Supporting Information

Perry et al. 10.1073/pnas.1419260111

SI Materials and Methods

Whole Genome, Whole Exome, RNA Sequencing, and Analysis.

Library construction. Libraries are constructed using the protocol described previously, with several modifications: first, initial genomic DNA input into shearing has been reduced from 3 μg to 100 ng in 50 μL of solution (1). Second, for adapter ligation, Illumina paired-end adapters have been replaced with palindromic forked adapters with unique eight-base index sequences embedded within the adapter. These index sequences enable pooling of libraries before sequencing. Third, custom sample preparation kits from Kapa Biosciences are now used for all enzymatic steps of the library construction process.

In-solution hybrid selection (for whole-exome libraries). In-solution hybrid selection was performed as described by Fisher et al. (1).

Size selection (for whole-genome shotgun libraries). For a subset of samples, size selection was performed using gel electrophoresis, with a target insert size of either 340 bp or 370 bp ± 10%. Multiple gel cuts were taken for libraries that required high sequencing coverage. For another subset of samples, size selection was performed using Sage's Pippin Prep.

Preparation of libraries for cluster amplification and sequencing. Following sample preparation, libraries were quantified using PicoGreen. Based on PicoGreen quantification, libraries were normalized to equal concentration and pooled by equal volume. Library pools were then quantified using a Sybr Green-based quantitative PCR (qPCR) assay, with PCR primers complementary to the ends of the adapters (kit purchased from Kapa Biosciences). After qPCR quantification, library pools were normalized to 2 nM, denatured using 0.2 N NaOH, and diluted to 20 pM, the working concentration for downstream cluster amplification and sequencing.

Cluster amplification and sequencing. Cluster amplification and sequencing of denatured templates was performed according to the manufacturer's protocol (Illumina) using v3 cluster amplification kits, v3 flowcells, v3 Sequencing-by-Synthesis kits, Multiplexing Sequencing Primer kits, and the latest version of Illumina's RTA software.

Exome analysis. Pair-ended reads were aligned to the hg19/GRCh37 build of the reference human genome using BWA 0.5.9. WES data were generated for 59 pairs using in-solution hybrid capture followed by Illumina sequencing. Reads were aligned to build hg19/GRCh37 of the human reference genome sequence BWA. PCR-duplicated reads were flagged using Picard (2). Alignments near putative indel sites were refined using GATK, using both the tumor and the normal samples. The degree of contamination by other samples was estimated using ContEst (3). Somatic point mutations were detected using MuTect (4). Somatic short insertions and deletions were identified using indelocator (www.broadinstitute.org/cancer/cga/indelocator). Artifactual mutations caused by the oxidative DNA damage during library preparation were removed using D-ToxoG (www.broadinstitute.org/cancer/cga/dtoxog). Somatic mutations were annotated using Oncotator (www.broadinstitute.org/oncotator). Total copy number ratios were computed as the ratio of tumor fraction read depth to the average fractional read depth in the normal samples in the region, followed by Circular Binary Segmentation (5, 6). Copy number profiles were analyzed using GISTIC2 (7). Absolute copy number, purity/ploidy, and clonality analysis was done using ABSOLUTE (8).

RNA sequencing and analysis. RNA Reads were aligned to the hg19/GRCh37 build of the reference human genome using an improved algorithm described previously, followed by PCR duplicate-read removal and base quality score recalibration using GATK (9). Quality of RNASeq was assessed using RNASeqQC (10). PathSeq was used to discover pathogen sequences. Somatic mutations were discovered using MuTect and annotated using Oncotator. Expression levels were estimated by computing the reads per kilobase of exon model per million mapped reads (11).

Detection of fusions was done using a previously described algorithm (9). Briefly, the approach first identifies recurrent (i.e., two or more) chimeric pairs with both ends aligned and mapping in two different genes, on different chromosomes, or at least 1 Mb apart if on the same chromosome. It is also required that the pair end aligned in their respective genes in the direction consistent with coding→coding 5′-3′ direction of the (putative) fusion mRNA transcript. Next, all unaligned reads are extracted, with the constraint that their mates were originally aligned and map into one of the genes in the gene pairs obtained as described above. An attempt is then made to align all such originally unaligned reads to the special "reference" built of all possible exon–exon junctions (full length, boundary-to-boundary, in coding 5′-3′ direction) between the discovered gene pairs. If such originally unaligned read maps onto a junctions between an exon of gene X and an exon of gene Y, and its mate was indeed mapped to one of the genes X, Y, then such read ("junction split-read") is counted toward evidence for X–Y fusion (the consistency of the orientation of this read and its mate is also checked).

Detection of germline variants. Germline variants were detected using the UnifiedGenotyper in the Genome Analysis Toolkit (www.broadinstitute.org/gatk), using default options, followed by filtering SNPs using Variant Quality Score Recalibration, and hard-filtering of indels (12, 13). Germline variants were annotated using SeattleSeq137 (snp.gs.washington.edu/SeattleSeqAnnotation137). Nonsilent variants were identified as those in classes: frameshift, frameshift-near-splice, missense, missense-near-splice, splice-3, splice-5, stop-gained, stop-gained-near-splice, stop-lost. Germline de novo variants were discovered using xBrowse (atgu.mgh.harvard.edu/xbrowse). Germline variants were defined as rare if they were present in <0.5% of the National Heart, Lung, and Blood Institute Exome Variant Server, or EVS. Fisher's exact test was used to determine whether germline variants were significantly associated with osteosarcoma compared with EVS samples.

Integrated analysis of somatic variants. Mutation significance (MutSigCV) algorithm was used to identify significantly somatically mutated genes by using somatic mutations detected in WGS and WES (14). MutSigCV identified genes with higher mutation frequencies than expected by chance given multiple covariates: the gene's base composition, length, background mutation rate, and sequencing coverage.

Mutation validation using RNAseq, WGS, and targeted resequencing. Whenever possible, we validated mutations detected in WES in RNASeq and WSG using a previously described method (15). Briefly the method examines the variant allele fraction of reads in the validation BAM file and compares to the expectation from the discovery BAM file, corrected for the sequencing noise. When using only well-powered sites (with at least 90% power to detect mutations, after taking into account the estimated allelic fraction of the mutation and the depth of coverage), this approach validated of 359 of 364 (96.8%) of mutations.

Additionally, 20 genes were selected for targeted resequencing and validation: *AHDC1*, *AKT1*, *ALK*, *AVIL*, *CEP164*, *CHEK2*, *CREBBP*, *ESR1*, *NF1*, *NF2*, *PDPK1*, *PLCZ1*, *PPAT*, *PRKDC*, *PTEN*, *ROBO2*, *SFI1*, *SQSTM1*, *TP53*, and *TSC2*. Targeted

resequencing was performed by PCR using a microfluidic device (Fluidigm), following the manufacturer's instructions. PCR primers were designed with 200-bp flanking tails around mutations of interest. All amplicons for a given sample were given the same barcode. Constructed libraries were loaded onto an Illumina MiSeq and sequenced using paired-end 150-bp reads, followed by the standard alignment pipeline. The resulting BAM files were used for validation in the sense of the method described above.

Identification of samples with kataegis. C > T and C > G mutations were plotted according to intermutation distance (WES), along with copy number calculated by SegSeq (WGS) and genomic rearrangements analyzed with dRanger (WGS) and ChainFinder (WES) (9, 16–18). Regions of characteristic co-occurrence of local hypermutation and genomic rearrangement were identified as "kataegis" (19).

Detection of viral sequences. WGS, WES, and RNASeq data were examined for the presence of viral nucleic acid sequences using PathSeq (20). Viral sequences were downloaded from National Center for Biotechnology Information Nucleotide (www.ncbi.nlm.nih.gov/nucleotide) using search term "Viruses[Organism] AND srcdb_refseq[PROP] NOT cellular organisms[ORGN]" and "Viruses[Organism] NOT srcdb_refseq[PROP] NOT cellular organisms[ORGN] AND nuccore genome samespecies [Filter] NOT nuccore genome[filter] NOT gbdiv syn[prop]" on June 2012.

Analysis of copy number, genomic breakpoints, and rearrangements. Somatic copy number alterations were assessed in WGS using SegSeq (17). Integrative analysis of genomic breakpoints and copy number in WES samples was performed using the ChainFinder algorithm (16). Somatic rearrangements were identified in WGS samples with dRanger (18). Rearrangements in 13 OS WGS samples were compared with those found in a pan-cancer dataset consisting of 275 tumor/normal WSG pairs from The Cancer Genome Atlas (TCGA). The distribution of tumor types across samples was as follows: 49 THCA (papillary thyroid carcinoma), 40 LUAD (lung adenocarcinoma), 31 LUSC (lung squamous cell carcinoma), 31 GBM (glioblastoma multiforme), 25 SKCM (skin cutaneous melanoma), 24 STAD (stomach adenocarcinoma), 20 PRAD (prostate adenocarcinoma), 18 BLCA (bladder urothelial carcinoma), 17 HNSC (head and heck squamous cell carcinoma), 16 LGG (brain lower grade glioma) and 4 KIRC (kidney renal clear cell carcinoma). All TCGA datasets are available through The Cancer Genome Atlas Data Portal (tcga-data.nci.nih.gov/tcga) and GCHub (cghub.ucsc.edu).

Identification of significantly altered pathways. Gene sets from the Molecular Signatures Database (MSigDB) Canonical Pathway set (GSEA) were treated analogously as single genes for the purpose of calculation of the footprint and the background mutation rate (i.e., gene territory and composition was combined in each gene set). MutSig2.0 was then used to identify significantly mutated gene sets.

Heuristic algorithm for analysis of clinically relevant somatic mutations (PHIAL). All exome-derived alterations (somatic point mutations, short insertions and deletions, and copy number alterations) were analyzed using a heuristic algorithm that interprets the clinical and biological significance of each alteration in the exome (21). Clinical significance was defined by whether a specific alteration may predict sensitivity or resistance to a treatment, or has prognostic or diagnostic ramifications. All alterations scored as being potentially clinically actionable were manually reviewed.

shRNA Screening.

In vitro shRNA screening. Primary mouse OS cells were seeded into 12-well dishes at a density of 1×10^6 cells per well, with a total of 3×10^7 cells infected per replicate (four replicates total). Cells were infected with a pool of lentivirally delivered shRNAs, composed of 40,021 shRNAs targeting ~8,400 mouse genes with a multiplicity of infection of 0.3–0.5. Cells were incubated overnight with virus and 5 μg/mL polybrene. The next day, cells from each replicate were pooled and cultured in 0.5 μg/mL puromycin for 18 population doublings. During propagation, 1×10^7 cells were passaged every 3–4 d to maintain initial representation, and remaining cells at each passage were stored in PBS at −80 °C. Genomic DNA was extracted from the final cell pellets, and 60 μg of gDNA was used as template for PCR amplification in eight parallel bar-coded reactions for each experimental replicate. PCR reactions were prepared for massively parallel sequencing (Illumina), as previously described (22, 23). All samples were sequenced to obtain at least 8e6 raw reads. The number of reads per individual shRNA was normalized between samples using the following calculation: Log_2[([raw read count for hairpin/ sum of raw reads for entire sample] × 1e6)+1]. shRNAs were rank ordered by their log-twofold change value, which was calculated as the average normalized $\log_2$ of the fold-change in the abundance of each shRNA in the average of endpoint samples compared with the initial pDNA reference pool. Next, the RNAi gene enrichment (RIGER) algorithm in the GENE-E program (www.broadinstitute.org/cancer/software/GENE-E) was used to collapse the normalized shRNA ranked list to gene rankings by two comp methods: (*i*) the weighted second best score (ranked top shRNA 25% weight + second best shRNA 75% weight) and (*ii*) a KS statistic, which is a Kolmogorov–Smirnov nonparametric rank statistic representing the positional distribution of a set of shRNAs within an ordered list of shRNAs (22, 24). Lists were generated from the top 500 genes from each ranking method. The lists were trimmed by *P* value ($P \leq 0.05$) and common essential genes, including ribosomal proteins proteosomal proteins, and splicing factors (22). A union of the remaining 348 genes (weighted second best) (Dataset S11) and 313 genes (KS) (Dataset S12) was taken (Dataset S13).

In vivo shRNA screening. Plasmids encoding shRNAs targeting *Pic3ca*, *Mtor*, and control genes listed in Dataset S15 were used to generate lentivirus-containing supernatants, as previously described (22). Equivalent amounts of supernatants were pooled and primary mouse OS cells were infected as described for the genomic screen. Cells were selected for 2 d with 0.5 $\mu g \cdot mL^{-1}$ puromycin. Next, 1×10^6 cells were injected in 100 μL PBS subcutaneously into the flanks of NCRNU-M mice (Taconic). Tumors were harvested 5 wk after implantation. Genomic DNA was isolated from tumors and all available genomic DNA from each tumor was prepared for massively parallel sequencing as described above. The log-twofold change values reported are the average log base 2 of the fold-change in the abundance of each shRNA in the tumors compared with the preinjection cells, $n = 5$ tumors. All experiments involving mice were carried out with approval from the Boston Children's Hospital Animal Use and Care Committee.

Primers for amplifying shRNAs encoded in genomic DNA. Barcoded forward primer (**N** indicates location of sample-specific barcode sequence): AATGATACGGCGACCACCGAGAAAGTATTTCGATTTCTTGGCTTTATATATCTTGTGGAA**NN**GACGAAAC. Common reverse primer: CAAGCAGAAGACGGCATACGAGCTCTTCCGA TCTTGTGGATGAATACTGCCATTTGTCTCGAGGTC. Illumina sequencing primer: AGTATTTCGATTT CTTGGCTTTATATATCTTGTGGAA.

1. Fisher S, et al. (2011) A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biol* 12(1):R1.
2. Li H, et al.; 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.
3. Cibulskis K, et al. (2011) ContEst: Estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* 27(18):2601–2602.
4. Cibulskis K, et al. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 31(3):213–219.

5. Dulak AM, et al. (2013) Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat Genet* 45(5):478–486.
6. Olshen AB, Venkatraman ES, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5(4):557–572.
7. Mermel CH, et al. (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 12(4):R41.
8. Carter SL, et al. (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 30(5):413–421.
9. Berger MF, et al. (2010) Integrative analysis of the melanoma transcriptome. *Genome Res* 20(4):413–427.
10. DeLuca DS, et al. (2012) RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* 28(11):1530–1532.
11. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7): 621–628.
12. McKenna A, et al. (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20(9):1297–1303.
13. DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43(5):491–498.
14. Lawrence MS, et al. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499(7457):214–218.
15. Ojesina AI, et al. (2014) Landscape of genomic alterations in cervical carcinomas. *Nature* 506(7488):371–375.
16. Baca SC, et al. (2013) Punctuated evolution of prostate cancer genomes. *Cell* 153(3):666–677.
17. Chiang DY, et al. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* 6(1):99–103.
18. Drier Y, et al. (2013) Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res* 23(2):228–235.
19. Nik-Zainal S, et al.; Breast Cancer Working Group of the International Cancer Genome Consortium (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* 149(5):979–993.
20. Kostic AD, et al. (2011) PathSeq: Software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* 29(5):393–396.
21. Van Allen EM, et al. (2014) Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med* 20(6):682–688.
22. Luo B, et al. (2008) Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci USA* 105(51):20380–20385.
23. Whittaker SR, et al. (2013) A genome-scale RNA interference screen implicates NF1 loss in resistance to RAF inhibition. *Cancer Discov* 3(3):350–362.
24. Cheung HW, et al. (2011) Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc Natl Acad Sci USA* 108(30):12372–12377.

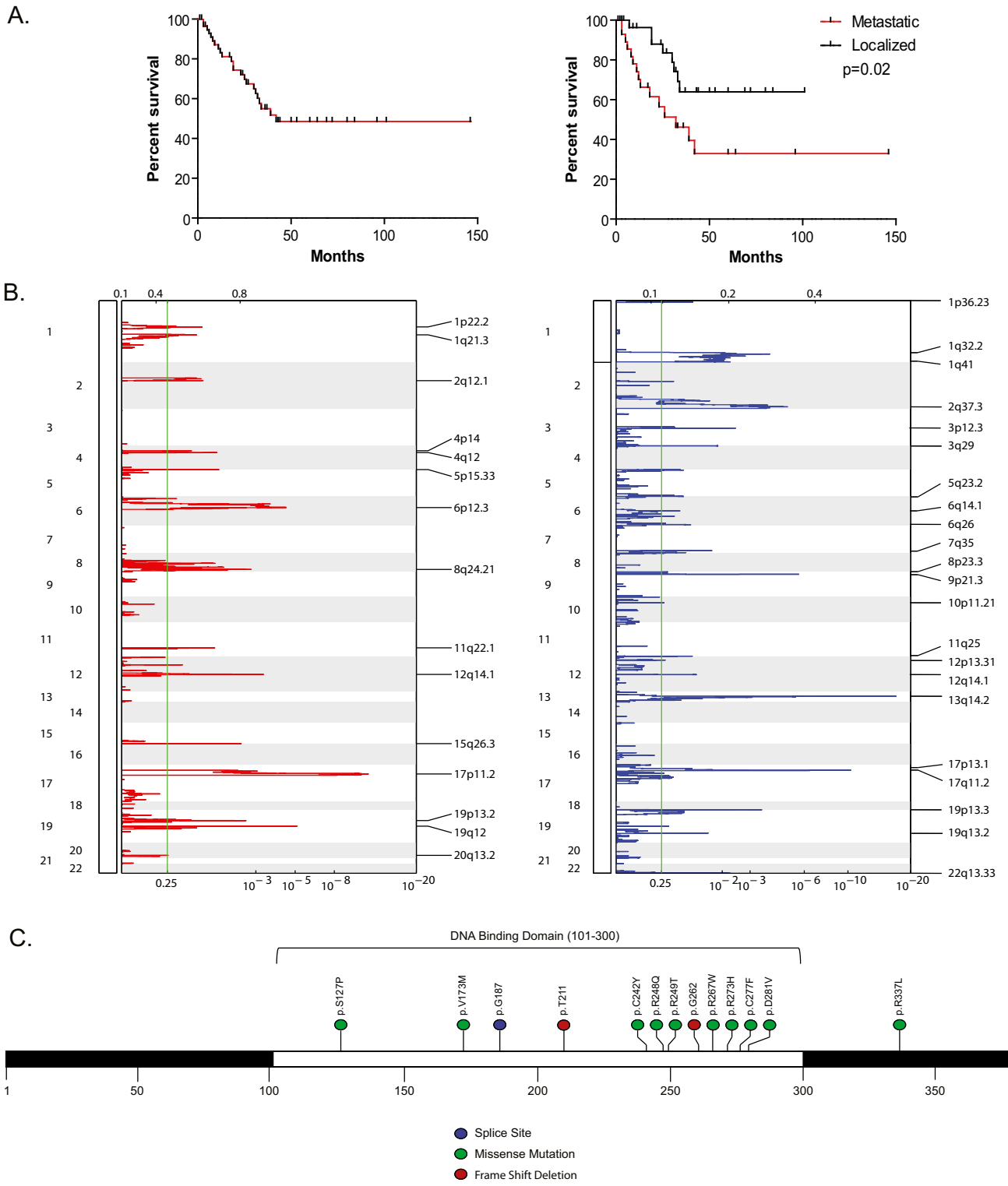

Fig. S1. (*A*) Overall survival (*Left*) and overall survival by disease status at time of diagnosis (*Right*) for 59 osteosarcoma patients in the study population. (*B*) Genomic positions of significant focal amplifications and deletions. Amplifications are depicted on the left (red) and deletions are depicted on the right (blue). Chromosomes are listed on the *y* axis. The *x* axis represents the normalized amplification signal (Top *x* axis) and the significance by *q* value (Bottom *x* axis). The green line represents the significance cutoff at *q* value = 0.25. (*C*) Schematic of the TP53 protein showing the distribution and type of somatic TP53 mutations (numbers indicate amino acid position). TP53 mutations were detected in 22% of samples by MutSigCV. Twelve of 13 TP53 mutations were in the DNA binding domain with 9 missence, 2 frameshift, and 1 splice site mutation. An additional missense mutation was found outside the DNA binding domain.

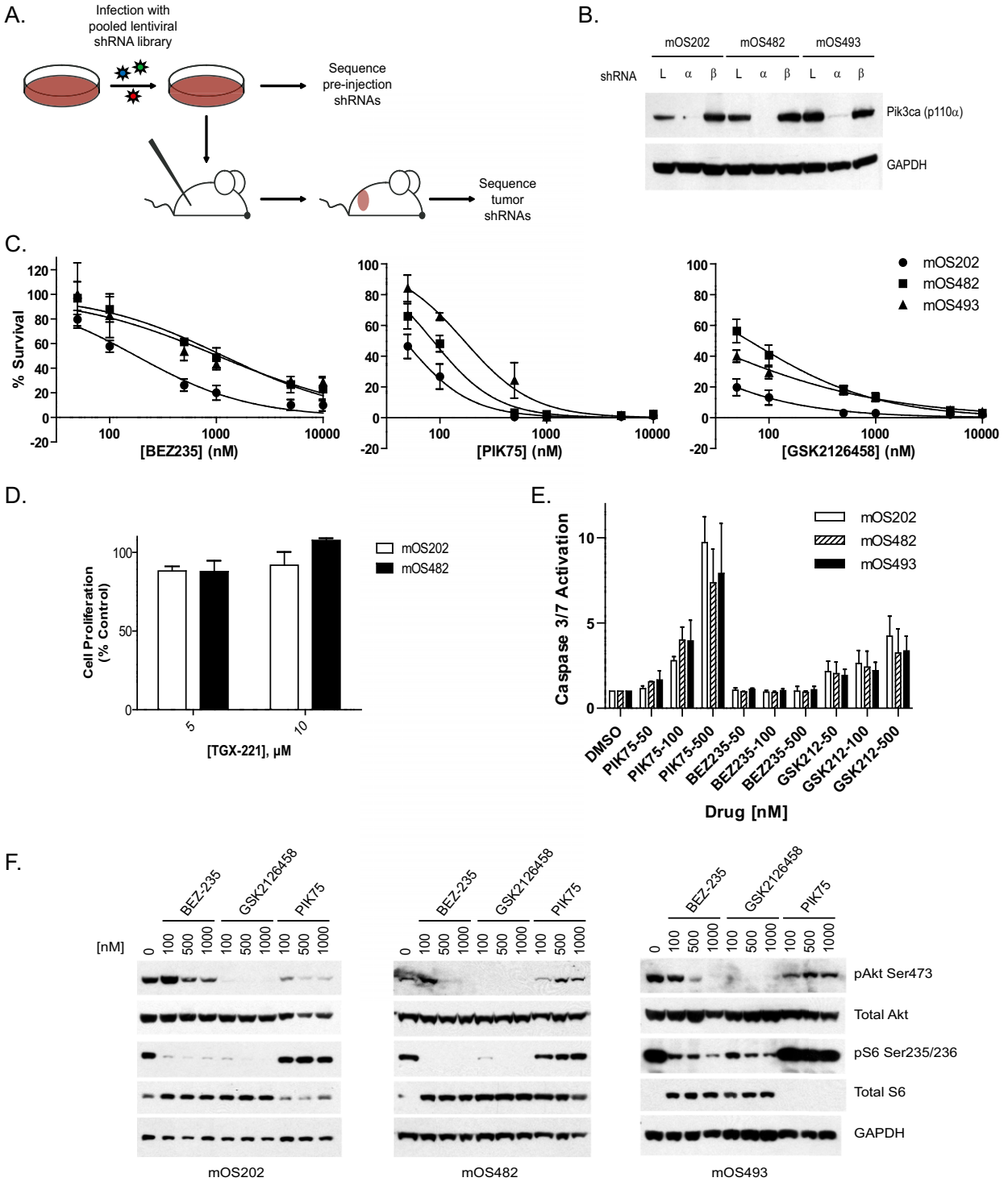

Fig. S2. (*A*) Outline of experimental design of targeted shRNA screen in osteosarcoma xenograft development. (*B*) Knockdown of PIK3CA in mOS cells. mOS cells were lentivirally transduced with shRNAs against luciferase (L), Pik3ca (α) or Pik3cb (β). Cell lysates were analyzed by Western blot 4 d postinfection and selection with puromycin. (*C*) mOS cells were exposed to BEZ235, PIK75, and GSK2126458 at the indicated concentrations for 72 h. Survival was measured by WST-1 assay using DMSO as the control. (*D*) mOS cells were exposed to 5 and 10 μM TGX-221, a Pik3cb-selective inhibitor, for 72 h and proliferation was measured as a percentage of control-treated (DMSO) cells. (*E*) Caspase 3/7 activation was measured in mOS cells after 16 h of exposure to indicated concentrations of PIK75, BEZ235, and GSK212645. (*F*) mOS cells were treated with BEZ235, GSK2126458, or PIK75 at the indicated drug concentrations. Cell extracts were analyzed by Western blotting with antibodies against phosphorylated AKT (Ser473), AKT, phosphorylated S6 (Ser235/236), S6, and GAPDH (loading control). Error bars are SEM, $n = 3$.

Table S1. Rare nonsilent deleterious mutations in candidate genes associated with OS

Candidate gene	Rationale	Mutation	Sample	LOH
RB1	Hereditary retinoblastoma	p.Q436K	MX04	Yes
TP53	Li Fraumeni Syndrome	p.T256P	SJ01	Yes
		p.C275Y	BZ22	Unk
		p.R290C	BZ36	Unk
		p.R337H	BZ39	Yes
		p.R337H	BZ34	Unk
		p.R337H	BZ15	Unk
		p.R342*	BZ24	Yes
WRN	Werner's syndrome	p.R1406*	SJD08	Unk
		p.S1292Y	BZ20	No
BLM	Bloom's syndrome	—	—	—
RECQL4	Rothmund-Thomson syndrome	—	—	—
SQSTM1	Paget's disease	p.A426V	BZ18	No
		p.K238E	SJD09	Unk
		p.K238E	MX02	No
		p.K238E	MX01	No
TNFRSF11A	Paget's disease	—	—	—
TNFRSF11B	Paget's disease	p.V281M	BZ29	No
VCP	Paget's disease	—	—	—
CSF1	Paget's disease susceptibility	—	—	—
OPTN	Paget's disease susceptibility	p.V161M	BZ07	No
TM7SF4	Paget's disease susceptibility	—	—	—
DCSTAMP	Paget's disease susceptibility	—	—	—
PML	Paget's disease susceptibility	p.R755H	SJ07	No
RIN3	Paget's disease susceptibility	p.R465Q	BZ03	Unk
		p.E241K	BZ10	No
GRM4	Osteosarcoma susceptibility	p.T303I	BZ11	No

Sixteen candidate genes were analyzed and 20 rare nonsilent variants were present after excluding missense variants classified as "benign" by PolyPhen2. For each germline mutated gene, the rationale, the specific mutation, and sample ID are listed. Genes with germline mutations were analyzed for somatic loss of heterozygosity (LOH) in tumor samples. Unk, unknown because of inability to determine LOH with data available.

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)
Dataset S3 (XLSX)
Dataset S4 (XLSX)
Dataset S5 (XLSX)
Dataset S6 (XLSX)
Dataset S7 (XLSX)
Dataset S8 (XLSX)
Dataset S9 (XLS)
Dataset S10 (XLS)
Dataset S11 (XLSX)
Dataset S12 (XLSX)
Dataset S13 (XLS)
Dataset S14 (XLSX)
Dataset S15 (XLS)