Online Methods

Soft tissue sarcoma specimens. 207 tumor and 205 matched normal DNA and RNA specimens (blood, fat, or adjacent tissue) were procured at Memorial Sloan-Kettering Cancer Center (MSKCC) from patients diagnosed with one of seven high-grade soft-tissue sarcoma subtypes (Table 1 and Supplementary Table 1). Specimens were macro-dissected from two 5mm H&E histological sections cut from the top and bottom of a cryomold to assess subtype, percent necrosis, cellularity, and stromal contamination. Analyses were conducted from regions of cellularity >75% and no mixed histology, necrosis, or fibrosis. This study was conducted with institutional review board approval and all individuals provided informed consent.

Sequencing. In total, we sequenced 226 protein-coding genes selected based on their validated or hypothesized role in tumorigenesis determined from the literature and the CancerGenes resource[1]. These included 90 tyrosine kinases, 90 oncogenes and tumor suppressors, 22 sarcoma-related genes, 6 telomere maintenance/stability genes, and 18 PI3K pathway genes not in the former groups (Supplementary Table 2). We also sequenced 496 microRNAs (87.5% overlap with release 11 of miRBASE)[2]. Samples sequenced (n=47, Supplementary Table 1) had LOH on at least one chromosome arm (by SNP array analysis) indicating sufficient tumor purity. In total, 3849 exons and adjacent intronic sequences were PCR amplified and Sanger sequenced. This produced ~27.9Mb of sequence (~0.58Mb/sample). Analysis was performed only on amplicons with data for 80% of bases in 80% of the samples, yielding 22.9Mb (82.1%) and 15Mb (53.8%) of sequence with 1 and 2x coverage respectively.

Mutation pipelines and validation. Mutations were initially called using the Broad Institute SNPCompare pipeline, as previously described[35]. Analysis of indels was facilitated by PolyTedh (Sharpe T. et al, unpublished). The following pipeline, independently developed at MSKCC, was used to supplement mutation detection. Bi-directional reads and sample/gene annotation were consumed from the initial Broad production run. Reads were excluded if their average Phred score was <20 (for bases 100-200) or if they lacked a valid sample, gene, direction, or amplicon mapping. Qualified reads were segregated to produce a reference sequence containing coding/UTR exons and ±5Kb up- and down-stream of the transcription start/stop for any transcripts representing the gene. Reads were assembled against the reference using Consed 16.0[36]. Assemblies were processed with PolyPhred 6.02b[37] and SNPdetector v2[38] and candidate mutation lists were merged without additional filtering. All putative mutations were annotated using the Genomic Mutation Consequence Calculator[39]. Only mutation calls that were supported in both forward and reverse reads, non-synonymous, indels affecting exonic sequence, or splice site variants (defined as ±4bp from an exon boundary) were merged with the Broad pipeline output (Supplementary Figure 1A). After eliminating intronic candidates (4355, 77.8%), synonymous mutations (633, 11.3%), and known SNPs (139, 2.5%; dbSNP build 128), 409 and 377 candidate mutations were submitted for mass spectrometric genotyping validation (from the two pipelines respectively).

To eliminate artifacts, false positives, or germline mutations, candidate mutations were independently confirmed in tumor and matched normal samples with mass spectrometry-based genotyping[40]. Of 786 candidate mutations genotyped, 376 (47.8%) were false positives, 296 (37.7%) germline mutations, 35 (4.5%) resulted from LOH in the tumor, 4 (0.5%) had somatic allele mismatch, 32 (4.1%) failed genotyping,

and manual review of traces/spectra eliminated 15 mutations (2%). Therefore, 28 (1.3%) candidate mutations were confirmed somatic. All these were genotyped in the remaining 160 tumor/normal pairs, yielding 1 additional somatic mutation. We also performed full exon re-sequencing of several mutated genes in all samples of selected subtypes (Supplementary Figure 1). Finally, an independent cohort of 50 myxoid/round-cell liposarcomas[16] was genotyped for 13 common *PIK3CA* point mutations (R88Q, N345K, C420R, E542K, E542Q, E545A, E545G, E545D, E545K, M1043I, H1047L, H1047R, H1047Y).

Predicted function impact of somatic mutations. To predict the likely functional impact of missense mutations we used a computation methodology, Mutation Assessor (http://mutationassessor.org/; Reva et al., in preparation), which predicts the affect of a given mutated residue on protein function using evolutionary information from protein-family sequence alignments[13] and residue placement in known or homology-deduced three-dimensional protein and complex structures (Supplementary Note).

SNP and expression arrays. Genomic DNA was hybridized to the StyI chip (~238,000 SNPs of 500K Mapping Array Set, Affymetrix) following manufacturer's instructions. Batches of 96 samples were processed as a plate, with tumor and normal pairs in adjacent wells. Samples/plates were tracked using ABGene 2D barcode tube/rack readers (ABGene). Transcript expression data was obtained for 149 tumors profiled on SNP arrays (Supplementary Table 1). Cryomold tissue specimens were processed for RNA extraction and this was hybridized on HG-U113A oligonucleotide arrays (Affymetrix) as previously described[15]. Expression of probe sets was estimated with Robust multi-array average (RMA)[41].

Array analysis. Two independent methods, GISTIC[22] and RAE[23], were used to assess the significance of genome-wide copy number aberrations (CNAs) in each of the seven subtypes. For GISTIC analysis, SNP array raw intensity files were processed and normalized as previously described[22] and as implemented in GenePattern[42]. Separately, RAE was run as previously described on the tumors of each subtype[23]. Known karyotypic complexity determined the false discovery rate (FDR) at which genomic regions were considered significantly altered and regions of interest determined (complex and simple karyotypes, FDR $\leq$ 1% and 10% respectively). Segmented copy number data used by either GISTIC or RAE analysis was visualized with the Integrative Genomics Viewer (see URLs). All mapping information was based on Affymetrix annotations and the hg17 build of the human genome sequence.

For RAE analysis, germline CNVs were determined from 173 qualifying normal samples (Supplementary Table 1) after removing those whose distribution of normalized segmentation (i) was grossly asymmetric (negative/positive skew in diploid peak), (ii) was incoherently multi-modal, (iii) bore isolated copy number genotypes similar to tumors (upon manual review), or (iv) >2% of autosomal genome was altered. Any segment exceeding A_0 or $D_0 \geq 0.5$ was considered altered and presumed polymorphic. These were combined with a subset of known variants (DGV, version 3)[43]. Tumor-specific regions identified by RAE were presumed polymorphic if they appeared in two or more of these sources (or a single matched normal) and had sequence coverage >50%. Putative polymorphisms were then manually reviewed. Finally, in regions of statistically significant CNA (Supplementary Table 5), likely aberration states were assigned to isoforms of resident RefSeq genes in each tumor: homozygous deletion ($D_1 \geq 0.9$), heterozygous loss ($D_0 \geq 0.9$ and $D_1 < 0.9$), copy-

neutral (D_0 and $A_0 < 0.9$), single-copy gains ($A_0 \geq 0.9$ and $A_1 < 0.1$), and multi-copy amplification ($A_0 \geq 0.9$ and $A_1 \geq 0.1$). We assessed the correlation between transcript expression and copy number with ANOVA and a nominal p-value < 0.05 was considered correlated.

We generated a consensus copy number profile in dedifferentiated liposarcoma (as in Figure 4A) from alterations detected by both GISTIC and RAE using the Fisher method. P-values were extracted from either method (representing 193,926 consensus SNPs). Assuming the null hypothesis is true, p-values are approximately uniformly distributed, so a statistic k is approximately distributed as a chi-square with $2n$ degrees of freedom. We let k equal:

$$k = \sum_{i=2}^{n} -2\log(p_i)$$

where p_i is the p-value for a SNP from one of the two methods i (n=2). A combined p-value was assigned to each locus reflecting the consensus significance of both methods. For multiple hypothesis correction (Benjamini and Hochberg FDR), we reduced these to putatively independent observations by calculating the median p-value of all SNPs spanning each region of the unified breakpoint profile derived by RAE (therefore correcting 14,814 versus 193,926 tests). The result is a single significance profile of SCNAs in dedifferentiated liposarcoma.

LOH analysis. LOH was determined for 205 tumor/normal sample pairs by the hidden Markov model (HMM) algorithm in dChipSNP[44]. The genotyping error rate was increased to 0.2 otherwise default parameters were used. Seven pairs were removed due to presumptive low data quality (>1000 segments of LOH segments per sample). For each subtype, GISTIC analysis of LOH was performed as described[22] and overlaid with the GISTIC results for copy loss (Supplementary Table 4).

Cell line analysis. LPS141 and DDLS8817 were derived from high-grade retroperitoneal dedifferentiated liposarcomas[15,45,46], as was FU-DDLS-1[47]. All were maintained in DMEM:F12 1:1 + 10% fetal bovine serum (FBS). DNA was extracted with the DNAeasy tissue kit (Qiagen, CA) and SNP array data obtained as described for the primary tumors and data analyzed with RAE as described above. The amplification status for the 385 genes with shRNA proliferation data was determined in each cell line as loci with A_0 and A_1 exceeding 0.9 and 0.01 respectively[23].

RNA interference screen. In total, 444 genes were identified as significantly amplified in dedifferentiated liposarcoma from a preliminary GISTIC analysis and an early-access version of RAE on a subset (80%) of the final tumor cohort. Of these, 385 genes had one or more shRNA targeting its transcript in the TRC library[48] and were selected for screening (Supplementary Table 2). After optimizing growth conditions, plate types, viral dose, and assay times, LPS141 and DDLS8817 cells were plated in 50ml at 700 cells/well and FU-DDLS-1 cells were plated at 900 cells/well in 384-well plates (Costar 3712) using the MicroFill (BioTek, VT) robot. Medium was replaced at 24hr with 8μg/ml polybrene using the RapidPlate (Qiagen) or Janus (PerkinElmerMA) automated pipetting workstations. Cells were then transduced with 1μl of shRNA lentiviral supernatants using the PerkinElmer Evolution P3 robot. Plates were spun at 2250rpm for 30 minutes and incubated overnight at 37°C. Infections were performed in quadruplicate, with two replicates placed under puromycin (2μg/ml) selection 24hrs post-transduction. Cell proliferation was measured 5 days post-infection with ATPLite

(PerkinElmer). Barcoded plates were read on a PerkinElmer EnVision plate reader. The RNAeye tool (A. Derr) was used to examine infection efficiency, data quality, run plate replicate comparisons, generate hairpin distributions by plate or batch, and export raw/normalized data by well, hairpin, or gene.

Screen analysis and validation. Raw hairpin/well proliferation values were first normalized to Z-scores based on the global plate mean and standard deviation of proliferation scores[49]. To select hits in each cell line, we used the probabilistic methodology RSA (Supplementary Note). Differential expression of the 385 genes in tumors relative to normal adipose tissue, or between tumors amplified for the gene and those diploid, was determined with an empirical Bayes t-test[50], or a two-sample Wilcoxon rank sum test respectively and after multiple hypothesis correction, genes were considered significant at q-value < 0.05 (FDR<5%)[51].

For validation experiments, shGFP-437 and a pLKO.1 virus containing a scrambled 18bp insert were used as controls, pLKO.1 lentivirus containing shRNAs targeting *CDK4* were from TRC, and those targeting *MDM2* and *YEATS4*, also generated by TRC, were obtained through Open Biosystems. The sequenced targeted by the following clones can be found at the TRC URL: YEATS4-Y1 (TRCN0000013143), YEATS4-Y3 (TRCN0000013145), YEATS4-Y4 (TRCN0000013146), CDK4-C1 (TRCN0000000362), CDK4-C4 (TRCN0000000363), CDK4-C8 (TRCN0000010520), MDM2-M76 (TRCN0000003376), MDM2-M77 (TRCN0000003377), MDM2-M78 (TRCN0000003380).

For *CDK4*, cells were trypsinized on day 0 and re-suspended at $5x10^5$ cells per 2ml medium with 8µg/ml polybrene (Sigma). 2ml were seeded into each well of a 24-well plate (Fisher) and appropriate titer of lentivirus was added to achieve ~70% cell infection. A spin infection was performed at 2000rpm for 2hrs at 30°C. Cells were trypsinized and transferred to individual 10cm dishes. After 24hrs, 1.5µg/ml puromycin was added and infected cells were selected for 72hrs and then $2x10^4$ cells were seeded per well in 6-well plates. Cells were trypsinized and counted using a Vi-Cell XR Cell Viability Analyzer (Beckman Coulter) on days 1, 3, 4, 5, and 11 post-seeding (day 5, 7, 8, 9, and 15 post-infection). For *MDM2* and *YEATS4*, cells were grown to 75% confluence and then infected with lentivirus in 6-well plates with polybrene. On day 1, infected cells were selected with 1µg/µl puromycin. After 72hrs, 750 cells were plated per well in 96-well plates (Day 4). Plates were collected on several days post-seeding, washed with PBS, and frozen. Cell quantification was performed using the CyQUANT cell proliferation kit (Invitrogen) on a SpectraMax M2 reader. For western blotting, lysates were prepared from 10cm dishes 7 days post-infection. Fifty µg of lysate per sample was run on a Tris-glycine gel. Anti-CDK4 (DCS156, Cell Signaling), anti-MDM2, (sc-965, Sigma), anti-YEATS4, (sc-81278, Sigma), and anti-vinculin (Sigma) antibodies were used.