Heterogeneity in the gene regulatory landscape of leiomyosarcoma.

Tatiana Belova[1], Nicola Biondi[2,†], Ping-Han Hsieh[1,†], Priya Chudasama[2] and Marieke L. Kuijjer[1,3,4]
[1]*Computational Biology and Systems Medicine Group, Centre for Molecular Medicine Norway, University of Oslo, Oslo, Norway*
[2]*Precision Sarcoma Research Group, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases, Heidelberg, Germany*
[3]*Department of Pathology, Leiden University Medical Center, Leiden, the Netherlands*
[4]*Leiden Center for Computational Oncology, Leiden University Medical Center, Leiden, the Netherlands*
†*Shared authorship*

Soft-tissue sarcomas are group of rare but highly aggressive malignancies. It is a tremendously heterogeneous group of tumors. Characterizing inter-tumor heterogeneity is crucial for selecting suitable cancer therapy as the presence of diverse molecular subgroups of patients can be associated with disease outcome or response to treatment. However, no methods have been developed to characterize heterogeneity based on genome-wide patient-specific regulatory networks. In this work, we propose a simple but efficient approach to characterize inter-tumor regulatory network heterogeneity, which we call PORCUPINE (Principal Components Analysis to Obtain Regulatory Contributions Using Pathway-based Interpretation of Network Estimates). PORCUPINE uses as input individual patient regulatory networks, represented by estimated regulatory interactions between transcription factors and their target genes, and a list of genes assigned to biological pathways in order to identify key pathways that drive heterogeneity among individuals. We used PORCUPINE to model regulatory heterogeneity in leiomyosarcoma, one of the most common soft-tissue sarcomas subtypes. We applied it to 80 genome-wide leiomyosarcoma regulatory networks modeled on data from The Cancer Genome Atlas and validated the results in an independent dataset of 37 leiomyosarcoma cases from the German Cancer Research Center. PORCUPINE identified 37 pathways, including pathways that represent potential targets for treatment of subgroups of leiomyosarcoma patients, such as FGFR and CTLA4 inhibitory signaling. PORCUPINE thereby provides a robust way of analyzing and interpreting patient-specific regulatory networks and is the first step towards implementing network-informed personalized medicine in leiomyosarcoma.

I. INTRODUCTION

Soft-tissue sarcomas are a group of rare but highly aggressive malignancies. While they account for less than 1% of all malignant tumors, soft-tissue sarcomas are a tremendously heterogeneous group of tumors and include more than 150 different histological subtypes [1]. Partly because of this heterogeneity, significant challenges exist in the management of sarcomas. Most soft-tissue sarcomas are treated similarly in the clinic, regardless of their site of origin [2]. Surgery with or without radiotherapy is the main treatment for localized disease. Several clinical trials have been conducted in soft-tissue sarcomas. However, until recently such trials included patients with many different histological subtypes in the same cohort, causing difficulties to conclude on the efficacy of these therapies in the individual subtypes [3]. Differences in clinical response among soft-tissue sarcoma subtypes lead to newer studies that only enrolled patients of certain histological subtypes [3], which have shown to result in better response and disease control.

Over the past years it has become evident that treatments tailored to a single patient, or group of patients belonging to a specific molecular subtype of cancer, can result in major improvements in cancer outcomes [4]. Characterizing inter-patient molecular tumor heterogeneity was shown to be crucial for selecting the most efficient cancer therapy, and the presence of diverse molecular subtypes can predict patient survival [5] and relapse or resistance to treatment [6]. Therefore, it is clear that the integration of personalized medicine into cancer treatment strategies requires extensive knowledge of inter-patient variability. Patients can, for example, be grouped into molecular subtypes based on "omics" data, including gene expression, microRNA, DNA methylation, somatic mutations, or proteomic profiles.

The molecular landscape of soft-tissue sarcomas has been characterized in several studies [7–10]. The Cancer Genome Atlas (TCGA) sarcoma project, one of the largest sarcoma sequencing projects to-date, performed a comprehensive and integrated analysis of 206 adult soft-tissue sarcomas, represented by six major subtypes, and showed that sarcomas vary greatly at the genetic, epigenetic, and transcriptomic levels [7]. More recently, some histological subtypes of soft-tissue sarcomas were further delineated into molecular subgroups according to

their genomic and transcriptomic profiles. For example, Guo *et al.*, characterized three molecular subtypes of leiomyosarcoma (LMS)—one of the most common subtypes of soft-tissue sarcomas—based on transcriptomic data. One of these subtypes was over-represented by uterine leiomyosarcoma, while the other two were over-represented in extra-uterine sites [11]. While these subtypes were not associated with tumor grade, they were somewhat related to patient survival.

Gene regulatory networks offer an in-depth view on the mechanisms that drive gene expression, but have so-far not been modeled for individual sarcoma patients. Through modeling interactions between transcription factors (TFs) or other regulators and their potential target genes, gene regulatory networks are useful tools to study regulatory landscapes. Recently, integrative methods have been developed that model these networks genome-wide. One of these methods is PANDA, which integrates putative TF-DNA binding with protein-protein interactions and target gene co-expression to infer a regulatory network for a specific condition [12]. More recently, we developed the LIONESS algorithm that can be combined with PANDA to infer patient-specific regulatory networks [13]. These patient-specific network models have been instrumental in capturing sex differences in gene regulation in healthy tissues [14] and colon cancer [15], as well as regulatory differences between glioblastoma patients with short-term and long-term survival [16].

In this work, we set out to map the genome-wide regulatory landscapes of 206 individual sarcomas obtained from TCGA by modeling large-scale gene regulatory networks using PANDA and LIONESS [7]. Analysis of heterogeneity among gene regulatory networks can facilitate stratification of patients into novel regulatory subtypes and identification of the regulatory programs that drive such heterogeneity. To characterize this inter-tumor regulatory heterogeneity, we propose a simple but efficient approach, which we call PORCUPINE (Principal Components Analysis to Obtain Regulatory Contributions Using Pathway-based Interpretation of Network Estimates). PORCUPINE detects statistically significant, key regulatory pathways that drive hetetero-geneity among patients.

We perform a detailed analysis of inter-patient heterogeneity in leiomyosarcoma by applying PORCUPINE to 80 genome-wide leiomyosarcoma regulatory networks modeled on data from TCGA (referred to below as TCGA-LMS). We validated the pathways detected by PORCUPINE in an independent dataset consisting of 37 leiomyosarcoma cases from the German Cancer Research Center (referred to below as DKFZ-LMS) [17]. This identified 37 shared pathways that define gene regulatory heterogeneity in both datasets, including pathways that play a known role in leiomyosarcoma biology, as well as pathways that have not been described before in the disease. Newly identified pathways, including FGFR signaling and CTLA4 inhibitory signaling, represent potential targets for treatment of subgroups of leiomyosarcoma patients. In addition, we show that the heterogeneity identified with PORCUPINE was not associated with methylation profiles or clinical features, thereby suggesting an independent mechanism of patient heterogeneity driven by the complex landscape of gene regulatory interactions.

MATERIALS AND METHODS

Gene expression data preprocessing

We downloaded expression data for all TCGA cases using the "recount" package in R [18]. The transcriptome data for 37 leiomyosarcoma cases obtained from the German Cancer Research Center (DKFZ) was preprocessed by the Omics IT and Data Management Core Facility (DKFZ ODCF) using the One Touch Pipeline [19]. We performed batch correction on the raw expression counts of the set of 206 TCGA soft-tissue sarcomas and the 37 DKFZ-LMS samples together, using the "Combat-seq" package in Bioconductor [20]. We then combined Combat-seq-adjusted counts with the raw expression counts of the remaining TCGA samples and performed smooth quantile normalization using "qsmooth" package in Bioconductor to preserve global differences in gene expression between the different cancer types [21], specifying each cancer type as a separate group level. Samples of 206 TCGA soft-tissue sarcomas and 37 DKFZ-LMS samples were specified as the same "soft-tissue sarcoma" group level.

Construction of individual patient gene regulatory networks

We used the MATLAB version of the PANDA network reconstruction algorithm (available in the netZoo repository `https://github.com/netZoo/netZooM`) to estimate an "aggregate" gene regulatory network, based on a total of 11,321 samples, 17,899 genes, and 623 TFs. These samples included 206 TCGA and 37 DKFZ soft-tissue sarcomas—remaining samples represented other cancer types available in TCGA. We used the entire TCGA dataset to build the aggregate network, as we previously found that LIONESS' estimates of single-sample edge weights are more robust when including a large, heterogeneous background of samples [13].

PANDA builds an aggregate network by incorporating three types of data—a "prior" regulatory network, which is based on a transcription factor motif scan to identify putative regulatory interactions between TFs and their target genes, protein-protein (PPI) interactions between TFs, and target gene expression data. The prior gene regulatory network was generated using a set of transcription factor motifs obtained from the Catalogue of Inferred Sequence Binding Preferences (CIS-

BP) [22], as described by Sonawane *et al.*, 2017 [23]. These motifs were scanned to promoters as described previously [24]. The prior network was intersected with the expression data to include genes and transcription factors with available expression data and at least one significant promoter hit. This resulted in initial map representing potential regulatory interactions between 623 transcription factors and 17,899 target genes. An initial protein-protein network was estimated between all TFs from motif prior map using interaction scores from StringDb v10 [25], which were scaled to be within range of [0,1], where self-interactions were set equal to one, as described previously [23]. To reconstruct patient-specific gene regulatory networks, we applied the LIONESS equation in MATLAB (available in the netZoo repository `https://github.com/netZoo/netZooM`).

UMAP visualization

To visualize the clustering distribution of the 206 TCGA soft-tissue sarcoma patient-specific gene regulatory networks, we applied dimensionality reduction with Uniform Manifold Approximation and Projection (UMAP), using the "uwot" package in R 3.6.1, setting the number of nearest neighbours to 20. We performed UMAP on the matrix of gene targeting scores obtained from the 206 individual sarcoma networks. Gene targeting scores are defined as the sum of all edge weights pointing to a gene and represent the amount of regulation a gene receives from the entire set of TFs available in a network [26]. These scores have previously been used to identify gene regulatory differences in various studies [15, 16, 26]. We visualized the results in two-dimensional UMAP space. To identify clusters in the data, we used the DBSCAN clustering algorithm on the UMAP coordinates from the first two embeddings [27], with the parameter "minPts" set to five.

Identifying regulatory heterogeneity using PORCUPINE

To capture inter-patient heterogeneity (referred to below as "heterogeneity") at the gene regulatory level, we developed a computational framework, which we call PORCUPINE. PORCUPINE is a Principal Components Analysis (PCA)-based approach that can be used to identify key pathways that drive heterogeneity among individuals in a dataset. It determines whether a specific set of variables—for example a set of genes in a specific pathway—have coordinated variability in their regulation.

PORCUPINE uses as input individual patient networks, for example networks modeled using PANDA and LIONESS, as well as a .gmt file (in MSigDb file format [28]) that includes biological pathways and the genes belonging to them. For each pathway, it extracts all edges connected to the genes belonging to that pathway and scales each edge across individuals. It then performs a PCA analysis on these edge weights, as well as on a null background that is based on random pathways. For the randomization (permutation), PORCUPINE creates a set of 1000 gene sets equal in size to the pathway of interest, where genes are randomly selected from all genes present in the .gmt file. The edges connected to these genes are then extracted. The amount of variance explained by the first principal component (PC1) in the pathway of interest is then compared to the amount of variance explained by PC1 in the random (permuted) data. To identify significant pathways, PORCUPINE applies a one-tailed t-test and calculates the effect size (ES). The latter is calculated as the difference between the variance explained by PC1 of the pathway of interest and the mean of the variance explained by PC1 corresponding to the random sets of pathways, divided by standard deviation of the variance explained by PC1 in the random sets using the cohensD function in the "lsr" package in R. P-values are adjusted for multiple testing with the Benjamini-Hochberg method [29] and significant pathways are returned based on user-defined thresholds of adjusted p-value and effect size. We developed PORCUPINE as R package and it is available as open-source code on GitHub (`https://github.com/kuijjerlab/PORCUPINE`).

We applied PORCUPINE to TCGA and DKFZ leiomyosarcoma data using Reactome pathways v7.1 from MSigDb, excluding pathways that consisted of more than 200 genes. Pathways with adjusted p-value less than 0.01 and effect size >=2 were reported as significant. As the number of genes in each pathway is different, we investigated whether the obtained results were biased towards pathways of smaller size. To test this, we split pathways in four groups based on their size, namely pathways containing less than 50, 50-100, 100-150, 150-200 genes. We then calculated the proportions of these groups among Reactome pathways and among the set of deregulated pathways identified in the TCGA-LMS and DKFZ-LMS datasets.

Clustering of pathways and identification of redundant aspects of gene regulatory heterogeneity

To investigate potential redundant patterns of heterogeneity captured by pathways identified with PORCUPINE, we computed the Pearson correlation coefficient for every pair of individuals for each pathway, based on the individual's TF-target edge weights in that pathway. We then combined pathway-level inter-individual correlations into a matrix for all pathways and performed clustering, visualizing the results using the "ComplexHeatmap" package in R. Additionally, to identify pathways with overlapping genes, we computed the Jaccard similarity between pairs of pathways.

Identification of top ranked target genes and transcription factors

To identify those genes and TFs that contribute most to the pathway's significance, we extracted the edge loadings of the first principal component (referred to below as the "edge contribution score"). Because the sum of the squares of all edge contribution scores for an individual principal component must be one, we calculated the expected edge contribution score, assuming that all edges contributed equally to that principal component. Edges with a contribution score $> 1.5\times$ the expected score were regarded as important contributors to that principal component. To identify transcription factors with many co-regulated genes, we then grouped transcription factors corresponding to these top edges according to the number of their targets.

Association of the significant pathways with clinical phenotypes

To investigate whether the heterogeneity captured by each pathway was associated with clinical features, we performed an association analysis of the coordinates of patients on the first principal component in each pathway (referred to below as the "pathway-based patient heterogeneity score") with the clinical data available for these patients. Clinical features for TCGA leiomyosarcoma patients were obtained using the "TCGAbiolinks" package from Bioconductor [30]. Clinical information for 37 DKFZ patients was obtained from the study by Chudasama *et al.* [17]. Since the clinical attributes represent a mix of categorical and numerical features, we applied Kruskal-Wallis and Pearson correlation tests for categorical and numerical features, respectively. We corrected p-values for multiple testing using the Benjamini-Hochberg approach and applied a threshold of 0.05 to identify significant associations.

In order to determine whether any of the identified pathways were associated with patient survival, we used the first principal component from each pathway in a Cox regression model to predict patient survival.

Association of the significant pathways with pathway-based mutation profiles

We downloaded and preprocessed leiomyosarcoma mutation data as previously described in Kuijjer *et al.* [31]. We used the SAMBAR algorithm [31] to obtain patient-specific pathway mutation scores for TCGA-LMS patients. Among 1,455 pathways, 954 pathways had mutation scores larger than zero in the TCGA-LMS dataset. To assess the association between pathways identified with PORCUPINE and these pathways mutation scores, we used a Kruskal Wallis test, comparing the pathway-based patient heterogeneity scores on the first principal component between two groups, i.e. mutated vs not mutated, for each mutated pathway. We used FDR <0.05 as threshold for reporting significant differences between the groups.

Association of the identified pathways with overall methylation profiles

DNA methylation data measured on the Illumina Infinium Human Methylation 450 BeadChip platform were downloaded for all sarcoma patients available in TCGA using the Bioconductor "TCGA biolinks" package in R. We downloaded raw methylation IDAT files and performed preprocessing and normalization with subset-quantile within array normalization (SWAN) using Bioconductor package "minfi." [32]. We calculated overall methylation profiles for each individual by using the mean value across all probes. We then correlated these values to the pathway-based patient heterogeneity scores in each pathway. Associations with FDR <0.05 were considered significant.

Validation of the pathways in healthy tissues

We obtained patient-specific regulatory networks for healthy smooth-muscle–derived tissues, represented by esophageal muscularis and uterus from the Genotype-Tissue Expression (GTEx) project, through the GRAND database of gene regulatory network models [33]. In total, 283 and 90 patient-specific networks were available for esophageal muscularis and uterus, respectively. We applied PORCUPINE to evaluate gene regulatory heterogeneity among the individuals in the merged set of 373 networks.

RESULTS

Pan-sarcoma clustering of patient-specific regulatory networks

In this study, we modeled genome-wide, patient-specific gene regulatory networks for 206 TCGA sarcoma patients using two computational algorithms, PANDA and LIONESS (Figure 1).

These patient-specific networks include information on likelihoods of regulatory interactions (represented as edge weights) between 623 transcription factors and 17,899 target genes. To explore and visualize patient heterogeneity based on the regulatory landscape of sarcomas, we first calculated gene targeting scores in these networks (see Methods), and then used uniform manifold approximation and projection (UMAP) for visualization. To investigate whether regulatory profiles cluster differently than expression data, we also performed UMAP on the expression data (Figure 2).

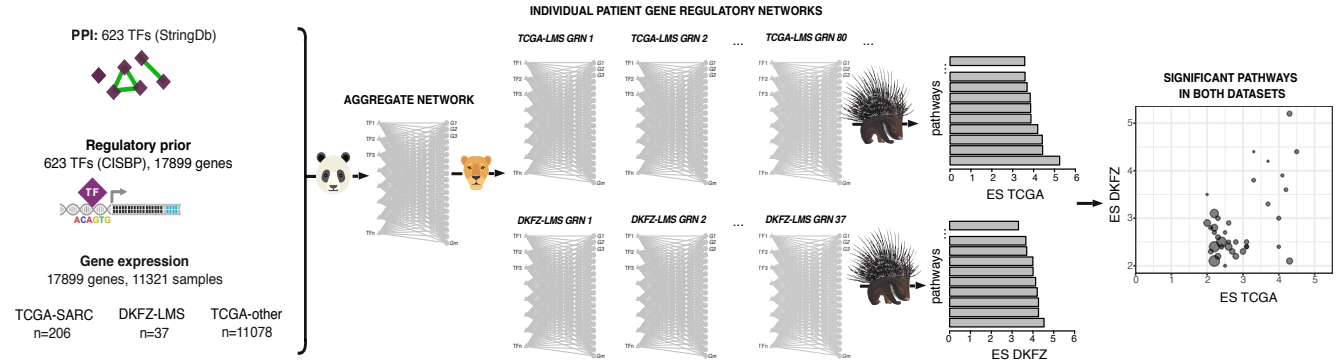


Figure 1. Schematic overview of the study. We modeled individual patient gene regulatory networks for leiomyosarcoma patients from two datasets (TCGA and DKFZ) with PANDA and LIONESS, integrating information on protein-protein interactions (PPI) between transcription factors (TF), prior information on TF-DNA motif binding, and gene expression data. We then developed and applied a new computational comparative network analysis tool (PORCUPINE) to identify significant pathways that capture heterogeneity in gene regulation across these datasets.

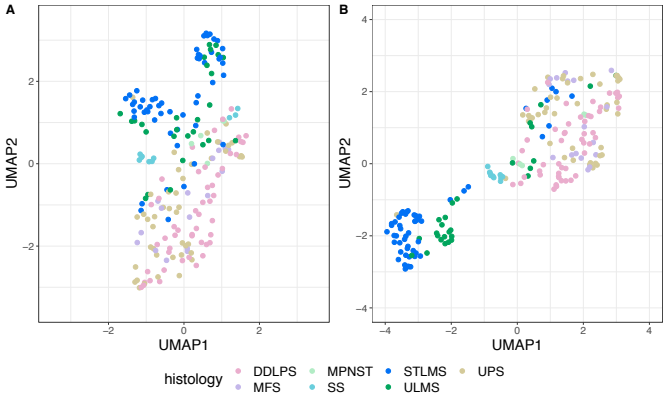


Figure 2. UMAP visualization of the distribution of 206 soft-tissue sarcomas, representing seven different histological subtypes (indicated with different colors) based on (A) gene targeting scores (B) expression.

In both cases, it is clear that the majority of leiomyosarcoma, represented by uterine (ULMS) and soft-tissue leiomyosarcoma (STLMS), cluster separately from other sarcoma subtypes, with a more distinct separation observed in the gene expression profiles. Interestingly, co-localization of uterine and soft-tissue leiomyosarcomas was different between the two UMAP embeddings. While ULMS samples show separation from STLMS in the expression data, clustering of leiomyosarcoma based on gene regulatory networks did not show such a separation. This indicates that, despite the apparent differences in gene expression between the two tissue-sites where leiomyosarcoma can develop, tumors from these sites do not have clearly distinct regulatory profiles. Additionally, the regulatory networks capture heterogeneity among leiomyosarcoma tumors that is not directly obvious from analysis of the expression data alone.

The remaining sarcoma subtypes were more spread out, with no clear co-localization of the gene regulatory networks derived from the same sarcoma histological subtype in distinct clusters, except for the cases of synovial sarcoma (SS). With the use of DBSCAN in 2D UMAP space, we next clustered the gene regulatory profiles for 206 sarcoma cases and identified ten clusters (Figure S1). Two clusters were mainly represented by leiomyosarcoma samples and contained 60% of all leiomyosarcoma samples. The remaining leiomyosarcoma samples were part of mixed clusters or belong to unclustered data.

In-depth analysis of gene regulatory heterogeneity in leiomyosarcoma with PORCUPINE

The distinct regulatory clusters we identified in leiomyosarcoma motivated us to perform an in-depth analysis of the regulatory heterogeneity of leiomyosarcoma. To do this, we developed PORCUPINE, a novel pathway-based approach that can be applied to patient-specific gene regulatory networks to identify biological pathways that capture regulatory heterogeneity in a patient population (Figure 3). PORCUPINE examines a pre-defined set of pathways, e.g. pathways from published resources such as Reactome [34], and then identifies those pathways that show a statistically significant excess of coordinated variability in gene regulation across individuals. The method performs PCA on all estimated regulatory interactions connected to genes from a specific pathway, obtained from a cohort of patient-specific networks. It then compares the variance captured by the first principal component in a pathway to the amount of variance that would be expected by chance. This process is repeated for each pathway. Significant pathways can then be selected based on user-defined thresholds of adjusted p-value and effect size.

We applied PORCUPINE to a set of 80 patient-specific leiomyosarcoma gene regulatory networks modeled on

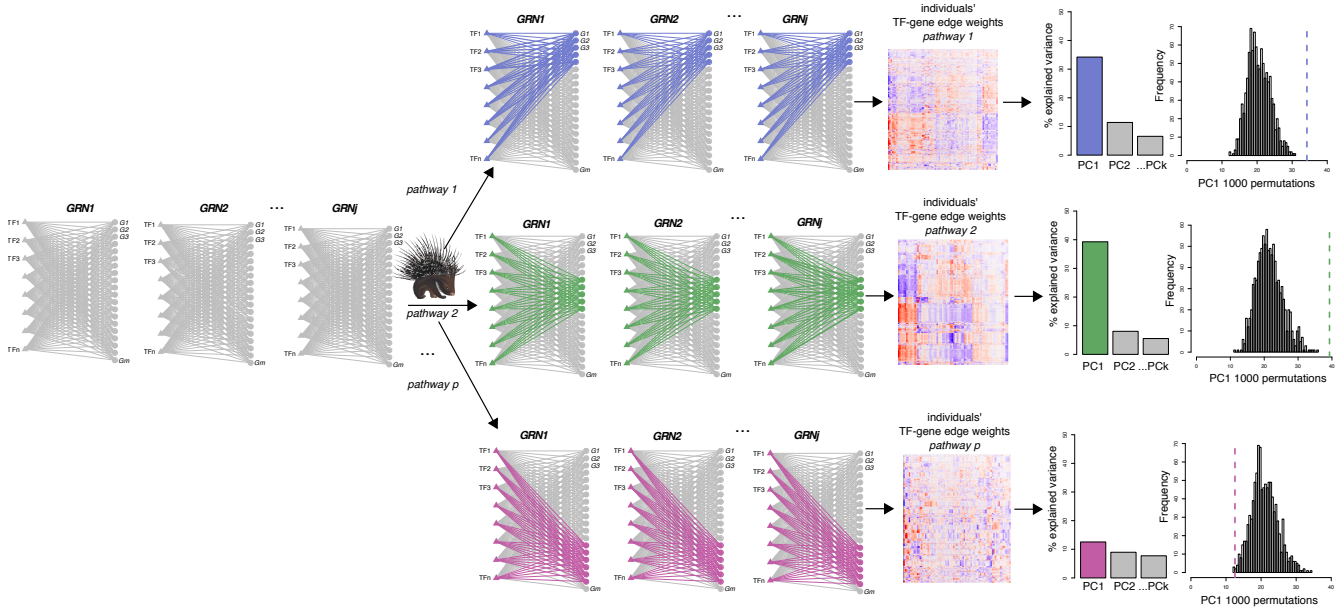


Figure 3. Overview of PORCUPINE (PCA to Obtain Regulatory Contributions Using Pathway-based Interpretation of Network Estimates). PORCUPINE applies the following steps: 1) TF-gene edge weight information is extracted from each individual gene regulatory network for all genes belonging to a certain pathway; 2) Principal Component Analysis is performed on the pathway-associated TF-gene weight matrix. The variance explained by the first principal component is extracted; 3) The amount of variance explained by PC1 is compared to the expected amount of variance explained, which is obtained by applying PCA on edge weights connected to 1,000 randomly generated gene sets of the same size as the selected pathway. Effect size is calculated. These steps are repeated for each pathway. P-values obtained from step 3 are then corrected for multiple testing with the Benjamini-Hochberg method.

data from TCGA, using 1,455 Reactome pathways from MSigDb (see Methods). This identified 72 significant pathways (adjusted p-value less than 0.01 and effect size >=2). We validated these results in an independent set of patient-specific networks modeled on 37 leiomyosarcoma samples from DKFZ. In the validation dataset, we identified 91 pathways, of which 37 were also identified in the networks modeled on TCGA. The overlap of 37 pathways is higher than expected by chance, with p-value <9.522e-29 based on a hypergeometric test, indicating that PORCUPINE's results are robust and highly reproducible across networks modeled on independent datasets. The 37 pathways that were detected in both datasets are shown in Figure 3, with corresponding effect sizes.

Notably, the pathways PORCUPINE identified varied in size, indicating that PORCUPINE analysis is not biased towards pathways of smaller or larger size (Table S1).

PORCUPINE identifies regulatory heterogeneity in pathways with known and new roles in leiomyosarcoma

The two most significant pathways that were identified in both datasets are "Inhibition of replication initiation of damaged DNA by RB1/E2F1" and "E2F mediated regulation of DNA replication," containing 13 and 22 genes, respectively. A closer examination of the genes in these pathways shows that all 13 genes in the first pathway are also part of the second pathway. PORCUPINE provides evidence of a coordinated change in the regulation of multiple genes in these pathways, which is not directly captured by expression data (Figure S2). These pathways are leiomyosarcoma-relevant, given that leiomyosarcomas are characterized by a high frequency of alterations in tumor suppressor gene *RB1*, which negatively regulates transcription factor E2F1 [17].

The 37 pathways can be further grouped into subcategories according to their cellular function (see Figure 4). Pathways with genes involved in cell cycle and signal transduction were the most frequent subcategories. Among others, we identified pathways such as "Negative regulation of FGFR2 signaling," "FGFRL1 modulation of FGFR1 signaling," "ERKs are inactivated," and "SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription."

Fibroblast growth factor receptors (FGFR) are tyrosine kinase receptors that are involved in several biological functions including regulation of cell growth, proliferation, survival, differentiation, and angiogenesis. Aberrant FGFR signaling has been shown to be associated with several human cancers and thus FGFRs are attractive druggable targets [35]. To our knowledge, among members of the FGFR family, only

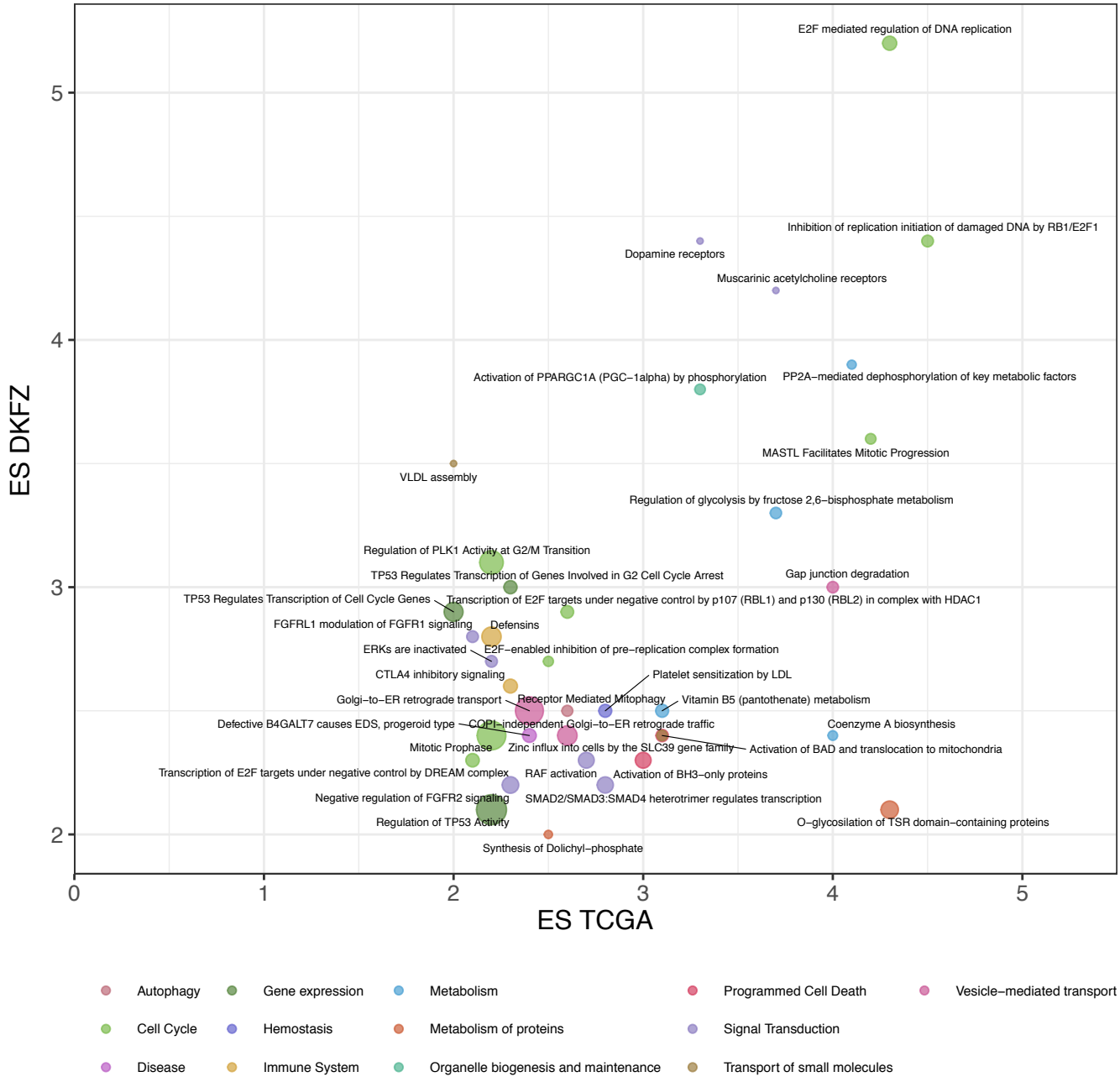


Figure 4. Pathways identified with PORCUPINE in both leiomyosarcoma datasets, based on FDR <0.05 and effect size >2. Pathways are colored according to their cellular function, with the size of the bubble reflecting the size of—or number of genes in—a pathway.

the inhibition of FGFR1 has been investigated in a patient with metastatic leiomyosarcoma, which showed clinical improvement [36]. However, there is an ongoing clinical trial testing the selective pan-FGFR inhibitor Rogaratinib, to treat patients with advanced sarcoma with alterations in FGFR 1-4 [37].

Two pathways were associated with TP53 regulation, including "TP53 regulates transcription of genes involved in G2 cell cycle arrest" and "TP53 regulates transcription of cell cycle genes." *TP53* mutations are frequently identified in leiomyosarcoma. It was shown that *TP53* is predictive of response to VEGFR inhibition in advanced sarcomas [38]. Gendicine, which represents a recombinant adenoviral vector expressing wild-type p53, was approved by the China Food and Drug Administration and is used in combination with chemotherapy to treat uterine leiomyosarcoma [39].

Other categories include pathways involved in

metabolism and metabolism of proteins and programmed cell death such as "PP2A-mediated dephosphorylation of key metabolic factors" and "Activation of BH3-only proteins."

Two pathways associated with immune system function were identified—"CTLA-4 inhibitory signaling" and "Defensins." CTLA-4 is an immune checkpoint, and monocolonal antibodies such as ipilimumab and tremelimumab have been developed to target CTLA-4. These CTLA-4 inhibitors have already been used in clinical studies for treatment of several cancers, including melanoma, mesothelioma, breast cancer, prostate cancer, pancreatic cancer, hepatocellular carcinoma, and non-small cell lung cancer [40]. The efficacy of immunotherapy with CTLA-4 inhibitors in sarcoma has only been evaluated in one study to-date, in which six patients with synovial sarcoma were treated with ipilimumab [41]. However, treatment with anti-CTLA-4 did not result in an immunological antitumor response and the disease progressed rapidly in all patients. To our knowledge, no clinical results testing the effect of anti-CTLA-4 in leiomyosarcoma are available or exist to-date.

In addition to the pathways described above, several pathways were associated with vesicle-mediated transport and transport of small molecules, among them the "Zinc influx into cells by the SLC39 gene family" pathway. The solute carrier (SLC) superfamily contains a large diversity of membrane-bound transporters, which mediate transport of substrates across various cellular membranes. The SLC39 gene family controls the influx of zinc and has been shown to play a critical role in several cancers. For example, upregulation of *SLC39A4* was associated with increased cell migration, cisplatin resistance, and poor survival in lung cancer [42]. Increased expression of *SLC39A6* was associated with shorter overall survival in esophageal carcinoma [43] and decreased expression of *SLC39A14* with aggressive tumor progression in prostate cancer [44]. To our knowledge, there are no studies regarding the role of the SLC39 gene family in sarcomas.

To evaluate whether the identified pathways capture similar patterns of regulatory heterogeneity, we performed clustering of pathways based on inter-individual correlations of edge weights, as described in the Methods section (Figure S3). Pathways were grouped into three main clusters, where pathways in each cluster stratified leiomyosarcoma tumors into similar subtypes. The clustering of pathways we observed was partly explained by gene overlap in these pathways (Figure S3). Pathways in cluster 1 had highest gene overlap (mean Jaccard index 0.18), followed by cluster 2 (mean Jaccard index 0.04), with almost no gene overlap between pathways in cluster 3 (mean Jaccard index 0.006) (Figure S3). Shared patterns of heterogeneity between pathways without apparent gene sharing can also indicate a higher order of co-regulation of these pathways. An example is the pathway "Negative regulation of FGFR2 signaling," which belongs to cluster 1, however, based on its Jaccard indices, this pathway does not cluster with remaining pathways from the same cluster.

Gene regulatory heterogeneity is not associated with clinical features in leiomyosarcoma

To investigate if the identified pathways were associated with clinicopathological features, we performed an association analysis of the pathway-based patient heterogeneity scores with clinical features available from the TCGA and DKFZ resources (Figure S4). As shown in Figure S4, there were no significant associations between the clinical features and the pathway-based patient heterogeneity scores on the first principal component (at FDR <5%).

To determine whether any of the identified pathways were related to patient survival, we used the pathway-based patient heterogeneity scores on the first principal component in Cox regression models to predict patient outcome. We did not identify any significant associations with survival.

Major genes and transcription factors contributing to leiomyosarcoma heterogeneity

We next selected those regulatory interactions in each of the 37 pathways that contributed most to the regulatory heterogeneity we observed in leiomyosarcoma (see Methods). Across all pathways, genes including *PPP2R1A, PPP2CB, TFDP2, CCNB1*, and *RB1* were frequently found among the top targets (Figure 5, Supplementary file 1). These genes are related to cell proliferation and growth. Noteworthy, *PPP2R1A* was among the top contributors in 13 out of 37 pathways and may therefore be a key player in driving leiomyosarcoma heterogeneity. It encodes for a subunit of protein phosphatase 2 (PP2), which plays a role in the negative control of cell growth and division. PP2A inactivation is a crucial step in malignant development [45]. It was previously shown that *PPP2R1A* mutation is frequent in uterine cancers [46]. However, we did not identify any association between the histological subtype of leiomyosarcoma and gene regulatory heterogeneity in pathways that had *PPP2R1A* among their major contributors. We also did not identify any significant association of patient heterogeneity scores with *PPP2R1A* mutation profiles, indicating that regulatory heterogeneity of *PPP2R1A* is not driven by mutations in the gene itself.

In addition to reporting the top target genes, we identified top transcription factors contributing to regulatory heterogeneity in each pathway. Transcription factors with coordinated variability in regulation of an enriched number of targets are shown in Figures 5C and S5. As can be seen from these figures, some TFs had a limited number of targets that they regulate in a coordinated manner, such as in seen in the pathway

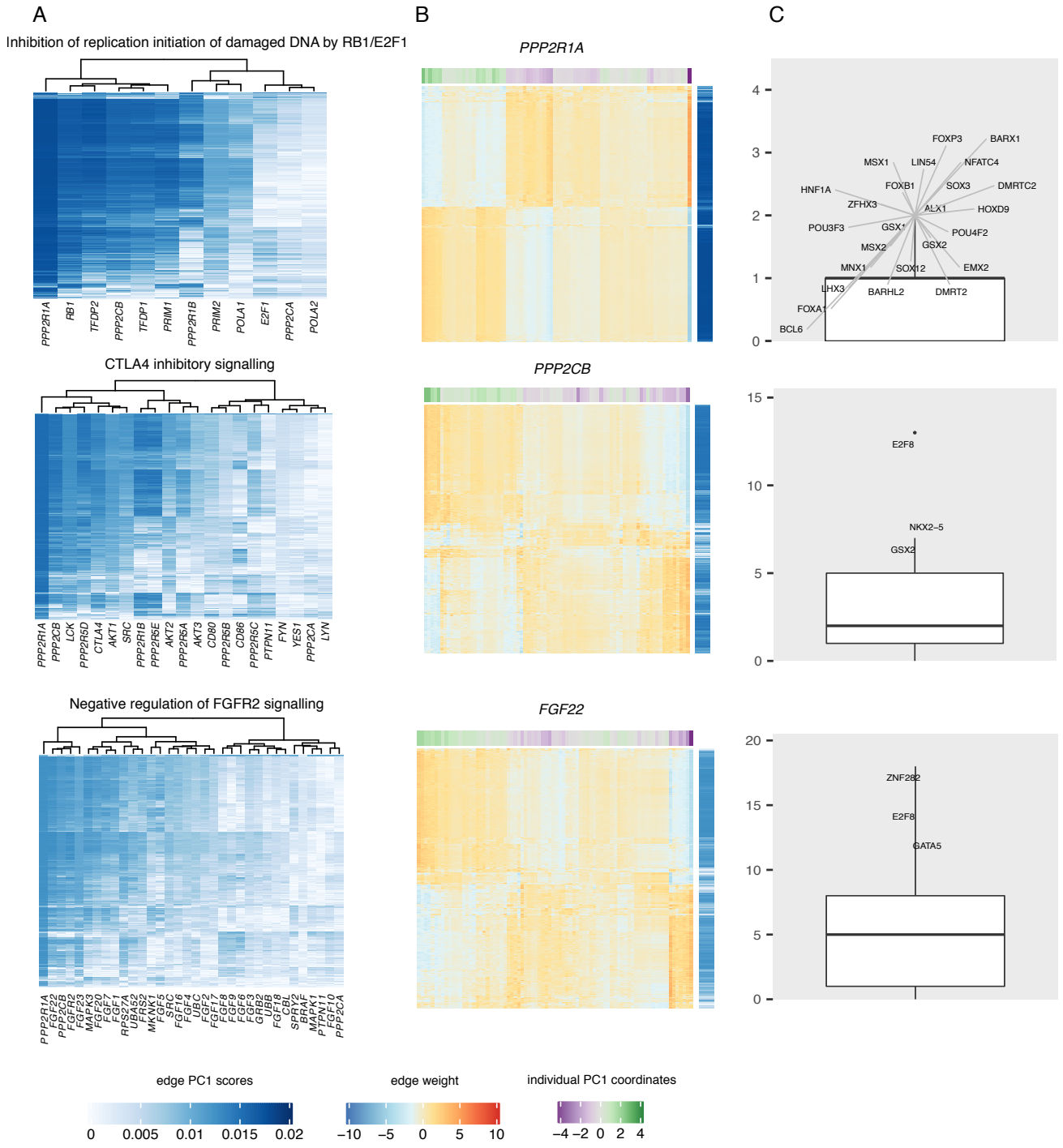


Figure 5. A. Heatmaps showing the contribution scores of genes and all TFs to the first principal component in three selected, significant pathways. B. Heatmaps showing the edge weights of selected genes to all TFs in these pathways. Edge weights are scaled across individuals. Row annotation shows the edge contribution scores to PC1 in each pathway. Column annotation indicates the patient heterogeneity scores in each pathway. C. Boxplots showing the number of targets for transcription factors with top edge contribution scores to PC1 in each pathway. Transcription factors with a number of targets greater than the 95th percentile in each pathway are labelled.

"Inhibition of replication initiation of damaged DNA by RB1/E2F1." Other TFs, such as ZNF282 and E2F8 in "Negative regulation of FGFR2 signalling," have a large number of targets (Figure 5C). Transcription factors that are frequent top regulators of heterogeneity among the identified pathways are E2F8, ZNF282, EMX2, GSX2, BARX1, and HNF1A. E2F8 and ZNF282 were the most frequent TFs that connected to a large number of targets across all identified pathways.

The E2F family of transcription factors contains eight members that play central roles in many biological processes, including cell proliferation, differentiation, DNA repair, cell cycle, and apoptosis. Several studies have shown that dysregulation of E2F8 is associated with oncogenesis and tumor progression in many cancers. For example, it was shown that expression of E2F8 is associated with tumor progression in breast cancer [47], human hepatocellular carcinoma [48], and lung cancer [49]. However, not much is known about the role and clinical significance of E2F8 in leiomyosaroma, nor in other sarcomas. Also, the role of ZNF282 (Zinc finger protein 282) in human cancers, including sarcomas, is unknown. In a study by Yeo *et al.* , it was shown that ZNF282 overexpression was associated with poor survival in esophageal squamous cell carcinoma, and depletion of ZNF282 inhibited cell cycle progression, migration, and invasion of cancer cells [50]. Additionally, the authors provided evidence that ZNF282 functions as an E2F1 co-activator in esophageal squamous cell carcinoma cells, highlighting a potential connection between this transcription factor and E2F signaling.

To illustrate the change in regulation of targets of these important transcription factors, we visualized E2F8 targets in the pathway "CTLA4 inhibitory signalling." Figure 6 shows the coordinated change in regulation of 13 genes in this pathway. Many of these genes show similar pattern of change in targeting by E2F8 across individuals. Interestingly, two genes—*AKT2* and *AKT3*—show an opposite regulatory pattern compared to other E2F8 target genes across the individuals.

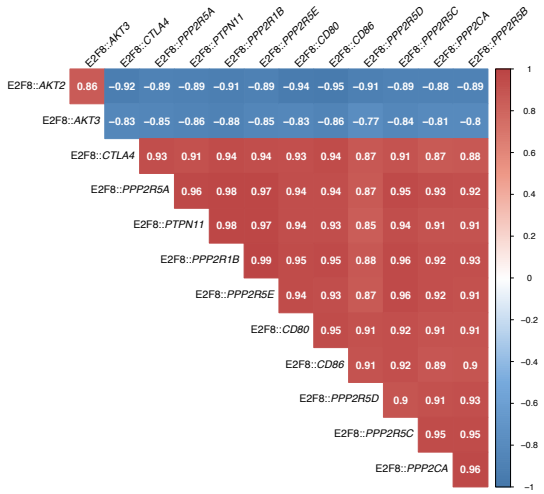


Figure 6. Pearson correlations among edge weights of target genes of the transcription factor "E2F8" in the pathway "CTLA4 inhibitory signalling."

Association of heterogeneously regulated pathways with pathway mutation scores

To evaluate if any of the identified pathways could classify patients with similar mutational profiles, we associated the first principal component in these pathways with pathway mutation scores. To do so, we downloaded and processed mutation data obtained from leiomyosarcoma tumors from TCGA (available for 72/80 patients) as described in Kuijjer *et al.* [31]. We performed a Kruskal Wallis test to compare the pathway-based patient heterogeneity scores on the first principal component in each of the 37 pathways between two groups, i.e. mutated compared to not mutated, for each mutated pathway. No significant differences were identified (FDR <0.05), indicating that the separation of leiomyosarcoma patients identified with PORCUPINE is independent of their mutation profiles and thus potentially a new, mutation-independent mechanism driving patient heterogeneity.

Regulatory heterogeneity is independent of epigenetic heterogeneity in leiomyosarcoma

To investigate if the patient heterogeneity profiles were associated with inter-individual differences in the tumor's methylation profiles, we performed correlation analysis of the pathway-based patient heterogeneity scores on PC1 with overall DNA methylation profiles of individual tumors. There were no significant associations (FDR <0.05), indicating that regulatory heterogeneity is independent of methylation status.

Validation of the identified pathways in healthy tissues

Finally, to explore if the 37 pathways we identified were cancer-specific, we assessed gene regulatory heterogeneity in healthy smooth muscle–derived tissues, represented by esophageal muscularis and uterus. In total, 283 esophageal muscularis and 90 uterus sample-specific gene regulatory networks, modeled with PANDA and LIONESS, were available from the GTEx project through the GRAND database [33]. We used PORCUPINE to characterize regulatory heterogeneity in this dataset. Among the 37 pathways identified to drive leiomyosarcoma heterogeneity, only one pathway, i.e "Gap junction degradation" was significant in these healthy tissues, showing that 36/37 pathways we identified are leiomyosarcoma-specific and that gene regulatory heterogeneity in these pathways likely develops during

sarcomagenesis.

Discussion

Soft-tissue sarcomas exhibit substantial heterogeneity in their clinical behavior, which exists not only between the known histological subtypes of these rare cancers, but also between tumors of the same histological subtype. This extreme heterogeneity represents a major challenge in the treatment of patients with the disease [51]. Understanding this heterogeneity at the molecular level may help explain variability in drug response between patients and potentially lead to the identification of new targets for treatment.

Genomic molecular heterogeneity, as well as heterogeneity in gene expression and DNA methylation levels of soft-tissue sarcomas have previously been described between different histological subtypes [7–10], as well as for tumors belonging to the same histological subtype [11]. However, classification of patients on the basis of gene regulatory networks has the potential to provide additional, novel information to stratify patients into clinically meaningful subgroups, to point to potential new targets for treatment, and to identify new biomarkers to guide selecting patients most likely to benefit from a specific treatment.

In this work, we profiled the genome-wide regulatory landscapes of 206 sarcomas obtained from TCGA, performing an in-depth analysis of heterogeneity occurring between individual gene regulatory profiles of leiomyosarcoma patients. To this end, we developed a novel pathway-based approach to analyze genome-wide networks. This approach, which we call PORCUPINE, identifies key pathways contributing to heterogeneity in gene regulation among the individuals in a dataset. PORCUPINE captures the coordinated variation of multiple functionally-related genes in a pathway by estimating if the variance explained by the first principal component is higher than expected by chance. Similar approaches have previously been successfully applied to study heterogeneity in cancer using gene expression profiles [52]. However, our approach differs from these methods as we specifically designed it to analyze large-scale, genome-wide gene regulatory networks.

We developed PORCUPINE as user-friendly R package that can be applied to single-sample networks. While we used PORCUPINE on networks modeled with PANDA and LIONESS, the tool is not limited to these specific methodologies, and could potentially also be used to analyze (bipartite) networks modeled with other single-sample approaches. Of course, when applying PORCUPINE, sufficient sample size and independent validation datasets, as we showed here by including an independent leiomyosarcoma dataset from DKFZ, are important to include to detect relevant and robust pathways. Additionally, it is important to note that, while the use of a large set of randomized pathways is beneficial, it comes with disadvantage of an increase in computational load.

Genome-wide gene regulatory networks represent high-dimensional data. Usually, network summary statistics, such as gene targeting scores, closeness centrality, or betweenness centrality, are calculated prior to any further analysis to reduce the dimensions of large-scale networks. Then, to identify heterogeneity across a cohort, unsupervised clustering approaches are widely used [53]. The advantage of PORCUPINE is that it can be directly applied to high-dimensional networks, as it uses as input the network's edge weights instead of a summary statistic. Moreover, as it does this per individual biological pathway, the output is not just a collection of significant differential edges that need to be further analyzed, but rather a list of differentially regulated pathways that are easy to interpret. Additionally, the method can capture significant aspects of heterogeneity among individuals in situations when no clear population structure with well defined clusters can be revealed. PORCUPINE estimates pathway-based patient heterogeneity scores that can facilitate the identification of either continuous gradients or discrete gene regulatory subtypes and that can be further used in association analyses with clinical covariates, or in survival analyses, as we have shown in this work.

Applying PORCUPINE to individual leiomyosarcoma patient networks modeled in two different cohorts, we identified 37 pathways that capture gene regulatory heterogeneity in leiomyosarcoma. Among these, we identified several pathways that could represent potential targets for treatment of subgroups of leiomyosarcoma patients, including RB1/E2F1 signaling as well as pathways involved in FGFR signaling and CTLA4 inhibitory signaling. The RB1/E2F1 pathway is essential in regulating cell growth and apoptosis, and is known to be disrupted in many cancers including leiomyosarcoma [54], where it is affected by frequent alterations in *RB1* gene. Therefore, the components of this pathway represent appealing targets for cancer therapy. RB1 pathway disruption has already been shown to be associated with resistance to therapies in several cancers, including sarcomas. For example, a study by Francis *et al.* (2017) showed that only Rb-positive sarcoma cells were sensitive to CDK4/6 inhibitor-based therapies [55]. In addition, a recent study by Hemming *et al.* [56] in patient-derived xenograft models of leiomyosarcoma showed that CDK inhibitors were found to inhibit tumor growth through decreasing expression levels of *E2F1* and other genes involved in the E2F-driven oncogenic transcriptional program.

Aberrant FGFR signaling has been associated with several human cancers and its inhibition is effectively applied in targeted therapies with small-molecule inhibitors. Its clinical significance has been described in several soft-tissue sarcomas, including synovial sarcoma, dedifferentiated liposarcoma, and other soft-tissue sarcoma types [57]. Pazopanib is a multitarget tyrosine kinase inhibitor and has inhibitory effect against VEGFR1

and VEGFR3, as well as other closely-related receptor-tyrosine kinases, including FGFR1, and may be used to treat advanced leiomyosarcoma [58]. It was also shown that Pazopanib provides comparable response in patients with uterine and non-uterine leiomyosarcoma [59].

In addition to targeting signal transduction pathways, immunotherapies may provide new treatment options and are under active investigation in sarcomas. Unfortunately, a study with anti-CTLA4, which showed great success in other cancer types, did not provide response in sarcoma patients [41]. However, the patients enrolled in this trial had synovial sarcoma, and to our knowledge no clinical studies that enrolled leiomyosarcoma patients and used anti-CTLA-4 treatment are available or exist so far. Stratifying patients based on CTLA-4 gene regulatory profiles could identify subgroups of patients that are more likely to respond to anti-CTLA4 treatment.

Finally, using PORCUPINE, we could also highlight genes and transcription factors that are important drivers of heterogeneity among leiomyosarcoma patients, including *RB1* and *PPP2R1A* as target genes, as well as the transcription factors E2F8 and ZNF282, which could potentially also be inhibited [60].

In summary, PORCUPINE allows to uncover patterns of inter-patient heterogeneity at the level of transcriptional regulation in the cell and identify pathways that can be potentially be targeted in the clinic. It thereby provides one of the first steps towards implementing network-informed personalized medicine in sarcomas.

ACKNOWLEDGEMENTS

The authors would like to thank Jing Yang and the Omics IT and Data Management Core Facility (ODCF) of the DKFZ for help with data preprocessing and transferring, as well as Ingrid Kjelsvik and Elisa Bjørgo for administrative support.

[1] Board WCoTE. Soft Tissue and Bone Tumours. vol. 3. 5th ed.; 2020.

[2] George S, Serrano C, Hensley ML, Ray-Coquard I. Soft Tissue and Uterine Leiomyosarcoma. Journal of Clinical Oncology. 2017;36(2):JCO.2017.75.984.

[3] Nakano K, Takahashi S. Precision Medicine in Soft Tissue Sarcoma Treatment. Cancers. 2020;12(1):221.

[4] Krzyszczyk P, Acevedo A, Davidoff EJ, Timmins LM, Marrero-Berrios I, Patel M, et al. The growing role of precision and personalized medicine for cancer treatment. TECHNOLOGY. 2019;06(03n04):79–100.

[5] Vijver MJvd, He YD, Veer LJvt, Dai H, Hart AAM, Voskuil DW, et al. A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. The New England Journal of Medicine. 2002;347(25):1999–2009.

[6] Zaretsky JM, Garcia-Diaz A, Shin DS, Escuin-Ordinas H, Hugo W, Hu-Lieskovan S, et al. Mutations Associated with Acquired Resistance to PD-1 Blockade in Melanoma. The New England Journal of Medicine. 2016;375(9):819–829.

[7] Network TCGAR, Abeshouse A, Adebamowo C, Adebamowo SN, Akbani R, Akeredolu T, et al. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. Cell. 2017;171(4):950–965.e28.

[8] Chibon F, Lagarde P, Salas S, Pérot G, Brouste V, Tirode F, et al. Validated prediction of clinical outcome in sarcomas and multiple types of cancer on the basis of a gene expression signature related to genome complexity. Nature Medicine. 2010;16(7):781–787.

[9] Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, O'Connell JX, et al. Molecular characterisation of soft tissue tumours: a gene expression study. The Lancet. 2002;359(9314):1301–1307.

[10] Segal NH, Pavlidis P, Antonescu CR, Maki RG, Noble WS, DeSantis D, et al. Classification and Subtype Prediction of Adult Soft Tissue Sarcoma by Functional Genomics. The American Journal of Pathology. 2003;163(2):691–700.

[11] Guo X, Jo VY, Mills AM, Zhu SX, Lee CH, Espinosa I, et al. Clinically Relevant Molecular Subtypes in Leiomyosarcoma. Clinical Cancer Research. 2015;21(15):3501–3511.

[12] Glass K, Huttenhower C, Quackenbush J, Yuan GC. Passing Messages between Biological Networks to Refine Predicted Interactions. PLoS ONE. 2013;8(5):e64832.

[13] Kuijjer ML, Tung MG, Yuan G, Quackenbush J, Glass K. Estimating Sample-Specific Regulatory Networks. iScience. 2019;14:226–240.

[14] Lopes-Ramos CM, Chen CY, Kuijjer ML, Paulson JN, Sonawane AR, Fagny M, et al. Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues. Cell Reports. 2020;31(12):107795.

[15] Lopes-Ramos CM, Kuijjer ML, Ogino S, Fuchs CS, DeMeo DL, Glass K, et al. Gene Regulatory Network Analysis Identifies Sex-Linked Differences in Colon Cancer Drug Metabolism. Cancer Research. 2018;78(19):5538–5547.

[16] Lopes-Ramos CM, Belova T, Brunner TH, Guebila MB, Osorio D, Quackenbush J, et al. Regulatory network of PD1 signaling is associated with prognosis in glioblastoma multiforme. Cancer Research. 2021;81(21):canres.0730.2021.

[17] Chudasama P, Mughal SS, Sanders MA, Hübschmann D, Chung I, Deeg KI, et al. Integrative genomic and transcriptomic analysis of leiomyosarcoma. Nature Communications. 2018;9(1):144.

[18] Collado-Torres L, Nellore A, Jaffe AE. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. F1000Research. 2017;6:1558.

[19] Reisinger E, Genthner L, Kerssemakers J, Kensche P, Borufka S, Jugold A, et al. OTP: An automatized system for managing and processing NGS data. Journal of Biotechnology. 2017;261:53–62.

[20] Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genomics and Bioinformatics. 2020;2(3):lqaa078–.

[21] Hicks SC, Okrah K, Paulson JN, Quackenbush J, Irizarry RA, Bravo HC. Smooth quantile normalization. Biostatistics. 2017;19(2):185–198.

[22] Weirauch M, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell. 2014;158(6):1431–1443.

[23] Sonawane AR, Platig J, Fagny M, Chen CY, Paulson JN, Lopes-Ramos CM, et al. Understanding Tissue-Specific Gene Regulation. Cell Reports. 2017;21(4):1077–1088.

[24] Hill KE, Kelly AD, Kuijjer ML, Barry W, Rattani A, Garbutt CC, et al. An imprinted non-coding genomic cluster at 14q32 defines clinically relevant molecular subtypes in osteosarcoma across multiple independent datasets. Journal of Hematology & Oncology. 2017;10(1):107.

[25] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research. 2015;43(D1):D447–D452.

[26] Weighill D, Guebila MB, Glass K, Platig J, Yeh JJ, Quackenbush J. Gene Targeting in Disease Networks. Frontiers in Genetics. 2021;12:649942.

[27] A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.

[28] Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov J, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Systems. 2015;1(6):417–425.

[29] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995;57(1):289–300.

[30] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research. 2016;44(8):e71–e71.

[31] Kuijjer ML, Paulson JN, Salzman P, Ding W, Quackenbush J. Cancer subtype identification using somatic mutation data. British Journal of Cancer. 2018;118(11):1492–1501.

[32] Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363–1369.

[33] Ben Guebila M, Lopes-Ramos CM, Weighill D, Sonawane A, Burkholz R, Shamsaei B, et al. GRAND: a database of gene regulatory network models across human conditions. Nucleic Acids Research. 2021;50(D1):gkab778–.

[34] Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. Nucleic Acids Research. 2016;44(D1):D481–D487.

[35] Porta R, Borea R, Coelho A, Khan S, Araújo A, Reclusa P, et al. FGFR a promising druggable target in cancer: Molecular biology and new drugs. Critical Reviews in Oncology/Hematology. 2017;113:256–267.

[36] Chudasama P, Renner M, Straub M, Mughal SS, Hutter B, Kosaloglu Z, et al. Targeting Fibroblast Growth Factor Receptor 1 for Treatment of Soft-Tissue Sarcoma. Clinical Cancer Research. 2017;23(4):962–973.

[37] U.S. National Library of Medicine;. Available from: `https://clinicaltrials.gov/ct2/show/NCT04595747`.

[38] Koehler K, Liebner D, Chen JL. TP53 mutational status is predictive of pazopanib response in advanced sarcomas. Annals of Oncology. 2016;27(3):539–543.

[39] Zhang WW, Li L, Li D, Liu J, Li X, Li W, et al. The First Approved Gene Therapy Product for Cancer Ad-p53 (Gendicine): 12 Years in the Clinic. Human Gene Therapy. 2018;29(2):160–179.

[40] Zhao Y, Yang W, Huang Y, Cui R, Li X, Li B. Evolving Roles for Targeting CTLA-4 in Cancer Immunotherapy. Cellular Physiology and Biochemistry. 2018;47(2):721–734.

[41] Maki RG, Jungbluth AA, Gnjatic S, Schwartz GK, D'Adamo DR, Keohan ML, et al. A Pilot Study of Anti-CTLA4 Antibody Ipilimumab in Patients with Synovial Sarcoma. Sarcoma. 2013;2013:168145.

[42] Wu Dm, Liu T, Deng Sh, Han R, Xu Y. SLC39A4 expression is associated with enhanced cell migration, cisplatin resistance, and poor survival in non-small cell lung cancer. Scientific Reports. 2017;7(1):7211.

[43] Cui XB, Shen Yy, Jin Tt, Li S, Li Tt, Zhang Sm, et al. SLC39A6: a potential target for diagnosis and therapy of esophageal carcinoma. Journal of Translational Medicine. 2015;13(1):321.

[44] Liao QD, Wang CG, Zhu YD, Chen WH, Shao SL, Jiang FN, et al. Decreased expression of SLC39A14 is associated with tumor aggressiveness and biochemical recurrence of human prostate cancer. OncoTargets and Therapy. 2016;Volume 9:4197–4205.

[45] Mazhar S, Taylor SE, Sangodkar J, Narla G. Targeting PP2A in cancer: Combination therapies. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 2019;1866(1):51–63.

[46] Haesen D, Asbagh LA, Derua R, Hubert A, Schrauwen S, Hoorne Y, et al. Recurrent PPP2R1A Mutations in Uterine Cancer Act through a Dominant-Negative Mechanism to Promote Malignant Cell Growth. Cancer Research. 2016;76(19):5719–5731.

[47] Ye L, Guo L, He Z, Wang X, Lin C, Zhang X, et al. Upregulation of E2F8 promotes cell proliferation and tumorigenicity in breast cancer by modulating G1/S phase transition. Oncotarget. 2016;7(17):23757–23771.

[48] Deng Q, Wang Q, Zong WY, Zheng DL, Wen YX, Wang KS, et al. E2F8 Contributes to Human Hepatocellular Carcinoma via Regulating Cell Proliferation. Cancer Research. 2010;70(2):782–791.

[49] Park SA, Platt J, Lee JW, López-Giráldez F, Herbst RS, Koo JS. E2F8 as a Novel Therapeutic Target for Lung Cancer. JNCI: Journal of the National Cancer Institute. 2015;107(9).

[50] Yeo SY, Ha SY, Yu EJ, Lee KW, Kim JH, Kim SH. ZNF282 (Zinc finger protein 282), a novel E2F1 co-activator, promotes esophageal squamous cell carcinoma. Oncotarget. 2014;5(23):12260–12272.

[51] Skubitz KM, Pambuccian S, Manivel JC, Skubitz AP. Identification of heterogeneity among soft tissue sarcomas by gene expression profiles from different tumors. Journal of Translational Medicine. 2008;6(1):23.

[52] Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nature Methods. 2016;13(3):241–244.

[53] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nature Reviews Genetics. 2015;16(6):321–332.

[54] Nevins JR. The Rb/E2F pathway and cancer. Human Molecular Genetics. 2001;10(7):699–703.

[55] Francis AM, Alexander A, Liu Y, Vijayaraghavan S, Low KH, Yang D, et al. CDK4/6 Inhibitors Sensitize Rb-positive Sarcoma Cells to Wee1 Kinase Inhibition through Reversible Cell-Cycle Arrest. Molecular Cancer Therapeutics. 2017;16(9):1751–1764.

[56] Hemming ML, Bhola P, Loycano MA, Anderson JA, Taddei ML, Doyle LA, et al. Preclinical modeling of leiomyosarcoma identifies susceptibility to transcriptional CDK inhibitors through antagonism of E2F-driven oncogenic gene expression. Clinical cancer research : an official journal of the American Association for Cancer Research. 2022.

[57] Napolitano A, Ostler AE, Jones RL, Huang PH. Fibroblast Growth Factor Receptor (FGFR) Signaling in GIST and Soft Tissue Sarcomas. Cells. 2021;10(6):1533.

[58] Graaf WTAvd, 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. The Lancet. 2012;379(9829):1879–1886.

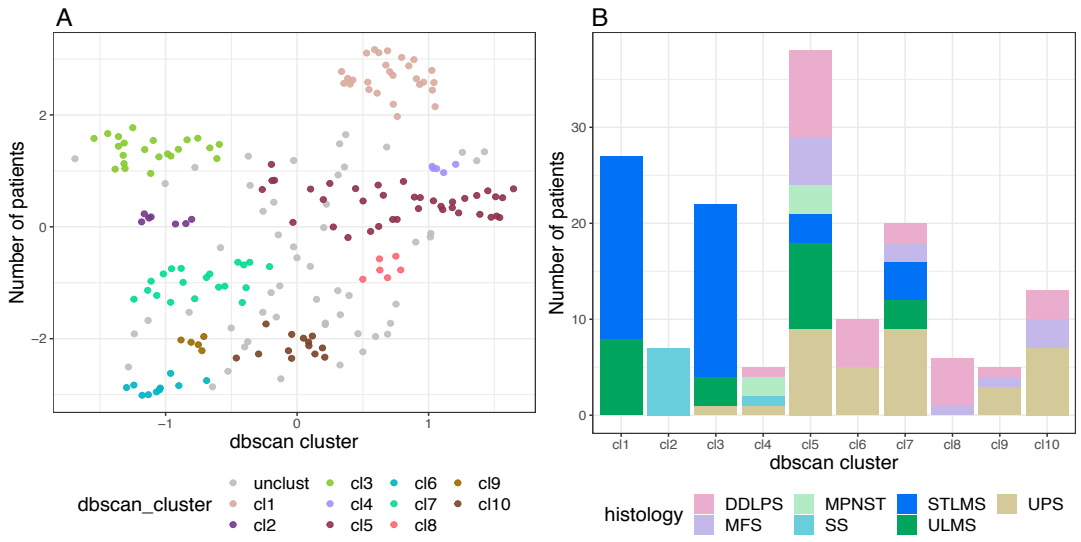
[59] Benson C, Ray-Coquard I, Sleijfer S, Litière S, Blay JY, Cesne AL, et al. Outcome of uterine sarcoma patients treated with pazopanib: A retrospective analysis based on two European Organisation for Research and Treatment of Cancer (EORTC) Soft Tissue and Bone Sarcoma Group (STBSG) clinical trials 62043 and 62072. Gynecologic Oncology. 2016;142(1):89–94.

[60] Bushweller JH. Targeting transcription factors in cancer — from undruggable to reality. Nature Reviews Cancer. 2019;19(11):611–624.

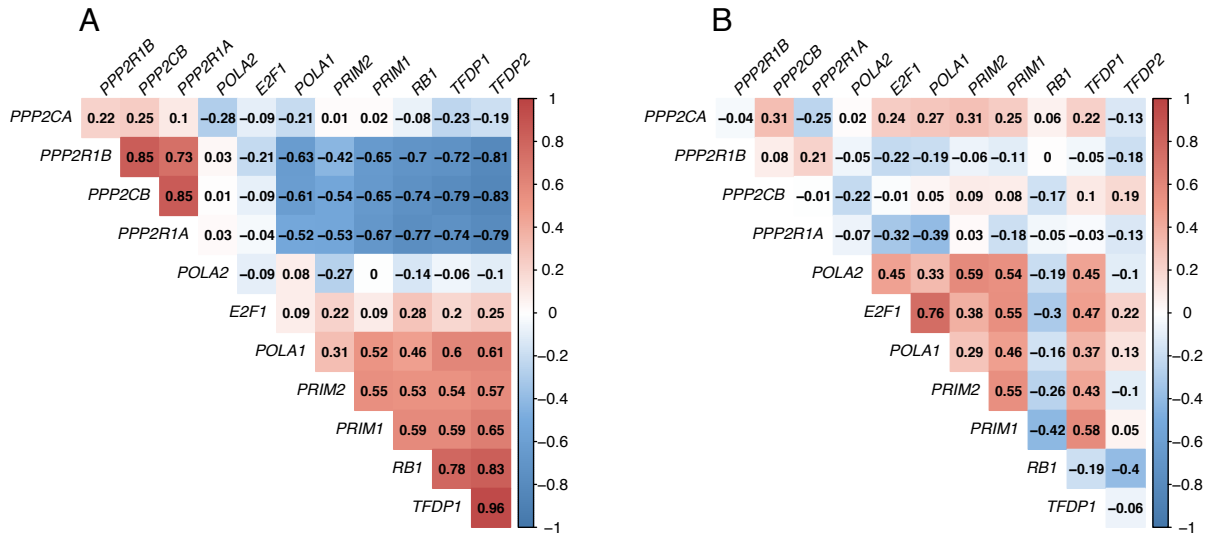
SUPPLEMENTARY FIGURES AND TABLES

ptw genes	Reactome%	TCGA%	DKFZ%
<50	75.8	81.9	78
$\geqslant 50 \& \leqslant 100$	16	12.5	12
>100 & <150	5.9	4.2	5.8
$\geqslant 150$	2.3	1.4	4.4

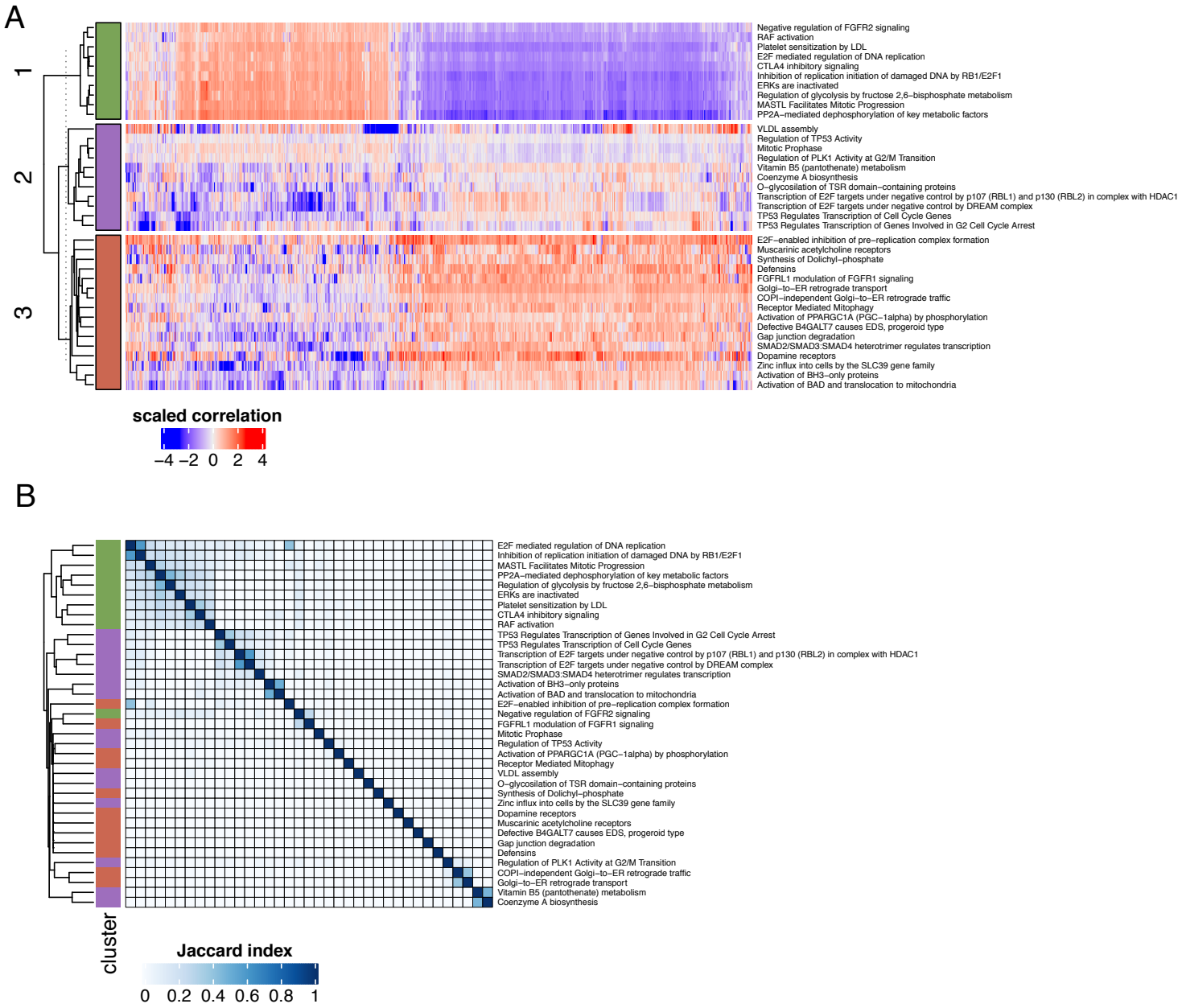
Supplemental Table S1. Proportions of pathways of different sizes among Reactome pathways and pathways identified with PORCUPINE in the TCGA-LMS and DKFZ-LMS datasets.



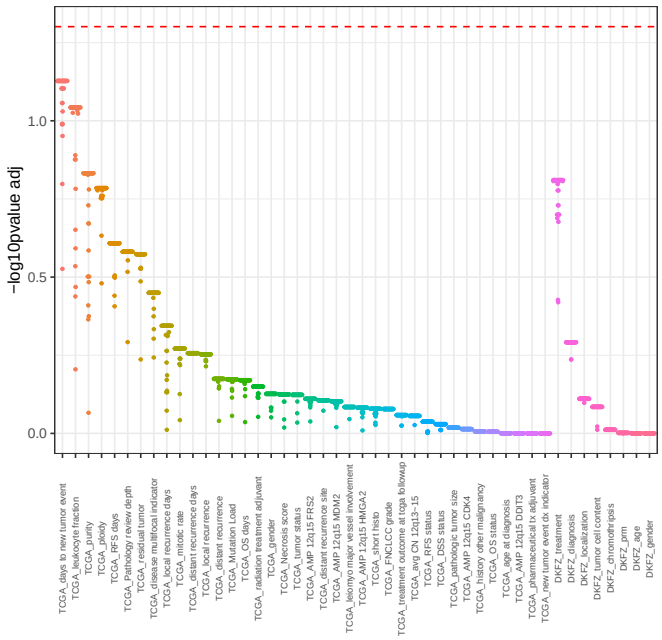
Supplementary figure S1. A. HDBSCAN clustering of 206 soft-tissue sarcomas on the first two UMAP dimensions obtained from applying UMAP on gene targeting scores. B. Distribution of STS histological subtypes across HDBSCAN clusters.



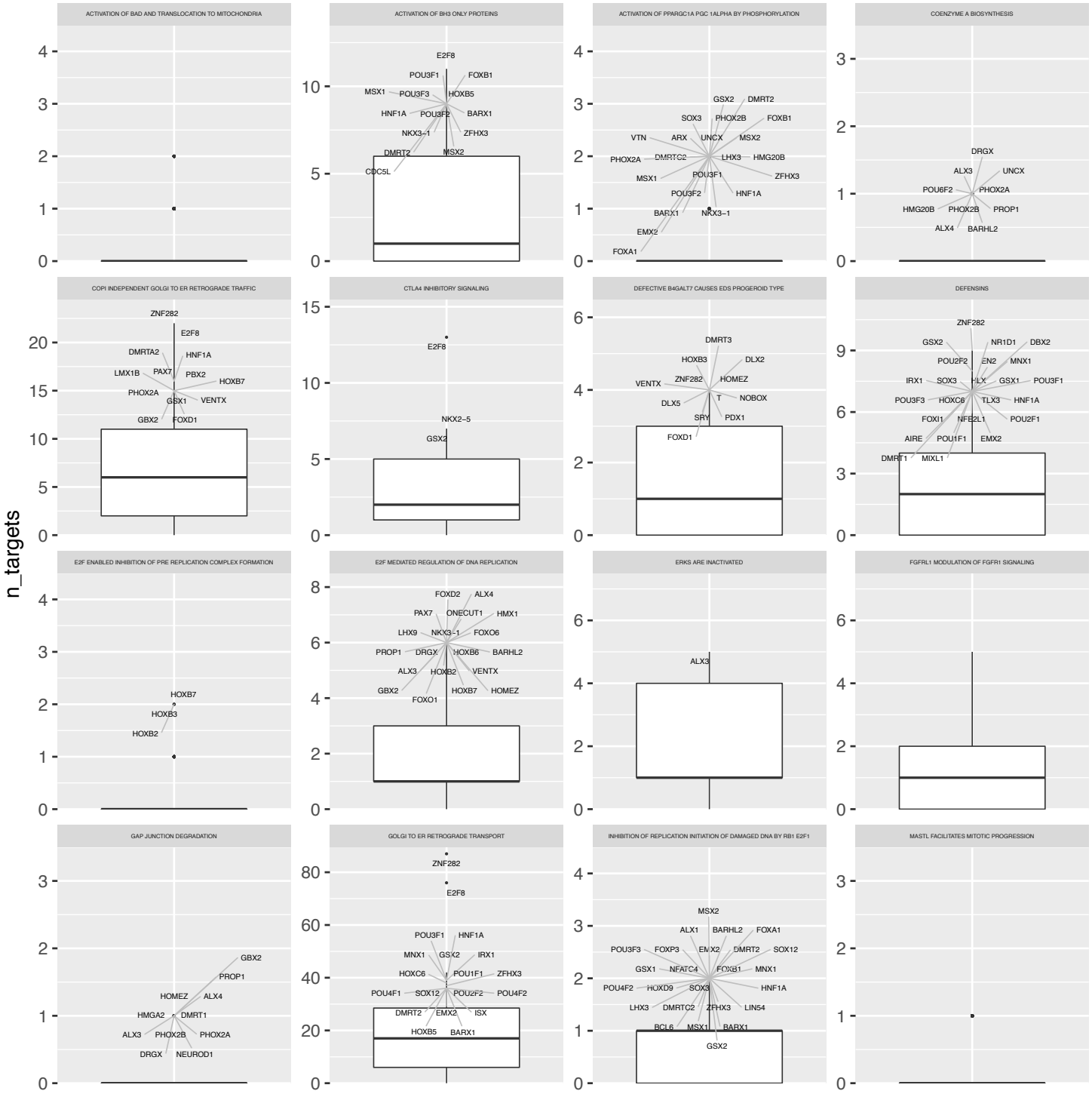
Supplementary figure S2. A. Pearson correlations between gene targeting scores. B. Correlations between gene expression levels in the pathway "Inhibition of replication initiation of damaged DNA by RB1/E2F1."

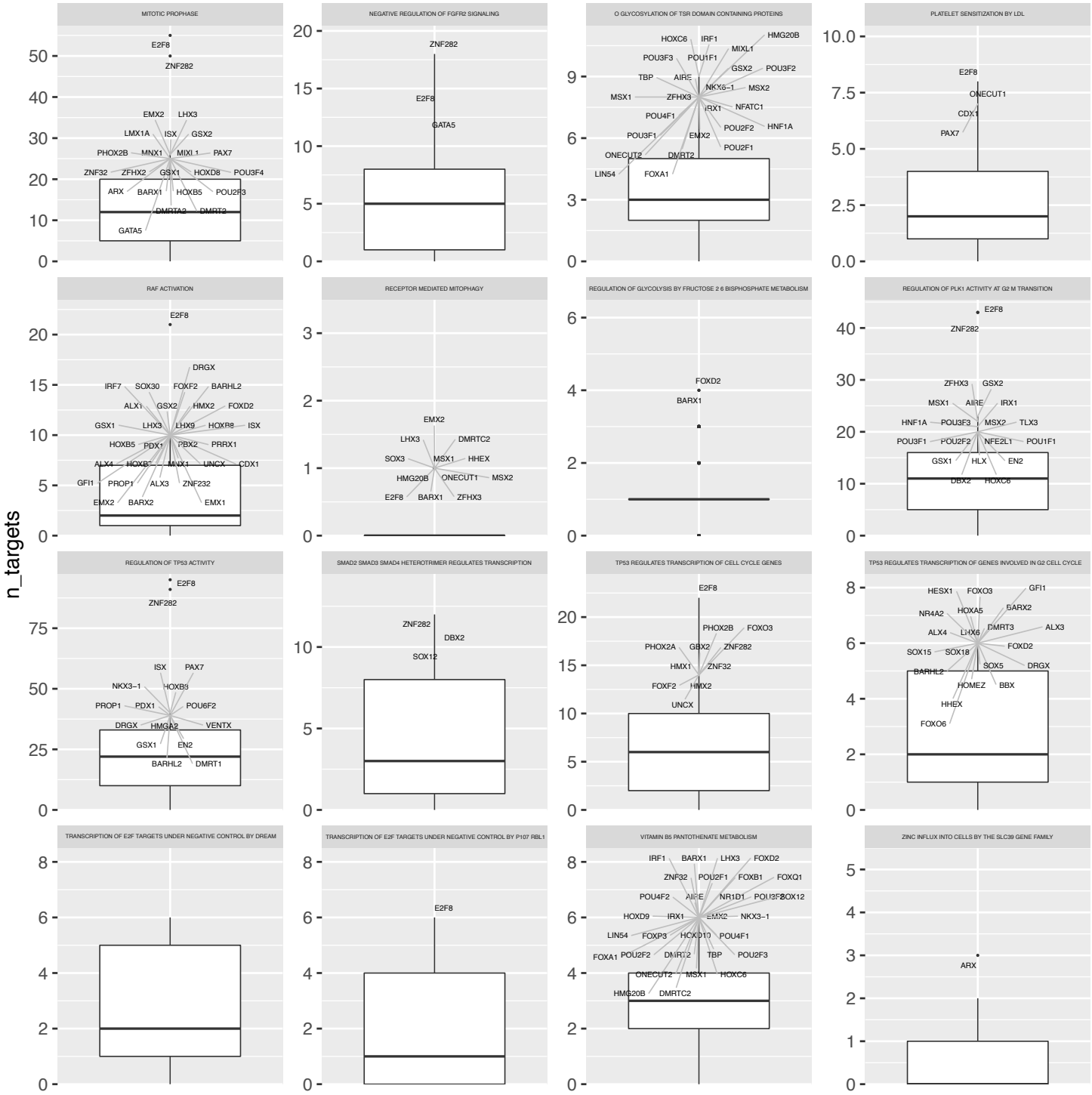


Supplementary figure S3. Clustering of 37 pathways based on A. pair-wise correlations between individual networks from the TCGA-LMS dataset; and B. the proportion of shared genes between the pathways.



Supplementary figure S4. Association of the clinical features of patients and pathway-based patient heterogeneity scores on PC1 in each of the 37 pathways.





Supplementary figure S5. Boxplots showing the number of targets for transcription factors with most highly weighted values to PC1 in each pathway. Transcription factors with number of targets greater than the 95th percentile are labelled.