Published in final edited form as:

Mol Cancer Res. 2016 April ; 14(4): 332–343. doi:10.1158/1541-7786.MCR-15-0368.

Integrative Genomic Analyses Yields Cell Cycle Regulatory Programs with Prognostic Value

Chao Cheng^{1,2,3,§}, Shaoke Lou¹, Erik H. Andrews¹, Matthew H. Ung¹, and Frederick S. Varn¹

¹Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire, 03755 USA

²Institute for Quantitative Biomedical Sciences, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, 03766 USA

³Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, 03766 USA

Abstract

Liposarcoma is the second most common form of sarcoma, which has been categorized into four molecular subtypes, which are associated with differential prognosis of patients. However, the transcriptional regulatory programs associated with distinct histological and molecular subtypes of liposarcoma have not been investigated. This study uses integrative analyses to systematically define the transcriptional regulatory programs associated with liposarcoma. Likewise, computational methods are used to identify regulatory programs associated with different liposarcoma subtypes as well as programs that are predictive of prognosis. Further analysis of curated gene sets was used to identify prognostic gene signatures. The integration of data from a variety of sources including gene expression profiles, transcription factor (TF) binding data from ChIP-seq experiments, curated gene sets, and clinical information of patients indicated discrete regulatory programs (e.g., controlled by E2F1 and E2F4) with significantly different regulatory activity in one or multiple subtypes of liposarcoma with respect to normal adipose tissue. These programs were also shown to be prognostic, wherein liposarcoma patients with higher E2F4 or E2F1 activity associated with unfavorable prognosis. A total of 259 gene sets were significantly associated with patient survival in liposarcoma, among which >50% are involved in cell cycle and proliferation.

§ Corresponding author (chao.cheng@dartmouth.edu).

The authors declare no conflict of interest

Implications

These integrative analyses provide a general framework that can be applied to investigate the mechanism and predict prognosis of different cancer types.

Competing interests

All authors declare to have no competing interests.

Authors' contributions

CC conceived the study and wrote the paper. SL participated in data analyses and manuscript writing. EHA, MHU and FSV were involved in microarray data processing and manuscript writing. All authors had final approval of the submitted and published versions.

Keywords

Liposarcoma; Integrative analysis; Transcription factor; Survival analysis; Regulatory program

Introduction

Soft tissue sarcoma is a group of malignant tumors that develop from the body's connective tissue such as muscles, tendons and fat. Based on statistics from 2009, there are ~10,000 new soft tissue sarcoma cases diagnosed each year in the United States (1), with the incidence in men 50% higher than in women and a mean survival time around 2.5 years, resulting in ~4,000 deaths each year (2). Soft tissue sarcomas are rare in adults with less than 1% of new cancer diagnoses, but relatively more common among children and account for over 20% of all pediatric malignant solid tumors (3). Based on histology, soft tissue sarcomas can be classified into more than 50 subtypes (4, 5), in which liposarcoma, a malignancy of adipocytes, makes up more than 20% of cases. Liposarcoma can be further categorized into five histological subtypes, including well-differentiated liposarcoma, dedifferentiated liposarcoma, myxoid/round cell liposarcoma, myxoid liposarcoma, and pleomorphic liposarcoma (4).

Genome-wide analyses have previously been performed to identify the genetic alterations in liposarcoma. Indeed, genomic alterations, including gene mutations, karyotype abnormalities, and copy number variations, have been identified in all liposarcoma subtypes (6-8). For example, it has been shown that 18% of myxoid/round-cell liposarcomas have *PIK3CA* mutations and 17% of pleomorphic liposarcomas have *TP53* mutations (9). Furthermore, gene expression profiling has also been used extensively to investigate the transcriptional dysregulation in liposarcoma (9-11). These studies have revealed genes that are differentially expressed between liposarcoma and normal samples, representing the down-stream output of genetic mutations in liposarcoma genomes. However, the transcriptional regulatory programs that connect genetic mutations with gene expression changes are largely unknown.

Traditionally, the histological subtype of a given liposarcoma is considered the major determinant of local and overall recurrence, survival outcome, and treatment approaches (4, 12, 13). For example, it has been shown that pleomorphic liposarcoma patients have a 3-fold higher risk of distant metastasis compared with dedifferentiated liposarcoma patients (14). Myxoid liposarcoma and myxoid/round cell liposarcoma are characterized as low and high grade, respectively, according to histological features and cell morphology (15, 16). For dedifferentiated and well-differentiated liposarcoma, subtype-specific therapeutic targets have been identified and different treatment approaches were suggested due to their difference in progression and regulatory mechanisms (4, 17). However, histological classification and stratification do not provide insight into the mechanisms of the varying liposarcoma malignancies. Further, improvement in prognostic prediction is an ongoing goal that a better understanding of the molecular pathology of liposarcoma promises to provide. A recent study showed that MYC protein expression was correlated with survival of liposarcoma patients (18), suggesting that the understanding of transcriptional regulatory

programs associated with patient survival will add substantial value to traditional histological prognosis analysis.

We have previously developed computational methods, REACTIN (Regulatory Activity Inference) and BASE (Binding Association with Sorted Expression), to investigate the regulatory programs associated with cancer (19-21). Both of the methods integrate gene expression data with ChIP-seq transcription factor (TF) binding data with the rationale that the regulatory activity of TFs can be reflected by the expression of their target genes. REACTIN compares the expression profiles between two groups of samples (e.g., tumor vs. normal) to identify TFs with significantly different regulatory activities (19). BASE can be used to infer the regulatory activity of TFs in each sample of a cancer gene expression dataset, and TFs associated with patient survival can subsequently be identified by survival analysis (20). In the past few years, a large number of ChIP-seq data have been generated for human TFs by the large-scale project ENCODE (the Encyclopedia of DNA Elements) (22) and case-by-case studies (23). The increasing availability of these datasets facilitates the application of the REACTIN and BASE algorithms to identify regulatory programs associated with tumorigenesis and prognosis of cancer in a systematic manner.

In this study, we performed systematic analyses in liposarcoma that integrate gene expression data, ENCODE ChIP-seq data, curated gene set data from MSigDB (24) and clinical information. We identified regulatory programs associated with liposarcoma and each subtype by using the REACTIN algorithm. Out of the 424 regulatory programs defined by ENCODE ChIP-seq experiments, the *E2F1*, *E2F4* and *NR3C1* programs are found to have significant activity difference between liposarcoma and normal fat tissue. Using BASE and survival analysis, we then systematically characterized regulatory programs that were predictive of liposarcoma patient survival. Interestingly, the *E2F1* and *E2F4* programs were among the most significant prognostic programs. Based on these results, we further refined the *E2F1* and *E2F4* regulatory programs by combining their target genes from multiple cell-lines. We find that the refined *E2F1* and *E2F4* signatures predict patient survival outcome with high accuracy in all samples as well as in dedifferentiated and well-differentiated subtypes. In addition, we extended our analysis to all 4,257 MSigDB gene sets and identified 261 gene sets that correlated with patient survival. These survival-associated gene sets group into five functional clusters, one of which is associated with “cell cycle and proliferation” and thus confirms the prognostic value of the *E2F1* and *E2F4* regulatory programs. The integrative analyses provide a framework that can be readily applied to other cancer types and diseases.

Materials and Methods

Gene expression processing

Whole-transcriptome expression datasets for soft tissue sarcoma using the Affymetrix Human Genome U133A Array platform (GSE21122 and GSE30929) were downloaded from the GEO database (9, 47). The former dataset consisted of 158 samples, including 9 normal fat tissues and sarcoma cancer samples with 5 major subtypes: dedifferentiated liposarcoma, pleomorphic liposarcoma, myxoid/round-cell liposarcoma, leiomyosarcoma, and myxofibrosarcoma (9). In our analysis, we only utilize the normal fat samples as control and

the three subtypes of liposarcoma tumor samples. The latter dataset consists of 140 samples across 5 subtypes of liposarcoma: dedifferentiated liposarcoma, myxoid liposarcoma, myxoid/round cell liposarcoma, pleomorphic liposarcoma, and well-differentiated liposarcoma. Distant recurrence-free survival (DRFS) information is also provided. For each dataset, we extracted the normalized expression data from series matrix files and transformed them log₁₀-based values. Probe set expression was collapsed into gene expression by choosing the probe with maximum average values in all samples to represent the gene expression for genes with multiple probe sets. The processed data contain expression levels of a total of 12,815 genes.

Processing of ChIP-seq data for human TFs

We downloaded and processed 424 TF binding profiles for a total of 119 human TFs under different cell lines from the ENCODE project (22). The data provide the binding signal of TFs in specific cell lines at each genome position, in the format of “bedGraph” files. Using a method called TIP (Target Identification from Profiles), we quantified the binding affinities and p-values of a TF with all genes. TIP takes into account TF binding signals within [-10kb, 10kb] surrounding transcription start site (TSS) of genes. In general, the contribution of binding events decays with the increase of their distances to the TSS of a gene. These binding affinities of TFs with all genes were further transformed to obtain 424 regulatory programs (e.g., the 6 programs shown in Figure 4), each representing the regulatory potential of all genes by a TF in the corresponding cell lines. These programs were utilized as the input for REACTIN and BASE analysis. More detailed description on defining regulatory programs based on ChIP-seq data can be found in Zhu et al (19).

Identification of regulatory programs with significant activity difference between liposarcoma and normal samples

The REACTIN (Regulatory Activity Inference) algorithm (19) was used to identify the regulatory programs that show significant activity difference between liposarcoma tumor samples and normal tissues. REACTIN takes three matrices as the inputs: a case gene expression matrix, containing expression levels of all genes in a number of case samples; a control gene expression matrix, containing expression levels of all genes in a number of control samples; and a TF binding matrix, containing the binding scores of all genes by a number of TFs under specific conditions/cell lines determined based on ChIP-seq/ChIP-chip experiments. The algorithm infers the regulatory activity difference of each TF/cell-line in case samples relative to control samples by examining the expression difference of each TF/cell-line's target genes between the cases and controls. For each TF/cell-line, REACTIN calculates a regulatory activity score (RAS), with positive scores indicating higher expression of target genes in case versus control and negative scores indicating lower expression. Thus, positive RAS suggests higher regulatory activity of a TF in the event that the TF mainly functions as a transcriptional activator, or lower regulatory activity if the TF is mainly a transcriptional repressor. Other than RAS, the algorithm also calculates the P-value and Q-value (FDR, false discovery rate) for each TF/cell-line based on a permutation procedure. In order to precisely model TF-gene interactions (i.e., target genes of TFs), REACTIN applies continuous quantitative values to represent TF binding affinities with all genes. More detailed description about the REACTIN algorithm can be found in (19).

The GSE21122 data contains expression profiles for normal fat tissue, dedifferentiated, myxoid/round-cell and pleomorphic liposarcoma subtypes. We thus performed four pairs of comparisons using the REACTIN algorithm: all liposarcoma samples (a combination of all three subtypes) versus normal fat samples, and samples from each of the three subtypes versus normal fat samples. A FDR=0.05 cutoff was used to select significant regulatory programs. The GSE30929 dataset was not used for REACTIN analysis, since it only contained tumor samples without normal controls.

Identification of regulatory programs associated with patient survival

The BASE (Binding Association with Sorted Expression) algorithm (20) is used to calculate the regulatory activity of a TF in each individual sample (denoted as iRAS). BASE takes two matrices as the inputs: an gene expression matrix, containing expression levels of all genes in a number of tumor samples; and a TF binding matrix, containing the binding scores of all genes by a number of TFs under specific conditions/cell lines (the same as the one used in REACTIN). To be flexible, the TF-binding matrix may also be filled with binary values, with 1 indicating target genes of a TF, thereby allowing pre-defined target gene sets for TFs to be used as the input to the BASE algorithm. In addition, more generally defined gene sets (e.g., genes in a pathway) can be used by BASE to calculate the “activity” of these gene sets in each sample. We applied the BASE algorithm to calculate the iRASs of the 424 regulatory programs defined based on ENCODE ChIP-seq data in each of the 140 liposarcoma tumor samples of the GSE30929 data.

For each set of iRASs across the 140 tumor samples for each regulatory program or gene set, a Cox PH regression model (48) was then used to examine survival association with patient survival. The resulting p-values for regulatory programs were adjusted into q-values by using the “Benjamini & Hochberg” multiple-testing correction method (49). The R package “survival” was used to implement survival analysis.

Defining *E2F1* and *E2F4* signatures by combining their target genes in multiple cell lines

ChIP-seq data of *E2F1* in HeLa S3 (two replicates from ENCODE data) and MCF7 cell lines, and *E2F4* in HeLa 3, GM06900 and K562b cell lines, were used to define the core target genes sets for *E2F1* and *E2F4* (*E2F1* and *E2F4* signatures), respectively. Target genes in each cell line were identified using the TIP method (50), with a FDR <0.1 being selected as the threshold. *E2F1* and *E2F4* signatures were defined by selecting genes that are targeted by them in all cell lines.

Identifying liposarcoma-specific target genes of *E2F1* and *E2F4* using ARACNE analysis

The ARACNE algorithm was applied to the GSE30929 to identify putative *E2F1* and *E2F4* target genes. This algorithm infers transcriptional regulatory interactions based on mutual information between expression profiles of genes. The expression data was pre-processed to remove genes with low expression levels and then choose 5,000 genes with the highest coefficient of variance. $P < 1e-7$ was used as the threshold to select significant target genes, resulting in 202 and 164 target genes for *E2F1* and *E2F4*, respectively.

Identification of gene sets associated with patient survival

The C2 curated dataset was downloaded from the MSigDB database (51), which contains 4,257 gene sets derived from pathway databases, literature, and expert knowledge. The iRASs of gene sets were calculated using the BASE algorithm and Cox PH regression performed to test the survival association for each gene set. The resulting significant gene sets (defined as meeting a FDR < 0.001 threshold) were combined with the *E2F1* and *E2F4* signatures to obtain a total of 261 survival-associated signatures.

Functional and network clustering of survival-associated signatures

For each pair of gene sets, we defined an overlap-ratio as the number of shared genes over the number of total genes in each gene set ($p1$ and $p2$). We subsequently connected all gene set pairs into a network by treating each gene set as a node and connecting two nodes if their overlap-ratio $\left(\frac{\sqrt{p1} \times \sqrt{p2}}{p1 + p2} \right) > 0.1$.

ClusterOne was used to cluster the network based on the cohesiveness of nodes (52). A plugin for Cytoscape (53) was downloaded from the Cytoscape app market and used to search clusters using the parameter settings of 1) p-value < 0.001 and 2) minimum number of gene set in a cluster = 10 (53). The common genes for each cluster were defined as those shared by at least 20% of the gene sets in the cluster, and then uploaded to the DAVID website for functional annotation (54). Based on the annotations to these common genes, the biological function of each network cluster was assigned.

Results

An overview of our analysis

A schematic flowchart of our analysis is presented in Figure 1. Briefly, we used REACTIN to integrate data from 424 ChIP-Seq experiments, which assayed the DNA binding profiles of 119 TFs across multiple cell lines, with gene expression data from a liposarcoma case-control dataset of liposarcomas and normal adipose soft tissue samples (9), to provide a list of TF regulatory programs that significantly differed between the liposarcoma and normal adipose tissue samples. Concurrently, using BASE we combined this same ChIP-seq data with another liposarcoma gene expression dataset (with patient survival information) (10), to calculate individual Regulatory Activity Scores (iRASs) for each TF in each sample. Using Cox PH models with clinical information containing distant recurrence-free survival (DRFS) information, we calculated which TFs' regulatory programs correlate with DRFS times. With these analyses, we characterized the TF regulatory programs that associate with liposarcoma oncogenesis and patient survival time, respectively, and identified two regulatory programs, *E2F1* and *E2F4*, that are both associated with liposarcoma and patient survival. To extend our examination of these two regulatory programs, we refined the *E2F1* and *E2F4* target gene signatures by combining the ChIP-Seq experimental data across cell lines, keeping only the target genes that are present across all cell lines, and used the refined signatures to predict the survival outcome of liposarcoma within and among its subtypes. Next, we applied BASE to the MSigDB gene sets and fitted each result to a Cox PH DRFS model, identifying 259 gene sets that correspond to liposarcoma patient survival. Finally,

with the *E2F1*, *E2F4* and 259 MSigDB survival-associated gene sets, we constructed a gene-set overlapping network to identify the overall structure of survival-associated pathways of liposarcoma. We classified the network into five functional clusters using clustering and Gene ontology annotation.

Regulatory programs associated with liposarcoma

First we identified the regulatory programs associated with liposarcoma by comparing all liposarcoma tumor samples with normal fat samples in GSE21122. Specifically, we examined the difference in regulatory activities of 424 programs defined based on ENCODE ChIP-seq data, which represent the DNA-binding profiles of 119 TFs across multiple cell lines and conditions. Suppl. Table S1 shows the four significant liposarcoma-associated regulatory programs identified by REACTIN at the significance level of 0.05 (FDR<0.05), with the regulatory activity score (RAS) indicating the activity difference between the liposarcoma tumor and the normal fat samples.

As shown in Suppl. Table S1, we identified four significant regulatory programs associated with liposarcoma. Among them, the HeLa S3 *E2F4*, K562 *E2F4*, and MCF7 *E2F1* programs have positive RAS values, indicating that the TFs regulating these programs exhibit higher activating capability or lower repressive capability in liposarcoma tumor samples than in normal fat tissue. In contrast, the A549 *NR3C1* program is associated with a negative RAS, suggesting lower activity of *NR3C1* (assuming it is a transcriptional activator) in liposarcoma tumor samples. The dataset contains expression profiles for three subtypes of liposarcoma: dedifferentiated, myxoid/RC (round cell), and pleomorphic Liposarcomas. We thus applied the REACTIN algorithm to identify significant regulatory programs associated with each subtype (Suppl. Table S2 for programs with FDR<0.01). The MCF7 *E2F1* (Figure 2A), HeLa S3 *E2F4*, and K562 *E2F4* (Figure 2B) programs are seen to be significant in all three subtypes. However, the A549 *NR3C1* program is significant in the dedifferentiated and the myxoid/RC subtypes, but not in the pleomorphic subtype (Figure 2C). In addition, the myxoid/RC and the pleomorphic subtypes are associated with a different set of other regulatory programs. For example, the MCF10 *STAT3* and the Huvec *CFOS* programs are significant in the myxoid/RC subtype (Suppl. Table S2). Taken together, our results suggest that different liposarcoma subtypes are associated with different sets of regulatory programs but they also share common programs (e.g., the *E2F4* programs). The *E2F4* programs are highly significant in all liposarcoma samples as well as in the three subtypes, suggesting that *E2F4* is a key factor and plays a crucial role in the tumorigenesis and progression of liposarcoma.

The REACTIN algorithm identifies liposarcoma associated regulatory programs based on the inferred regulatory activity difference between tumor and normal samples. Alteration of TF activity might result from different mechanisms, including changes in mRNA expression, protein expression, or post-translational modification (e.g., phosphorylation). Thus, we investigated whether the identified regulatory programs associated with liposarcoma can be revealed at the mRNA expression level of the corresponding TFs. We compared the expression levels of *E2F1*, *E2F4*, and *NR3C1* between liposarcoma tumor samples and normal fat samples (Figure 2). As can be seen, *E2F1* regulatory activity is reflected by its

mRNA expression (Figure 2A vs. 2D), with significant expression changes of *E2F1* mRNA overall and within subtypes of liposarcoma compared to normal samples. For *E2F4*, while we observed significantly enhanced activity in all subsets of liposarcoma samples (Figure 2B), *E2F4* mRNA expression levels were found to be down-regulated in dedifferentiated and pleomorphic subtypes, but without significant expression change in the myxoid/RC subtype (Figure 2E). This suggests that *E2F4* protein may function mainly as a transcription repressor and that in different liposarcoma subtypes the deregulation of *E2F4* activity involves different mechanisms than altered mRNA expressivity. The inconsistency between mRNA expression and regulatory activity is more apparent for the *NR3C1* regulatory program (Figure 2C vs. 2F). As shown, we have detected significant activity changes of *NR3C1* in all liposarcoma subtypes, but significant down-regulation of *NR3C1* mRNA expression is only observed in the myxoid/RC subtype. Taken together, our results indicate that by inferring TF activity changes REACTIN is more sensitive in identifying cancer associated regulatory programs than looking at mRNA expression levels of TFs alone. When used in conjunction, REACTIN-inferred regulatory activity and mRNA expression of TFs provide increased biological insight into transcriptional regulation in cancer.

Regulatory programs predictive to patient survival in liposarcoma

We next investigated whether regulatory programs defined using ChIP-seq data for TFs are informative for predicting clinical outcomes of liposarcoma patients. Given the 424 ENCODE TF regulatory programs, we applied the BASE algorithm to calculate the regulatory activity of each program in each liposarcoma case of the GSE30929 dataset, which includes clinical survival and mRNA expression data for 140 liposarcoma cases. To discriminate from the regulatory activity scores (RAS) calculated by REACTIN, which summarize TF activity difference between two groups of samples, we called these scores (calculated by BASE) for individual samples individual Regulatory Activity Scores (iRASs). We calculated iRASs for each of the 424 programs in the 140 liposarcoma samples.

We performed hierarchical clustering analysis to these iRAS scores to examine whether iRAS can discriminate different liposarcoma subtypes (Figure 3). As shown, liposarcoma samples of the well-differentiated subtype, which is known to have a favorable prognosis, tend to have lower iRASs (region in green), indicating that across the assayed TFs, they tend to have less TF regulatory activity. Likewise, patient tumors of the pleomorphic subtype, which has unfavorable prognosis, tend to have higher iRASs compared to all other subtypes. This suggests that the iRAS reflects the underlying molecular characteristics of tumors, and that poorer-prognosis subtypes have globally more TF transcriptional regulatory activity, at least among the TF regulatory programs studied. However, this analysis is unable to clearly stratify the myxoid and myxoid/round cell subtype tumors (Figure 3). This is in line with prior findings showing that myxoid and myxoid/round-cell tumors are difficult to classify (15). Additionally, dedifferentiated liposarcoma tumors show no clear clustering patterns, indicating a more complicated regulatory mechanism responsible for its poor prognosis (17).

To identify the regulatory programs that are prognostic in liposarcoma, we applied univariate Cox PH regression models with iRAS as the single variable to identify the programs correlated with distant recurrence-free survival (DRFS) of patients. Table 1 lists the top 15

most significant regulatory programs (FDR<0.001). Interestingly, the *E2F1* programs (defined in MCF7 and HeLa S3) and the *E2F4* programs (defined in HeLa S3 and K562b), which show differential activity between liposarcoma tumor and normal fat samples, are also identified to be predictive to the survival of liposarcoma patients. To demonstrate the capability of these programs for prognostic prediction, we stratified patient samples into two groups with positive and negative iRAS values, respectively (Figure 4A-D). As shown, stratification based on iRASs of these programs separate patients into better and worse prognostic groups with significantly different survival times (FDR<0.001). Beyond the *E2F1* and the *E2F4* programs, our analysis also identified other prognostic regulatory programs such as the *IRF3* program (Figure 4E) and the *FOS* program (Figure 4F) defined in HeLa S3 cell line.

Define *E2F1* and *E2F4* signatures for prognostic prediction in liposarcoma

As shown in Suppl. Table S1, *E2F1* and *E2F4* regulatory programs are found to be significantly associated with both liposarcoma development and liposarcoma survival outcome. Thus, we pursued further investigation of these programs in prognostic prediction of liposarcoma. As each TF's regulatory program was inferred from the activity of target genes derived from multiple cell culture ChIP-seq experiments, none of which were conducted within adipocytes, we decided to refine the target gene lists for each TF by taking the intersection of the target genes across all experiments: the genes that were identified as targets in all cell culture experiments. As these genes were identified as targets across multiple cell lines, we reasoned they represented core target gene sets that would remain targets within adipocytes as well. After this refinement, the *E2F1* signature contains 116 genes and the *E2F4* signature contains 199 genes, among which only 21 are shared.

We then calculated the iRASs for *E2F1* and *E2F4* (denoted as *E2F1* scores and *E2F4* scores, respectively) in the 140 liposarcoma tumor samples by applying the BASE algorithm to the two signatures. Cox PH survival regression shows a significant hazardous effect of the *E2F1* and *E2F4* scores. When patients are stratified based on *E2F1* or *E2F4* scores (setting 0 as the cut-off between strata), the resulting positive and negative groups show significantly different DRFS prognosis: the positive groups of both are associated with worse prognosis (Suppl. Figure S1A and S1C). As expected, the proportion of good prognosis samples varies substantially between different liposarcoma subtypes. The majority of samples in well-differentiated subtype are classified as good prognosis (*E2F1*<0), while most of the samples in myxoid/round cell and pleomorphic subtypes are classified as poor prognosis (*E2F1*>0) (Suppl. Figure S1B). Similar distribution is observed when *E2F4* score is used to stratify samples (Suppl. Figure S1D), but it seems that the *E2F1* and *E2F4* signatures are different in discriminating dedifferentiated liposarcoma samples.

In addition, we have applied the ARACNE algorithm (25) to identify putative *E2F1* and *E2F4* target genes in liposarcoma samples. In GSE30929 data, we have identified 202 and 164 liposarcoma-specific target genes for *E2F1* and *E2F4*, respectively. We then applied the BASE algorithm to obtain *E2F1* scores and *E2F4* scores liposarcoma samples based on these target gene sets. When patients were stratified into high- and low-risk groups based on these scores, we observed significant survival difference between them (Suppl. Figure S2). This

result indicate that our analysis can be combined with condition-specific TF target genes prediction methods, such as ARACNE, to identify prognostic regulatory programs, which relaxes the requirement of our analysis for ChIP-seq data.

Application of *E2F1* and *E2F4* signatures to prognostic prediction in liposarcoma subtypes

Distinct liposarcoma subtypes are associated with different prognoses. Dedifferentiated and pleomorphic subtypes are associated with poor prognosis; well-differentiated and myxoid subtypes are associated with good prognosis, while the myxoid/round cell subtype shows intermediate prognosis (Suppl. Figure S3A, GSE30929 data). We compared the five liposarcoma subtypes in their *E2F1* scores (Suppl. Figure S3B) and *E2F4* scores (Suppl. Figure S3C) in GSE30929. *E2F1* and *E2F4* scores approximately reflect the prognosis of different subtypes. As shown, the well-differentiated subtype and the pleomorphic subtype exhibit the lowest and the highest *E2F1/E2F4* scores, respectively, consistent with their prognosis. In the dedifferentiated and myxoid subtypes, *E2F1* score shows bimodal distributions with a significantly proportion of samples associated with lower scores. This pattern is not observed for *E2F4* scores in the two subtypes, which suggests that the *E2F1* and *E2F4* programs, though both involved in cell cycle and proliferation, capture different information regarding transcriptional regulation in liposarcoma.

We next examined whether the *E2F1* and *E2F4* scores can be used to predict survival outcome in individual subtypes. Limited by sample size, we only choose the dedifferentiated (40 samples) and the well-differentiated (52 samples) subtypes for prognostic analysis. As shown, The *E2F1* program is able to classify good and poor prognosis samples in the dedifferentiated subtype ($P=0.03$, Figure 5A), but not in the well-differentiated subtype ($P>0.1$, Figure 5B). In contrast, the *E2F4* program can discriminate good from poor prognosis samples with moderate significance in dedifferentiated subtype ($P=0.08$, Figure 5C) and ($P=0.05$, Figure 5D). The difference between *E2F1* and *E2F4* might be explained by their functional difference in cell differentiation: *E2F1* suppresses cell differentiation while *E2F4* promotes cell differentiation (26).

Gene sets associated with liposarcoma prognosis

Up to now in this analysis *E2F1* and *E2F4* regulatory programs have been defined based on ChIP-seq experiments from the ENCODE project. Regulatory programs associated with them might also be characterized by other approaches (e.g., differentially expressed genes in *E2F1/E2F4* knockdown experiments) and other sources beyond ENCODE. The MSigDB database is one such source of alternative approaches, having collected 4,257 curated gene sets that are compiled based on previous publications and known biological pathways (24) and are highly informative for transcriptional regulatory programs. Therefore, we performed a systematic analysis of MSigDB data to identify gene sets that are associated with patient survival in liposarcoma. Specifically, we calculated the iRASs of all gene sets in all samples in the GSE30929 dataset using the BASE algorithm, and examined their correlation with survival using univariate Cox PH regression models. At a $FDR<0.001$ significance level, 259 significant gene sets are identified (Suppl. Table S3).

To further investigate these significant survival prognostic gene sets, we connected them into a network by linking sets based on their number of shared genes (Figure 6). As shown, the gene sets group into 5 clusters: “cell cycle and proliferation” (138 gene sets), “PPAR pathway and fatty-acid metabolism” (19 gene sets), “cAMP dependent and regulation related” (14 gene sets), “MAPK pathway and cell growth” (10 gene sets), and “oxidation reduction and drug metabolism” (10 gene sets), respectively, according to their gene annotations. Notably, 138 out of 259 (53%) survival-associated gene sets are involved in cell cycle and proliferation, highlighting the importance of this biological process in determining liposarcoma prognosis (27). In the network construction we also incorporated the *E2F1* and *E2F4* programs defined based on ChIP-seq data. As expected, genes in these programs are enriched in “cell cycle and proliferation” and thus these two programs fall into that gene set cluster (Figure 6, red dots).

Strikingly, the gene set analysis also identified three drug-related gene sets that are significantly correlated with liposarcoma patient survival (Figure 6, black dots). These gene sets are “Martinez response to Trabectedin DN” (genes down-regulated in sarcoma cell lines after treatment with Trabectedin) (28), “Keen response to Rosiglitazone UP” (genes up-regulated in aorta biopsies from mice treated with Rosiglitazone) (29), and “Ruan response to TNF Troglitazone UP” (genes up-regulated in 3T3-L1 cells in response to Troglitazone and TNF) (30). Interestingly, Trabectedin is a known anti-tumor drug and has been approved for the treatment of advanced soft tissue sarcoma (31). Rosiglitazone and Troglitazone are anti-diabetic drugs that belong to the thiazolidinedione class and act as PPAR gamma agonists (32, 33). Our result suggests the program regulated by these drugs might be associated with survival of patients with liposarcoma. Interestingly, both of the two drugs are undergoing phase II clinical trials (34, 35) for liposarcoma treatment. This suggests that the framework introduced in this study might be used to reposition old drugs for treating liposarcoma.

Discussion

In this article, we propose a new method to integrate multi-level data sources to identify the regulatory programs behind survival outcome in liposarcoma patients. Instead of directly referring to the mRNA expression of transcriptional regulatory factors, we infer the regulatory activities of TFs in samples based on the mRNA expression of their target genes, which are in turn determined by high-throughput ChIP-seq experiments.

In this study, we utilized the ChIP-seq data generated by ENCODE to define TF regulatory programs. In total, 424 TF binding profiles for 119 human TFs were investigated. However, these TFs only represent <10% of all human TFs. Thus, a caveat of this analysis is that many TFs critical for liposarcoma oncogenesis have not been included for the lack of data. In addition, many TFs regulate different sets of target genes under different conditions and in different tissues. Another caveat of the study is the lack of liposarcoma tissue-specific ChIP-Seq data. To address this issue, we define target gene sets for *E2F1* and *E2F4* by including the common targets identified by ChIP-seq across multiple different cell lines. These gene sets might represent a core set of target genes that are regulated by TFs in different tissues, and thus can be used for investigating transcriptional regulation in liposarcoma samples.

Furthermore, we did not identify any significant prognostic TF regulatory programs using the TCGA liposarcoma RNA-seq data, possibly due to the short follow-up times of the patients in the dataset.

Our analysis suggests that TF binding profiles in other cell lines might be informative for inferring TF activity in a specific tissue. For example, we show that *E2F4* programs defined based on ChIP-seq data in HeLa S3 and K562b can be used to characterize *E2F4* regulatory activity in liposarcoma and detect the association with patient survival. The utility of a ChIP-seq dataset is determined by (i) the similarity between the tested tissue and the interested tissue, and (ii) the degree of tissue-specificity of the interested TF. On the other hand, the usage of core target genes may also be problematic. Since they represent a subset rather than a complete list of target genes for E2F1 or E2F4, it is possible that the resulting scores do not reflect E2F1/E2F4 activities. Rather than, they reflect the activities of other cell cycle regulatory TFs that share a significantly large number of these core target genes with E2F1 or E2F4.

E2F1 and *E2F4* both belong to the E2F family and play related but different roles in the control and balance of cell cycle regulation. In accordance with other studies, *E2F1* is over-expressed in liposarcoma as reflected by its high regulatory score. In the case of *E2F4*, there are no significant changes in mRNA expression but our analysis shows it exhibits significantly increased regulatory activity. This suggests that the regulatory activity of *E2F4* cannot be reflected by mRNA level presumably due to post-transcriptional and post-translational regulation (36, 37). In fact, it has been shown that E2F phosphorylation and localization can be regulated by MEK/ERK activation, leading to changes in its transcriptional activity (36). Although *E2F1* and *E2F4* have different regulatory effects (i.e. activation/ repression), both exhibit enhanced activity in liposarcoma. This is consistent with prior findings that *E2F4* is involved in complex, wide-ranging functions and not solely limited to cell cycle regulation (38). Furthermore, over-expression of *E2F1* has been found to suppress cell differentiation while *E2F4* has been found to promote cell differentiation. This may explain why the *E2F1* and *E2F4* signatures exhibit differing predictive preferences in the dedifferentiated and well-differentiated liposarcoma subtypes.

As key factors in cell cycle regulation, *E2F1* and *E2F4* play critical roles in different cancer types. We and other researchers have shown the prognostic value of E2F4 in breast cancer (21, 39) and bladder cancer (40). Molina-Privado et al. have investigated its function in Burkitt lymphoma tumorigenesis, and found reduced protein but not mRNA levels of E2F4 in sporadic Burkitt lymphoma cell lines and tumor samples (41). Hijacking of cell cycle control happens in almost all malignant tumors (42). Factors in cell cycle related pathways have attracted much attention, with some of them having been considered as potential drug targets (43). For example, CDK4/6, when activated, promotes the transition from G1 to S phase via two mechanisms: (i) phosphorylation of the pRB protein to release and activate *E2F1*, and (ii) by phosphorylation of p107 and p130 to disrupt the DNA binding of *E2F4* (44, 45). It has been shown that sustained knockdown of CDK4 significantly inhibits the proliferation of tumor cells (9). In liposarcoma a notable overexpression of CDK4/6 has been observed. Currently, a CDK4/CDK6 inhibitor is under phase II clinical trials (46), which down-regulates *E2F1* and *E2F4* activity and arrests tumor cells in G1 phase. Using

the *E2F1* and *E2F4* signatures proposed in this study, it should be possible to predict the sensitivity of liposarcoma samples to this drug and other CDK4/CDK6 inhibitors.

The importance of *E2F1* and *E2F4* programs in prognosis has been further validated by gene set analysis. More than 50% of gene sets that are associated with prognosis in liposarcoma are involved in cell cycle and proliferation. In addition, gene set analysis also revealed other biological processes that are associated with liposarcoma prognosis, such as the PPAR pathway and the fatty-acid metabolism pathway. Moreover, we have identified three drug-related sets, one of which has been approved for treating sarcoma and the other two are under clinical trials for the treatment of liposarcoma.

Conclusions

In summary, we introduce a framework in this study that has successfully identified *E2F1* and *E2F4* regulatory signatures and applied them to predict prognosis in liposarcoma. It provides a useful tool that can be generally applied to other cancer types and human diseases as well.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the American Cancer Society Research Grant, #IRG-82-003-30, and the Dartmouth Clinical and Translational Science Institute under award number UL1TR001086 and KL2TR001088 from the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH).

Abbreviations

DRFS	Distant recurrence-free survival
REACTIN	Regulatory Activity Inference
BASE	Binding Association with Sorted Expression
ENCODE	the Encyclopedia of DNA Elements
ChIP-seq	Chromatin Immunoprecipitation followed by Massively Parallel Sequencing
iRAS	individual Regulatory Activity Score
FDR	False Discovery Rate

References

1. Jemal A, Ward E, et al. Cancer statistics. CA Cancer J Clin. 2009; 59:225–49. SR. [PubMed: 19474385]
2. Howlader, N.; Noone, AM.; Krapcho, M.; Garshell, J.; Miller, D.; Altekruse, SF., et al. SEER Cancer Statistics Review, 1975-2012. National Cancer Institute; Bethesda, MD: http://seer.cancer.gov/csr/1975_2012/, based on November 2014 SEER data submission, posted to the SEER web site, April 2015. ed2014

3. Burningham Z, Hashibe M, Spector L, Schiffman JD. The epidemiology of sarcoma. *Clinical sarcoma research*. 2012; 2:14. [PubMed: 23036164]
4. Singer S, Socci ND, Ambrosini G, Sambol E, Decarolis P, Wu Y, et al. Gene expression profiling of liposarcoma identifies distinct biological types/subtypes and potential therapeutic targets in well-differentiated and dedifferentiated liposarcoma. *Cancer research*. 2007; 67:6626–36. [PubMed: 17638873]
5. Fletcher, CD.; Mertens, F. Pathology and genetics of tumors of soft tissue and bone. IARC Press; 2002. UK
6. Fletcher CD, Akerman M, Dal Cin P, de Wever I, Mandahl N, Mertens F, et al. Correlation between clinicopathological features and karyotype in lipomatous tumors. A report of 178 cases from the Chromosomes and Morphology (CHAMP) Collaborative Study Group. *The American journal of pathology*. 1996; 148:623–30. [PubMed: 8579124]
7. Willmore-Payne C, Holden J, Turner KC, Proia A, Layfield LJ. Translocations and amplifications of chromosome 12 in liposarcoma demonstrated by the LSI CHOP breakapart rearrangement probe. *Archives of pathology & laboratory medicine*. 2008; 132:952–7. [PubMed: 18517278]
8. Gebhard S, Coindre JM, Michels JJ, Terrier P, Bertrand G, Trassard M, et al. Pleomorphic liposarcoma: clinicopathologic, immunohistochemical, and follow-up analysis of 63 cases: a study from the French Federation of Cancer Centers Sarcoma Group. *The American journal of surgical pathology*. 2002; 26:601–16. [PubMed: 11979090]
9. Barretina J, Taylor BS, Banerji S, Ramos AH, Lagos-Quintana M, Decarolis PL, et al. Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nature genetics*. 2010; 42:715–21. [PubMed: 20601955]
10. Gobble RM, Qin LX, Brill ER, Angeles CV, Ugras S, O'Connor RB, et al. Expression profiling of liposarcoma yields a multigene predictor of patient outcome and identifies genes that contribute to liposarcomagenesis. *Cancer research*. 2011; 71:2697–705. [PubMed: 21335544]
11. De Cecco L, Negri T, Brich S, Mauro V, Bozzi F, Dagrada G, et al. Identification of a gene expression driven progression pathway in myxoid liposarcoma. *Oncotarget*. 2014; 5:5965–77. [PubMed: 25115389]
12. Evans HL. Liposarcoma: a study of 55 cases with a reassessment of its classification. *The American journal of surgical pathology*. 1979; 3:507–23. [PubMed: 534388]
13. Asano N, Susa M, Hosaka S, Nakayama R, Kobayashi E, Takeuchi K, et al. Metastatic patterns of myxoid/round cell liposarcoma: a review of a 25-year experience. *Sarcoma*. 2012; 2012:345161. [PubMed: 22550416]
14. Eilber FC, Brennan MF, Eilber FR, Eckardt JJ, Grobmyer SR, Riedel E, et al. Chemotherapy is associated with improved survival in adult patients with primary extremity synovial sarcoma. *Annals of surgery*. 2007; 246:105–13. [PubMed: 17592298]
15. Smith TA, Easley KA, Goldblum JR. Myxoid/round cell liposarcoma of the extremities. A clinicopathologic study of 29 cases with particular attention to extent of round cell liposarcoma. *The American journal of surgical pathology*. 1996; 20:171–80. [PubMed: 8554106]
16. Singer S, Antonescu CR, Riedel E, Brennan MF. Histologic subtype and margin of resection predict pattern of recurrence and survival for retroperitoneal liposarcoma. *Annals of surgery*. 2003; 238:358–70. discussion 70-1. [PubMed: 14501502]
17. Lahat G, Anaya DA, Wang X, Tuvlin D, Lev D, Pollock RE. Resectable well-differentiated versus dedifferentiated liposarcomas: two different diseases possibly requiring different treatment approaches. *Annals of surgical oncology*. 2008; 15:1585–93. [PubMed: 18398663]
18. Tran D, Verma K, Ward K, Diaz D, Kataria E, Torabi A, et al. Functional Genomics Analysis Reveals a MYC Signature Associated with a Poor Clinical Prognosis in Liposarcomas. *The American journal of pathology*. 2015; 185:717–28. [PubMed: 25622542]
19. Zhu M, Liu CC, Cheng C. REACTIN: regulatory activity inference of transcription factors underlying human diseases with application to breast cancer. *BMC genomics*. 2013; 14:504. [PubMed: 23885756]
20. Cheng C, Yan X, Sun F, Li LM. Inferring activity changes of transcription factors by binding association with sorted expression profiles. *BMC bioinformatics*. 2007; 8:452. [PubMed: 18021409]

21. Khaleel SS, Andrews EH, Ung M, DiRenzo J, Cheng C. E2F4 regulatory program predicts patient survival prognosis in breast cancer. *Breast cancer research : BCR*. 2014; 16:486. [PubMed: 25440089]
22. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*. 2012; 489:91–100. [PubMed: 22955619]
23. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research*. 2013; 41:D991–5. [PubMed: 23193258]
24. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011; 27:1739–40. [PubMed: 21546393]
25. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics*. 2006; 7(Suppl 1):S7. [PubMed: 16723010]
26. Paramio JM, Segrelles C, Casanova ML, Jorcano JL. Opposite functions for E2F1 and E2F4 in human epidermal keratinocyte differentiation. *The Journal of biological chemistry*. 2000; 275:41219–26. [PubMed: 11005809]
27. Varn FS, Ung MH, Lou SK, Cheng C. Integrative analysis of survival-associated gene sets in breast cancer. *BMC medical genomics*. 2015; 8:11. [PubMed: 25881247]
28. Martinez N, Sanchez-Beato M, Carnero A, Moneo V, Tercero JC, Fernandez I, et al. Transcriptional signature of Ecteinascidin 743 (Yondelis, Trabectedin) in human sarcoma cells explanted from chemo-naïve patients. *Molecular cancer therapeutics*. 2005; 4:814–23. [PubMed: 15897246]
29. Keen HL, Ryan MJ, Beyer A, Mathur S, Scheetz TE, Gackle BD, et al. Gene expression profiling of potential PPARgamma target genes in mouse aorta. *Physiological genomics*. 2004; 18:33–42. [PubMed: 15054141]
30. Ruan H, Pownall HJ, Lodish HF. Troglitazone antagonizes tumor necrosis factor-alpha-induced reprogramming of adipocyte gene expression by inhibiting the transcriptional regulatory functions of NF-kappaB. *The Journal of biological chemistry*. 2003; 278:28181–92. [PubMed: 12732648]
31. Brodowicz T. Trabectedin in soft tissue sarcomas. *Future oncology*. 2014; 10:s1–5. [PubMed: 25048043]
32. Yau H, Rivera K, Lomonaco R, Cusi K. The future of thiazolidinedione therapy in the management of type 2 diabetes mellitus. *Current diabetes reports*. 2013; 13:329–41. [PubMed: 23625197]
33. Koffarnus RL, Wargo KA, Phillippe HM. Rivoglitazone: a new thiazolidinedione for the treatment of type 2 diabetes mellitus. *The Annals of pharmacotherapy*. 2013; 47:877–85. [PubMed: 23632280]
34. Demetri GD, Fletcher CD, Mueller E, Sarraf P, Naujoks R, Campbell N, et al. Induction of solid tumor differentiation by the peroxisome proliferator-activated receptor-gamma ligand troglitazone in patients with liposarcoma. *Proceedings of the National Academy of Sciences of the United States of America*. 1999; 96:3951–6. [PubMed: 10097144]
35. Debrock G, Vanhentenrijk V, Sciort R, Debiec-Rychter M, Oyen R, Van Oosterom A. A phase II trial with rosiglitazone in liposarcoma patients. *British journal of cancer*. 2003; 89:1409–12. [PubMed: 14562008]
36. Paquin MC, Cagnol S, Carrier JC, Leblanc C, Rivard N. ERK-associated changes in E2F4 phosphorylation, localization and transcriptional activity during mitogenic stimulation in human intestinal epithelial crypt cells. *BMC cell biology*. 2013; 14:33. [PubMed: 23919615]
37. Morillo SM, Abanto EP, Roman MJ, Frade JM. Nerve growth factor-induced cell cycle reentry in newborn neurons is triggered by p38MAPK-dependent E2F4 phosphorylation. *Molecular and cellular biology*. 2012; 32:2722–37. [PubMed: 22586272]
38. Lee BK, Bhinge AA, Iyer VR. Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. *Nucleic acids research*. 2011; 39:3558–73. [PubMed: 21247883]

39. Bertucci F, Finetti P, Birnbaum D. The E2F4 prognostic signature is also predictive of the pathological response of breast cancer to chemotherapy. *Breast cancer research : BCR*. 2015; 17:54. [PubMed: 25887620]
40. Cheng C, Varn FS, Marsit CJ. E2F4 Program Is Predictive of Progression and Intravesical Immunotherapy Efficacy in Bladder Cancer. *Molecular cancer research : MCR*. 2015; 13:1316–24. [PubMed: 26032289]
41. Molina-Privado I, Jimenez PR, Montes-Moreno S, Chiodo Y, Rodriguez-Martinez M, Sanchez-Verde L, et al. E2F4 plays a key role in Burkitt lymphoma tumorigenesis. *Leukemia*. 2012; 26:2277–85. [PubMed: 22475873]
42. Jones RG, Thompson CB. Tumor suppressors and cell metabolism: a recipe for cancer growth. *Genes & development*. 2009; 23:537–48. [PubMed: 19270154]
43. Lapenna S, Giordano A. Cell cycle kinases as therapeutic targets for cancer. *Nature reviews Drug discovery*. 2009; 8:547–66. [PubMed: 19568282]
44. Scime A, Li L, Ciavarrà G, Whyte P. Cyclin D1/cdk4 can interact with E2F4/DP1 and disrupts its DNA-binding capacity. *Journal of cellular physiology*. 2008; 214:568–81. [PubMed: 17894419]
45. Nevins JR. The Rb/E2F pathway and cancer. *Human molecular genetics*. 2001; 10:699–703. [PubMed: 11257102]
46. Dickson MA, Tap WD, Keohan ML, D'Angelo SP, Gounder MM, Antonescu CR, et al. Phase II trial of the CDK4 inhibitor PD0332991 in patients with advanced CDK4-amplified well-differentiated or dedifferentiated liposarcoma. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2013; 31:2024–8. [PubMed: 23569312]
47. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*. 2002; 30:207–10. [PubMed: 11752295]
48. Cox DR. Regression Models and Life-Tables. *Journal of the Royal Statistical Society, Series B*. 1972; 34:187–220.
49. 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*. 1995; 57:289–300.
50. Cheng C, Min R, Gerstein M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics*. 2011; 27:3221–7. [PubMed: 22039215]
51. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*. 2005; 102:15545–50. [PubMed: 16199517]
52. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature methods*. 2012; 9:471–2. [PubMed: 22426491]
53. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, et al. A travel guide to Cytoscape plugins. *Nature methods*. 2012; 9:1069–76. [PubMed: 23132118]
54. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*. 2009; 37:1–13. [PubMed: 19033363]

Figure 1. The overflow of our analysis

Figure 2. Differential activity and mRNA expression relationship of *E2F1*, *E2F4* and *NR3C1* between liposarcoma tumor samples and normal fat tissue samples.

(A-C) The regulatory activities calculated by the REACTIN method for *E2F1*, *E2F4*, and *NR3C1* in different subtypes of liposarcoma. Numbers above each plot are FDR values of the difference between samples in the corresponding subtype and the normal fat control. For *E2F4* (panel B), data from experiments in two cell lines, HeLaS3 and K562, are represented. (D-E) Distributions of mRNA expression levels in normal fat controls and samples from different liposarcoma subtypes. Numbers above each plot are adjusted p-values (FDR) of the difference between that subtype and the normal fat control. Ctrl: normal fat; All: all liposarcoma; DD: Dedifferentiated, Mrc: Myxoid/round-cell; Ple: Pleomorphic. The horizontal dotted lines indicate the median of normal fat expression.

Figure 3. Heatmap of iRAS score among 140 liposarcoma samples

The color map is shown on the top left and the column side bar represents subtypes of liposarcoma, in different colors, as indicated by legend on the left. Liposarcoma samples are clustered by iRAS, with each row corresponding to the regulatory activity of each ChIP-Seq Experiment's TF binding profile.

Figure 4. Survival association for *IRF3*, *FOS*, *E2F1* and *E2F4* iRAS

The liposarcoma samples are dichotomized into iRAS>0 and iRAS<0 groups. The adjusted p-value is indicated in the bottom left of each Kaplan-Meier curve.

Figure 5. Survival prediction for dedifferentiated and well-differentiated liposarcoma
 (A-B) Patient stratification based on their *E2F1* scores. (C-D) Patient stratification based on their *E2F4* scores. In well-differentiated samples, the peak value of in the distribution of *E2F4* scores was chosen as the threshold to stratify samples into high and low *E2F4* groups (B and D), since *E2F4* scores are negative in most samples.

Figure 6. Survival-associated gene sets clustered by function

Gene sets (nodes) shared significantly large number of genes are connected in the network. The gene sets cluster into 5 main groups (colored) by function as indicated by the legend. The *E2F1* and *E2F4* gene signatures are highlighted in red. Gene sets associated with drug response are colored black and described as indicated by the dotted arrows.

Table 1

Regulatory programs significantly associate with liposarcoma patient survival.

TF	Cell line	ENCODE ID	P-value	FDR	Hazard ratio (95% CI)
IRF3	HeLa S3	SydhTfbsHelas3Irf3IggrabAln	5E-07	2E-04	1.42 _(1.24,1.63)
<i>E2F1</i>	HeLa S3	SydhTfbsHelas3Ha <i>E2F1</i> StdAln	3E-06	5E-04	1.35 _(1.19,1.53)
CFOS	HeLa S3	SydhTfbsHelas3CfosStdAln	4E-06	5E-04	1.52 _(1.27,1.81)
P300	Sknshra	HaibTfbsSknshraP300V0416102Aln	5E-06	5E-04	1.48 _(1.25,1.76)
KAP1	U2OS	SydhTfbsU2osKap1UcdAln	9E-06	7E-04	1.49 _(1.25,1.77)
REST	U87	HaibTfbsU87NrsfPcr2xAln	1E-05	7E-04	1.39 _(1.20,1.61)
<i>E2F1</i>	MCF7	SydhTfbsMcf7Ha <i>E2F1</i> UcdAln	1E-05	7E-04	1.18 _(1.10,1.28)
GTF2F1	HeLa S3	SydhTfbsHelas3Gtf2f1rapIggrabAln	1E-05	7E-04	1.60 _(1.30,1.98)
<i>E2F1</i>	HeLa S3	SydhTfbsHelas3 <i>E2F1</i> StdAln	2E-05	9E-04	1.33 _(1.17,1.51)
REST	Pfsk1	HaibTfbsPfsk1NrsfPcr2xAln	2E-05	9E-04	1.34 _(1.17,1.53)
<i>E2F4</i>	HeLa S3	SydhTfbsHelas3 <i>E2F4</i> StdAln	2E-05	9E-04	1.14 _(1.07,1.21)
TAF7	H1hESC	HaibTfbsH1hescTaf7sq8V0416102Aln	3E-05	9E-04	1.40 _(1.20,1.64)
JUND	H1hESC	HaibTfbsH1hescJundV0416102Aln	4E-05	1E-03	1.59 _(1.28,1.98)
<i>E2F4</i>	K562b	SydhTfbsK562b <i>E2F4</i> UcdAln	4E-05	1E-03	1.18 _(1.09,1.27)
IRF3	HepG2	SydhTfbsHepg2Irf3IggrabAln	4E-05	1E-03	1.41 _(1.20,1.66)

The hazard ratio and 95% confidence interval (CI) are calculated using univariate cox proportional hazards models with iRAS score as the single variable.