message("    -o ", msi_result_ext)
message("")
message("  Then re-run this script — Step 3 will run automatically.")

# ============================================================================
# STEP 3 — LOAD RESULTS AND PLOT (runs only when output exists)
```

```
# ==============================================================================

# ── Internal ──────────────────────────────────────────────────────────────
if (!file.exists(msi_result_int)) {
  message("  Internal: MSI_Results_Internal.txt not found — run Step 2 first.")
} else {
  message("--- MSI Step 3: plotting internal ---")

  shared_int <- intersect(colnames(mat_int_sym), rownames(meta_int))
  grp_int <- setNames(
    ifelse(as.character(meta_int[shared_int, "Sporadic_vs_Radiotherapy_induced"]) ==
INTERNAL_RTI_LABEL,
           "RTI", "Non-RTI"),
    shared_int)

  msi_int <- read_tsv(msi_result_int, show_col_types = FALSE) %>%
    rename(Sample_ID = 1, MSI_Score = `probability_of_MSI-H`) %>%
    mutate(Group = factor(grp_int[Sample_ID], levels = c("Non-RTI", "RTI"))) %>%
    filter(!is.na(Group))

  write.csv(msi_int, file.path(OUTPUT_DIR, "MSI_Internal.csv"), row.names = FALSE)

  p_int <- boxplot_2grp(msi_int, "MSI_Score", "Internal (n=32)") +
    labs(title = "MSI probability score (MSIsensor-RNA)")
  ggsave(file.path(OUTPUT_DIR, "MSI_Internal_plot.pdf"), p_int, width = 5, height = 6)
  message("  -> MSI_Internal_plot.pdf")

  write.csv(wilcox_tbl(msi_int, "MSI_Score", "MSIsensor-RNA", "Internal"),
            file.path(OUTPUT_DIR, "MSI_Internal_wilcoxon.csv"), row.names = FALSE)
}

# ── External ──────────────────────────────────────────────────────────────
if (!file.exists(msi_result_ext)) {
  message("  External: MSI_Results_External.txt not found — run Step 2 first.")
} else {
  message("--- MSI Step 3: plotting external ---")

  grp_ext <- setNames(
    ifelse(clin_ext$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA[
             match(valid_cols, clin_ext$Sample_ID)] == EXTERNAL_RTI_LABEL,
           "RTI", "Non-RTI"),
    valid_cols)

  msi_ext <- read_tsv(msi_result_ext, show_col_types = FALSE) %>%
    rename(Sample_ID = 1, MSI_Score = `probability_of_MSI-H`) %>%
    mutate(Group = factor(grp_ext[Sample_ID], levels = c("Non-RTI", "RTI"))) %>%
    filter(!is.na(Group))

  write.csv(msi_ext, file.path(OUTPUT_DIR, "MSI_External.csv"), row.names = FALSE)

  p_ext <- boxplot_2grp(msi_ext, "MSI_Score", "External (n=55)") +
    labs(title = "MSI probability score (MSIsensor-RNA)")
  ggsave(file.path(OUTPUT_DIR, "MSI_External_plot.pdf"), p_ext, width = 5, height = 6)
  message("  -> MSI_External_plot.pdf")

  write.csv(wilcox_tbl(msi_ext, "MSI_Score", "MSIsensor-RNA", "External"),
            file.path(OUTPUT_DIR, "MSI_External_wilcoxon.csv"), row.names = FALSE)
}

message("06_msi.R complete.")
```

07_checkpoints. R

Immune checkpoint gene expression: CD274, PDCD1, CTLA4, LAG3, HAVCR2, TIGIT — RTI vs Non-RTI comparison.

```
# ==============================================================================
# 07_checkpoints.R — Immune checkpoint gene expression: RTI vs Non-RTI
#
# Immune checkpoint proteins act as inhibitory regulators of T-cell activity.
# Tumour cells exploit these pathways to evade immune destruction. We measure
```

```
# the expression of six key checkpoint genes directly from log2(TPM+1) and
# compare RTI vs Non-RTI using the Wilcoxon rank-sum test — a non-parametric
# test chosen because expression data in small cohorts rarely follows a normal
# distribution and makes no assumption about the underlying data shape.
#
# Genes analysed:
#   CD274  — PD-L1  (programmed death-ligand 1)
#   PDCD1  — PD-1   (programmed cell death protein 1)
#   CTLA4  — CTLA-4 (cytotoxic T-lymphocyte associated protein 4)
#   LAG3   — LAG-3  (lymphocyte activation gene 3)
#   HAVCR2 — TIM-3  (T cell immunoglobulin and mucin domain 3)
#   TIGIT  — TIGIT  (T cell immunoreceptor with Ig and ITIM domains)
#
# Runs on:
#   A. Internal cohort (32 patients, log2 TPM+1, Ensembl IDs extracted directly)
#   B. External cohort (55 QC patients, raw TPM → log2 TPM+1, Hugo symbols)
#
# Outputs:
#   Checkpoints_Internal.csv / Checkpoints_Internal_panels.pdf
#   Checkpoints_External.csv / Checkpoints_External_panels.pdf
#   Checkpoints_Internal_wilcoxon.csv / Checkpoints_External_wilcoxon.csv
# ==============================================================================

source("config/config.R")
source("config/utils.R")

library(dplyr)
library(tibble)
library(ggplot2)
library(patchwork)

# ── checkpoint gene definitions ──────────────────────────────────────────────
CP_ENSEMBL <- c(
  "ENSG00000120217",  # CD274  — PD-L1
  "ENSG00000188389",  # PDCD1  — PD-1
  "ENSG00000163599",  # CTLA4  — CTLA-4
  "ENSG00000089692",  # LAG3   — LAG-3
  "ENSG00000135077",  # HAVCR2 — TIM-3
  "ENSG00000181847"   # TIGIT
)

CP_SYMBOLS <- c("CD274", "PDCD1", "CTLA4", "LAG3", "HAVCR2", "TIGIT")

# Clinical display labels for plot titles
CP_LABELS <- c(
  CD274  = "PD-L1 (CD274)",
  PDCD1  = "PD-1 (PDCD1)",
  CTLA4  = "CTLA-4 (CTLA4)",
  LAG3   = "LAG-3 (LAG3)",
  HAVCR2 = "TIM-3 (HAVCR2)",
  TIGIT  = "TIGIT"
)

# ── boxplot with clinical gene label as title ────────────────────────────────
boxplot_cp <- function(df, gene, subtitle) {
  title <- CP_LABELS[gene]
  boxplot_2grp(df, gene, subtitle) +
    labs(title = title, y = "Expression log2(TPM+1)")
}

# ==============================================================================
# A. INTERNAL
# ==============================================================================
message("--- Checkpoints: internal ---")

tpm_int  <- read.csv(INTERNAL_TPM_LOG_FILE, row.names = 1, check.names = FALSE)
meta_int <- read.csv(INTERNAL_METADATA_FILE, row.names = 1)

# Extract checkpoint genes directly by Ensembl ID — data is already log2(TPM+1)
found_ens <- intersect(CP_ENSEMBL, rownames(tpm_int))
message("  Found ", length(found_ens), "/6 checkpoint genes by Ensembl ID")
```

```
# Build per-sample dataframe
shared_int <- intersect(colnames(tpm_int), rownames(meta_int))
grp_int <- setNames(
  ifelse(as.character(meta_int[shared_int, "Sporadic_vs_Radiotherapy_induced"]) ==
INTERNAL_RTI_LABEL,
         "RTI", "Non-RTI"),
  shared_int)

# Map Ensembl IDs to Hugo symbols for column names
ens_to_sym <- setNames(
  CP_SYMBOLS[match(found_ens, CP_ENSEMBL)],
  found_ens)

cp_int <- as.data.frame(t(tpm_int[found_ens, shared_int])) %>%
  rownames_to_column("Sample_ID") %>%
  rename(!!!ens_to_sym) %>%
  mutate(Group = factor(grp_int[Sample_ID], levels = c("Non-RTI", "RTI")))

present_int <- intersect(CP_SYMBOLS, colnames(cp_int))

write.csv(cp_int, file.path(OUTPUT_DIR, "Checkpoints_Internal.csv"), row.names = FALSE)

p_int <- wrap_plots(
  lapply(present_int, \(g) boxplot_cp(cp_int, g, "Internal (n=32)")),
  ncol = 3)
ggsave(file.path(OUTPUT_DIR, "Checkpoints_Internal_panels.pdf"),
       p_int, width = 14, height = 10)

write.csv(wilcox_tbl(cp_int, present_int, "Checkpoint", "Internal"),
          file.path(OUTPUT_DIR, "Checkpoints_Internal_wilcoxon.csv"), row.names =
FALSE)
message("  -> Checkpoints_Internal_panels.pdf")

# ============================================================================
# B. EXTERNAL
# ============================================================================
message("--- Checkpoints: external ---")

clin_ext <- read.delim(EXTERNAL_CLINICAL_55PT_FILE, stringsAsFactors = FALSE)
tpm_raw  <- read.delim(EXTERNAL_TPM_FILE, stringsAsFactors = FALSE)

valid_cols <- intersect(colnames(tpm_raw), clin_ext$Sample_ID)

# Extract checkpoint genes by Hugo symbol, apply log2(TPM+1)
found_sym <- intersect(CP_SYMBOLS, tpm_raw$Hugo_Symbol)
message("  Found ", length(found_sym), "/6 checkpoint genes by Hugo symbol")

mat_cp_ext <- tpm_raw %>%
  filter(Hugo_Symbol %in% found_sym) %>%
  dplyr::select(Hugo_Symbol, all_of(valid_cols)) %>%
  group_by(Hugo_Symbol) %>%
  summarise(across(everything(), mean), .groups = "drop") %>%
  column_to_rownames("Hugo_Symbol") %>%
  as.matrix()

mat_cp_ext <- log2(mat_cp_ext + 1)

grp_ext <- setNames(
  ifelse(clin_ext$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA[
           match(valid_cols, clin_ext$Sample_ID)] == EXTERNAL_RTI_LABEL,
         "RTI", "Non-RTI"),
  valid_cols)

cp_ext <- as.data.frame(t(mat_cp_ext)) %>%
  rownames_to_column("Sample_ID") %>%
  mutate(Group = factor(grp_ext[Sample_ID], levels = c("Non-RTI", "RTI")))

present_ext <- intersect(CP_SYMBOLS, colnames(cp_ext))

write.csv(cp_ext, file.path(OUTPUT_DIR, "Checkpoints_External.csv"), row.names = FALSE)
```

```
p_ext <- wrap_plots(
  lapply(present_ext, \(g) boxplot_cp(cp_ext, g, "External (n=55)")),
  ncol = 3)
ggsave(file.path(OUTPUT_DIR, "Checkpoints_External_panels.pdf"),
       p_ext, width = 14, height = 10)

write.csv(wilcox_tbl(cp_ext, present_ext, "Checkpoint", "External"),
          file.path(OUTPUT_DIR, "Checkpoints_External_wilcoxon.csv"), row.names =
FALSE)
message("  -> Checkpoints_External_panels.pdf")

message("07_checkpoints.R complete.")
```

08_oncopredict. R

oncoPredict: drug sensitivity prediction using GDSC2 pharmacogenomics training data (ridge regression).

```
# ==============================================================================
# 08_oncopredict.R — oncoPredict: drug sensitivity prediction (GDSC2)
#
# oncoPredict (Maeser et al. 2021) predicts drug sensitivity (IC50) for each
# patient by training a ridge regression model on the GDSC2 pharmacogenomics
# dataset (cancer cell line gene expression + drug response) and projecting
# it onto patient tumour expression data. A lower predicted IC50 means the
# tumour is predicted to be more sensitive to that drug. We compare predicted
# IC50 between RTI and Non-RTI using the Wilcoxon rank-sum test and visualise
# the top differentially sensitive drugs as a lollipop plot.
#
# NOTE: This script takes ~2 hours per cohort. Run with nohup:
#   nohup Rscript scripts/06_downstream/08_oncopredict.R > logs/oncopredict.log 2>&1 &
#
# Requires GDSC2 training files in DATA_DIR (download from osf.io/c6tfx):
#   GDSC2_Expr_RMA_Log.rds
#   GDSC2_Res.rds
#
# Runs on:
#   A. Internal cohort (32 patients, log2 TPM+1, Ensembl IDs → Hugo via org.Hs.eg.db)
#   B. External cohort (55 QC patients, raw TPM → log2 TPM+1, Hugo symbols)
#
# Outputs:
#   OncoPredict_Internal.csv / OncoPredict_Internal_lollipop.pdf
#   OncoPredict_External.csv / OncoPredict_External_lollipop.pdf
#   OncoPredict_Internal_stats.csv / OncoPredict_External_stats.csv
# ==============================================================================

source("config/config.R")
source("config/utils.R")

library(oncoPredict)
library(AnnotationDbi)
library(org.Hs.eg.db)
library(dplyr)
library(tibble)
library(ggplot2)

# ── check GDSC2 training files ────────────────────────────────────────────────
if (!file.exists(GDSC2_EXPR_FILE) || !file.exists(GDSC2_RES_FILE)) {
  stop("GDSC2 training files not found in DATA_DIR.\n",
       "Download from: https://osf.io/c6tfx/\n",
       "Expected:\n  ", GDSC2_EXPR_FILE, "\n  ", GDSC2_RES_FILE)
}

trainingExprData <- readRDS(GDSC2_EXPR_FILE)
trainingPtype    <- readRDS(GDSC2_RES_FILE)

# ── run calcPhenotype and return IC50 dataframe ──────────────────────────────
run_oncopredict <- function(expr_matrix, grp, cohort_label) {
  message("  Running calcPhenotype (", ncol(expr_matrix), " patients)...")
  calcPhenotype(
```

```
    trainingExprData         = trainingExprData,
    trainingPtype            = trainingPtype,
    testExprData             = expr_matrix,
    batchCorrect             = "eb",
    powerTransformPhenotype  = TRUE,
    removeLowVaryingGenes    = 0.2,
    minNumSamples            = 10,
    printOutput              = FALSE,
    removeLowVaringGenesFrom = "homogenizeData"
  )
  ic50 <- read.csv("calcPhenotype_Output/DrugPredictions.csv",
                   row.names = 1, check.names = FALSE) %>%
    rownames_to_column("Sample_ID") %>%
    mutate(Group = factor(grp[Sample_ID], levels = c("Non-RTI", "RTI")))
  write.csv(ic50, file.path(OUTPUT_DIR, paste0("OncoPredict_", cohort_label, ".csv")),
            row.names = FALSE)
  ic50
}

# ── Wilcoxon per drug → stats table ──────────────────────────────────────
drug_stats <- function(ic50_df, cohort_label) {
  drug_cols <- setdiff(colnames(ic50_df), c("Sample_ID", "Group"))
  drug_cols <- drug_cols[sapply(ic50_df[drug_cols], is.numeric)]

  stats <- bind_rows(lapply(drug_cols, function(d) {
    rti     <- na.omit(ic50_df[[d]][ic50_df$Group == "RTI"])
    non_rti <- na.omit(ic50_df[[d]][ic50_df$Group == "Non-RTI"])
    if (length(rti) < 3 || length(non_rti) < 3) return(NULL)
    w <- wilcox.test(rti, non_rti, exact = FALSE)
    data.frame(
      Drug              = d,
      P_Value           = round(w$p.value, 4),
      More_Sensitive_In = ifelse(mean(rti) < mean(non_rti), "RTI", "Non-RTI"),
      RTI_Mean          = round(mean(rti), 4),
      NonRTI_Mean       = round(mean(non_rti), 4)
    )
  }))
  stats$FDR    <- round(p.adjust(stats$P_Value, method = "BH"), 4)
  stats$Log10P <- -log10(stats$P_Value)
  write.csv(stats, file.path(OUTPUT_DIR, paste0("OncoPredict_", cohort_label,
"_stats.csv")),
            row.names = FALSE)
  stats
}

# ── lollipop plot: top 10 RTI-sensitive + top 10 Non-RTI-sensitive ──────────
lollipop_plot <- function(stats, cohort_label) {
  top_rti    <- stats %>% filter(More_Sensitive_In == "RTI")    %>% arrange(P_Value)
%>% head(10)
  top_nonrti <- stats %>% filter(More_Sensitive_In == "Non-RTI") %>% arrange(P_Value)
%>% head(10)

  top20 <- bind_rows(top_rti, top_nonrti) %>%
    mutate(
      Directional = ifelse(More_Sensitive_In == "RTI", Log10P, -Log10P),
      Drug        = reorder(Drug, Directional)
    )

  n_label <- ifelse(cohort_label == "Internal", "n=32", "n=55")

  ggplot(top20, aes(x = Drug, y = Directional, color = More_Sensitive_In)) +
    geom_segment(aes(xend = Drug, y = 0, yend = Directional), linewidth = 1.2) +
    geom_point(size = 4) +
    geom_hline(yintercept = 0, linewidth = 0.8) +
    coord_flip() +
    scale_color_manual(
      values = c("RTI" = COL_RTI, "Non-RTI" = COL_NONRTI),
      name   = "More sensitive in:") +
    theme_classic(base_size = 12) +
    labs(
```

```
      title    = paste("Top 20 differentially sensitive drugs —", cohort_label,
paste0("(", n_label, ")")),
      subtitle = "Lower IC50 = more sensitive | oncoPredict GDSC2",
      x        = "GDSC2 drug",
      y        = "Directional significance (-log10 p)") +
    theme(
      legend.position = "top",
      plot.title      = element_text(face = "bold"),
      axis.text       = element_text(color = "black", face = "bold"))
}

# ============================================================================
# A. INTERNAL
# ============================================================================
message("--- oncoPredict: internal ---")

tpm_int  <- read.csv(INTERNAL_TPM_LOG_FILE, row.names = 1, check.names = FALSE)
meta_int <- read.csv(INTERNAL_METADATA_FILE, row.names = 1)
mat_int  <- as.matrix(tpm_int)

# Ensembl → Hugo via org.Hs.eg.db (offline)
syms_int <- mapIds(org.Hs.eg.db,
                   keys      = rownames(mat_int),
                   column    = "SYMBOL",
                   keytype   = "ENSEMBL",
                   multiVals = "first")

mat_int_sym <- as.data.frame(mat_int) %>%
  mutate(SYM = syms_int) %>%
  filter(!is.na(SYM), SYM != "") %>%
  group_by(SYM) %>%
  summarise(across(everything(), mean), .groups = "drop") %>%
  column_to_rownames("SYM") %>%
  as.matrix()

shared_int <- intersect(colnames(mat_int_sym), rownames(meta_int))
grp_int <- setNames(
  ifelse(as.character(meta_int[shared_int, "Sporadic_vs_Radiotherapy_induced"]) ==
INTERNAL_RTI_LABEL,
         "RTI", "Non-RTI"),
  shared_int)

ic50_int   <- run_oncopredict(mat_int_sym[, shared_int], grp_int, "Internal")
stats_int  <- drug_stats(ic50_int, "Internal")
p_loll_int <- lollipop_plot(stats_int, "Internal")
ggsave(file.path(OUTPUT_DIR, "OncoPredict_Internal_lollipop.pdf"),
       p_loll_int, width = 12, height = 10)
message("  -> OncoPredict_Internal_lollipop.pdf")

# ============================================================================
# B. EXTERNAL
# ============================================================================
message("--- oncoPredict: external ---")

clin_ext <- read.delim(EXTERNAL_CLINICAL_55PT_FILE, stringsAsFactors = FALSE)
tpm_raw  <- read.delim(EXTERNAL_TPM_FILE, stringsAsFactors = FALSE)

valid_cols <- intersect(colnames(tpm_raw), clin_ext$Sample_ID)

mat_ext <- tpm_raw %>%
  dplyr::select(Hugo_Symbol, all_of(valid_cols)) %>%
  group_by(Hugo_Symbol) %>%
  summarise(across(where(is.numeric), mean), .groups = "drop") %>%
  column_to_rownames("Hugo_Symbol") %>%
  as.matrix()

mat_ext <- log2(mat_ext + 1)

grp_ext <- setNames(
  ifelse(clin_ext$MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA[
           match(valid_cols, clin_ext$Sample_ID)] == EXTERNAL_RTI_LABEL,
```

```
        "RTI", "Non-RTI"),
  valid_cols)

ic50_ext   <- run_oncopredict(mat_ext, grp_ext, "External")
stats_ext  <- drug_stats(ic50_ext, "External")
p_loll_ext <- lollipop_plot(stats_ext, "External")
ggsave(file.path(OUTPUT_DIR, "OncoPredict_External_lollipop.pdf"),
       p_loll_ext, width = 12, height = 10)
message("  -> OncoPredict_External_lollipop.pdf")

message("08_oncopredict.R complete.")
```

A.9 Clinical Analysis

scripts/07_clinical/01_clinical. R

Clinical analysis: age at diagnosis, Kaplan-Meier disease-specific survival (internal), and latency from radiotherapy (external).

```
# ============================================================================
# 09_clinical.R — Clinical analysis: internal and external cohorts
#
# Internal cohort:
#   - Age at diagnosis (boxplot)
#   - Disease-specific survival (Kaplan-Meier)
#
# External cohort:
#   - Age at diagnosis (boxplot from binned data)
#   - Latency: prior radiotherapy → angiosarcoma diagnosis (RTI only, lollipop)
#
# Outputs:
#   Clinical_Internal_age.pdf
#   Clinical_Internal_KM.pdf
#   Clinical_Internal.csv
#   Clinical_External_age.pdf
#   Clinical_External_latency.pdf
#   Clinical_External.csv
#   Latency_External.csv
# ============================================================================

source("config/config.R")
source("config/utils.R")

library(ggplot2)
library(dplyr)
library(survival)
library(survminer)

# ── shared publication theme ──────────────────────────────────────────────
theme_pub <- theme_classic(base_size = 13) +
  theme(
    plot.title      = element_text(face = "bold", size = 14),
    plot.subtitle   = element_text(size = 10, color = "grey40"),
    axis.text       = element_text(color = "black", size = 11),
    legend.position = "top",
    legend.title    = element_blank()
  )

# ============================================================================
# A. INTERNAL
# ============================================================================
message("--- Clinical: internal ---")

clin <- read.csv(INTERNAL_CLINICAL_FILE, stringsAsFactors = FALSE) %>%
  mutate(
    Group = factor(
      ifelse(Sporadic_vs_Radiotherapy_induced.x == INTERNAL_RTI_LABEL,
             "RTI", "Non-RTI"),
      levels = c("Non-RTI", "RTI")),
    Birth_date_real = as.Date(as.numeric(Birth_date), origin = "1899-12-30"),
```

```
    Diag_date_real  = as.Date(Angiosarcoma_diagnosis_date),
    Age             = as.numeric(difftime(Diag_date_real, Birth_date_real,
                                          units = "days")) / 365.25,
    Last_date_real  = as.Date(as.numeric(Last_contact_or_Death_date),
                              origin = "1899-12-30"),
    OS_months       = as.numeric(difftime(Last_date_real, Diag_date_real,
                                          units = "days")) / 30.44,
    event_disease   = as.integer(Last_known_vital_status == "Dead_(disease)")
  )

message("  n=", nrow(clin),
        "  RTI=", sum(clin$Group == "RTI"),
        "  Non-RTI=", sum(clin$Group == "Non-RTI"))

n_counts  <- clin %>% count(Group)
label_map <- setNames(paste0(n_counts$Group, "\n(n=", n_counts$n, ")"),
                      n_counts$Group)

# ── Age at diagnosis — descriptive boxplot, no p-value ──────────────────────
p_age_int <- ggplot(clin, aes(x = Group, y = Age, fill = Group)) +
  geom_boxplot(alpha = 0.6, outlier.shape = NA, width = 0.5, linewidth = 0.8) +
  geom_jitter(shape = 21, fill = "white", color = "black",
              width = 0.1, size = 2) +
  scale_fill_manual(values = GROUP_COLS) +
  scale_x_discrete(labels = label_map) +
  labs(
    title    = "Age at angiosarcoma diagnosis",
    subtitle = "Internal cohort — all female patients",
    x        = NULL,
    y        = "Age (years)") +
  theme_pub +
  theme(legend.position = "none")

ggsave(file.path(OUTPUT_DIR, "Clinical_Internal_age.pdf"),
       p_age_int, width = 5, height = 6)
message("  -> Clinical_Internal_age.pdf")

# ── Kaplan-Meier disease-specific survival ──────────────────────────────────
clin_surv <- clin %>% filter(!is.na(OS_months), !is.na(event_disease))
message("  KM n=", nrow(clin_surv), "  events=", sum(clin_surv$event_disease))

km_fit <- survfit(Surv(OS_months, event_disease) ~ Group, data = clin_surv)

p_km <- ggsurvplot(
  km_fit,
  data                = clin_surv,
  palette             = GROUP_COLS,
  pval                = TRUE,
  pval.size           = 4.5,
  conf.int            = TRUE,
  conf.int.alpha      = 0.15,
  risk.table          = TRUE,
  risk.table.height   = 0.25,
  risk.table.fontsize = 3.5,
  xlab                = "Time (months)",
  ylab                = "Disease-specific survival",
  title               = "Disease-specific survival — RTI vs Non-RTI (Internal)",
  legend.labs         = c("Non-RTI", "RTI"),
  legend.title        = "",
  ggtheme             = theme_pub,
  tables.theme        = theme_classic(base_size = 11)
)

pdf(file.path(OUTPUT_DIR, "Clinical_Internal_KM.pdf"), width = 7, height = 7)
print(p_km)
dev.off()
message("  -> Clinical_Internal_KM.pdf")

write.csv(
  clin %>% dplyr::select(Sample_ID, Group, Age, OS_months, event_disease),
  file.path(OUTPUT_DIR, "Clinical_Internal.csv"),
```

```
  row.names = FALSE)

# ============================================================================
# B. EXTERNAL
# ============================================================================
message("--- Clinical: external ---")

ext_clin <- read.delim(EXTERNAL_CLINICAL_55PT_FILE,
                       stringsAsFactors = FALSE) %>%
  rename(SampleID = 1)

clin_ext <- read.table(EXTERNAL_CLINICAL_PATIENT,
                       header = TRUE, sep = "\t",
                       stringsAsFactors = FALSE,
                       comment.char = "#") %>%
  rename(SampleID = 1)

# ── convert age bins to midpoints ──────────────────────────────────────────
bin_to_mid <- function(x) {
  sapply(x, function(b) {
    if (is.na(b))    return(NA)
    if (b == "80+")  return(82)
    if (b == "0-20") return(10)
    mean(as.numeric(strsplit(b, "-")[[1]]))
  })
}

ext_55 <- ext_clin %>%
  left_join(clin_ext %>% dplyr::select(SampleID, SEX,
                                       MRD_AGE_AT_DIAGNOSIS_GROUP),
            by = "SampleID") %>%
  mutate(
    Group   = factor(
      ifelse(MRD_RADIATION_ASSOCIATED_ANGIOSARCOMA == EXTERNAL_RTI_LABEL,
             "RTI", "Non-RTI"),
      levels = c("Non-RTI", "RTI")),
    Age_mid = bin_to_mid(MRD_AGE_AT_DIAGNOSIS_GROUP)
  )

message("  n=", nrow(ext_55),
        "  RTI=", sum(ext_55$Group == "RTI"),
        "  Non-RTI=", sum(ext_55$Group == "Non-RTI"))

n_counts_ext  <- ext_55 %>% count(Group)
label_map_ext <- setNames(paste0(n_counts_ext$Group, "\n(n=", n_counts_ext$n, ")"),
                          n_counts_ext$Group)

# ── Age at diagnosis — descriptive boxplot, no p-value ─────────────────────
p_age_ext <- ggplot(ext_55, aes(x = Group, y = Age_mid, fill = Group)) +
  geom_boxplot(alpha = 0.6, outlier.shape = NA, width = 0.5, linewidth = 0.8) +
  geom_jitter(shape = 21, fill = "white", color = "black",
              width = 0.1, size = 2) +
  scale_fill_manual(values = GROUP_COLS) +
  scale_x_discrete(labels = label_map_ext) +
  labs(
    title    = "Age at angiosarcoma diagnosis",
    subtitle = "External cohort (n=55) — age shown as bin midpoint",
    x        = NULL,
    y        = "Age (years)") +
  theme_pub +
  theme(legend.position = "none")

ggsave(file.path(OUTPUT_DIR, "Clinical_External_age.pdf"),
       p_age_ext, width = 5, height = 6)
message("  -> Clinical_External_age.pdf")

write.csv(
  ext_55 %>% dplyr::select(SampleID, Group, Age_mid, SEX,
                           MRD_AGE_AT_DIAGNOSIS_GROUP),
  file.path(OUTPUT_DIR, "Clinical_External.csv"),
  row.names = FALSE)
```

```
# ── Latency: prior RT → angiosarcoma (RTI patients only) ──────────────────────
tl_file <- file.path(DATA_DIR, "data_timeline_treatment.txt")

if (!file.exists(tl_file)) {
  message("  Latency skipped: data_timeline_treatment.txt not found in DATA_DIR.")
} else {
  tl <- read.table(tl_file, header = TRUE, sep = "\t",
                   stringsAsFactors = FALSE, comment.char = "#") %>%
    rename(PATIENT_ID = 1)

  breast_locs <- c("LEFT BREAST", "RIGHT BREAST",
                   "LEFT CHEST WALL", "RIGHT CHEST WALL",
                   "SUPRACLAVICULAR REGION")

  rti_patients <- ext_55 %>% filter(Group == "RTI") %>% pull(SampleID)

  latency_df <- tl %>%
    filter(
      PATIENT_ID %in% rti_patients,
      TREATMENT_TYPE == "RADIATION THERAPY",
      START_DATE < 0,
      LOCATION %in% breast_locs
    ) %>%
    group_by(PATIENT_ID) %>%
    slice_max(START_DATE, n = 1) %>%
    ungroup() %>%
    mutate(Latency_years = abs(START_DATE) / 365.25) %>%
    arrange(Latency_years) %>%
    mutate(PATIENT_ID = factor(PATIENT_ID, levels = PATIENT_ID))

  message("  RTI patients with latency data: ", nrow(latency_df))

  p_lat <- ggplot(latency_df, aes(x = Latency_years, y = PATIENT_ID)) +
    geom_segment(aes(x = 0, xend = Latency_years,
                     y = PATIENT_ID, yend = PATIENT_ID),
                 color = COL_RTI, linewidth = 0.8) +
    geom_point(color = COL_RTI, size = 3.5) +
    geom_vline(xintercept = median(latency_df$Latency_years),
               linetype = "dashed", color = "grey40", linewidth = 0.6) +
    annotate("text",
             x     = median(latency_df$Latency_years) + 0.3,
             y     = 0.9,
             label = paste0("Median: ",
                            round(median(latency_df$Latency_years), 1), " yrs"),
             hjust = 0, size = 3.8, color = "grey30") +
    labs(
      title    = "Latency: prior RT to angiosarcoma diagnosis",
      subtitle = paste0("RTI patients — External cohort (n=",
                        nrow(latency_df), ")"),
      x        = "Latency (years)",
      y        = NULL) +
    theme_pub +
    theme(axis.text.y = element_text(size = 9),
          legend.position = "none")

  ggsave(file.path(OUTPUT_DIR, "Clinical_External_latency.pdf"),
         p_lat, width = 6, height = 8)

  write.csv(
    latency_df %>% dplyr::select(PATIENT_ID, START_DATE, LOCATION, Latency_years),
    file.path(OUTPUT_DIR, "Latency_External.csv"),
    row.names = FALSE)

  message("  -> Clinical_External_latency.pdf")
}

message("09_clinical.R complete.")
```

A.10 R Session Information

envs/session_info.txt

R version, package versions, and system information for full reproducibility.

```
# ================================================================================
# envs/session_info.txt
# R session information captured on the analysis server (alucard).
# Regenerate after a full pipeline run with:
#   sink("envs/session_info.txt"); sessionInfo(); sink()
# ================================================================================

# This file will be populated after the first complete pipeline run.
# It records the exact R version, package versions, and system info used
# for the analysis to ensure reproducibility.

# Expected key packages and minimum versions:
#   DESeq2          >= 1.42.0
#   apeglm          >= 1.24.0
#   clusterProfiler >= 4.10.0
#   msigdbr         >= 7.5.1
#   ComplexHeatmap  >= 2.18.0
#   MCPcounter      >= 1.2.0
#   estimate        >= 1.0.13
#   oncoPredict     >= 0.2
#   singscore       >= 1.22.0
#   survival        >= 3.5.7
#   survminer       >= 0.4.9
#   ggplot2         >= 3.5.0
#   ggrepel         >= 0.9.5
#   patchwork       >= 1.2.0
#   biomaRt         >= 2.58.0
#   org.Hs.eg.db    >= 3.18.0
```