Recurrent *MSC*^E116K^ mutations in ALK-negative anaplastic large cell lymphoma

R.A. Luchtel, et al.

SUPPLEMENTAL MATERIAL

Table of Contents

SUPPLEMENTAL METHODS

DNA extraction

Tissue for sequencing was snap-frozen in OCT directly following surgical removal. Genomic

DNA was extracted from frozen sections by phenol-chloroform extraction and ethanol

precipitation as previously published.[1] For 22 of the 62 patients, paired normal DNA was

extracted from mononuclear cells isolated from peripheral blood samples uninvolved by tumor

under protocols conducted by the Molecular Epidemiology Resource of the University of

Iowa/Mayo Clinic Lymphoma Specialized Program of Research Excellence.[2]

Whole exome sequencing and variant calling

Tumor and paired normal genomic DNA was captured using SureSelect Human All Exon kit V2

or V4+UTR (Agilent Technologies) and sequenced as 101 bp x 2 paired-end reads at 3

samples/lane on a HiSeq 2000 (Illumina). Additional exome sequencing data reflecting

peripheral blood samples from 50 healthy patients without cancer or known family history of

genetic disease were obtained from the Mayo Clinic Biobank.[3] FASTQ files from DNA

sequencing were aligned to the human genome build GRCh37 using NovoAlign.[4] GATK was

used for processing aligned reads, including local realignment and base quality score

recalibration according to published Best Practices.[5-7]

 An inclusive set of variants for each sample was generated by taking the union of the

results of multiple variant callers (GATK, SNVmix,[8] JointSNVMix,[9] SomaticSniper,[10] and

MuTect[11]; Supplemental Figure 1). Somatic calling was used for paired samples. GATK

UnifiedGenotyper was used to gather all evidence for (backfill) each variant across all samples

(tumor, paired normal control, and healthy control). We utilized a variant filtering strategy to

leverage population allele frequencies, within-patient and healthy-patient controls, as well as technical features, to identify the variants with highest biological and technical confidence. Variants that were outside the capture regions, polymorphic (≥0.1% in ExAC[12]), within repetitive elements as identified using RepeatMasker[13] and UCSC annotation (https://genome.ucsc.edu/), or identified in paired germline or healthy control samples were excluded. Recurrent variants were defined as being identified in at least two tumor samples. Variants were quality controlled by manual review of read depths and fractions of supporting reads on a per-sample basis. Copy number alterations were evaluated using PatternCNV as previously published.[14]

Sanger sequencing and targeted resequencing

For Sanger sequencing of *MSC*[E116K], 10 ng genomic DNA from clinical tissue samples was amplified with KAPA HiFi HotStart ReadyMix (Kapa Biosystems) per the manufacturer's recommendations using the PCR primers 5'-CTCAGCCGCAGAGTGCAA-3' (forward) and 5'-GTGTCCAGCTTGGAGAGCTT-3' (reverse). The reverse primer was used as a sequencing primer. For targeted resequencing, DNA was amplified using the primers 5'-CTCCCGGCCAAGGGCTCA-3' (forward) and 5'-CTTTGCTCAGCACGCGCATC-3' (reverse) that included a 5' oligo tail containing the sequencing primer region of the Illumina NGS adapters. PCR products were purified using Agencourt AMPure (Beckman Coulter) and underwent a 2nd PCR reaction using index primers to incorporate a sample-specific barcode sequence with the remaining Illumina adapter sequence. After final AMPure purification, samples were pooled and sequenced on a MiSeq instrument (Illumina). Data were analyzed using a custom bioinformatics pipeline to detect variants with >5% mutant allele frequency.

Immunohistochemistry and fluorescence *in situ* hybridization (FISH)

Immunohistochemistry was performed on FFPE tissue samples using musculin (N-20, Santa

Cruz Biotechnology; 1:50), CD30 (JCM182, Leica; 1:250), MYC (Y69, Epitomics; 1:100), or

Ki67 (MIB-1, Dako; 1:300) primary antibodies using previously published methods.[15]

Immunohistochemistry was scored by decile as previously.[16] FISH was performed for *DUSP22*

and *TP63* rearrangements as previously published.[17]

MSC **cloning, mutagenesis, and siRNA**

The *MSC* gene was cloned into pCR2.1-TOPO (Thermo Fisher Scientific) using the primers 5'-

CCACCATGTACCCATACGACGTCCCAGACTACGCTTCCACGGGCTCGGTGAGT-3' and

5'-ATCCCATCAAGTGAGTTCCAGTCCGA-3' to introduce a 5' HA tag. The E116K

mutation was introduced by in situ mutagenesis QuickChange II Site-Directed Mutagenesis Kit

(Agilent). Both wild-type and mutant constructs were verified by Sanger sequencing and

shuttled into pcDNA3.1 (Thermo Fisher Scientific), pLEX (Open Biosystems), and pLVX-

ZsGreen (Clontech) vectors for functional experiments. pLEX-GFP was created by shuttling

GFP from pGIPZ (Open Biosystems) into pLEX using SnaBI and NotI restriction enzymes.

siRNA targeting *MSC* was purchased from Dharmacon (ON-TARGETplus SMARTpool).

siRNA targeting *IRF4* was purchased from Santa Cruz (sc-270125). AllStars negative control

siRNA was purchased from Qiagen.

Cell lines

Karpas 299 and HEK-293T cell lines were obtained from ATCC. Mac-1 and Mac-2A cells were derived by M.E.K. from a patient with multiple T-cell neoplasms.[18] FE-PD was provided by Dr. K. Pulford, Oxford, United Kingdom, with permission from Dr. A. Del Mistro, Padova, Italy. ALCL cells were maintained in RPMI and HEK-293T cells in DMEM, all with 10% FBS (Clontech) and 1% penicillin/streptomycin (Gibco). CD4$^+$ T cells were isolated from normal donor peripheral blood using a human CD4$^+$ T-cell enrichment kit (STEMCELL Technologies) and cultured in X-VIVO 15 medium (Lonza). Normal T cells were stimulated with anti-CD3/CD28 beads (Fisher Scientific).

Transfection and lentiviral transduction

T-cell lymphoma cell lines were transiently transfected with siRNA or pcDNA vectors by electroporation as previously described[19,20] or using the Amaxa Nucleofector Kit V (Amaxa) according to the manufacturer's instructions. Lentiviral particles were generated in HEK-293T cells and concentrated using Lenti-X Concentrator (Clontech). Karpas 299, FE-PD, and Mac-1 cells were transduced with pLEX lentivirus and puromycin was added to the medium 48 h after transduction to selectively culture transduced cells. Normal T cells were transduced with pLVX-ZsGreen lentivirus and sorted for GFP expression 72 h post-transduction. GFP-positive cells were cultured in the presence of anti-CD3/CD28 beads (Fisher Scientific) and assayed at 3 days (RNA) or 8 days (protein) post-transduction.

To evaluate normal T-cells expressing both *MSC*^{wt} and *MSC*^{E116K}, we inserted *MSC*^{E116K} into pLVX-mCherry (Clontech). We then co-transduced CD4$^+$ T-cells with pLVX-ZsGreen-*MSC*^{wt} and pLVX-mCherry-*MSC*^{E116K} or their corresponding empty vectors and flow-sorted the

resultant cells for ZsGreen and mCherry double-positivity. Sorted cells were activated with anti-CD3/CD28 beads as above for further experiments.

Western blotting and protein stability assay

Western blotting was performed as previously described.[19,20] Nuclear and cytoplasmic fractions were extracted using the NE-PER Nuclear and Cytoplasmic Extraction Kit (Thermo Scientific) following the manufacturer's instructions. Membranes were incubated with the following primary antibodies: MSC (N-20; Santa Cruz), HA (12CA5; Roche), β-Actin (AC-15; Novus), E2A (V-18; Santa Cruz), HEB (H-4; Santa Cruz), E2F2 (2302C4a; Abcam), MYC (Cell Signaling), GAPDH (14C10; Cell Signaling), α-tubulin (2144S, Cell Signaling), Lamin A (133A2, Cell Signaling), CD30 (EPR4102; Abcam), or IRF4 (M-17; Santa Cruz). Protein detection was performed on a LI-COR Odyssey CL_X using IRDye 680W and 800CW secondary antibodies (LI-COR).

To assess the stability of wild-type and mutant musculin protein, we treated Karpas 299 cells transduced with MSC^{wt} or MSC^{E116K} with cycloheximide (100 μg/mL; Thermo Fisher) and the caspase inhibitor Q-VD-OPh (5 μM; SM Biochemicals) and assessed proteins by Western blot as above at time intervals from 0 h to 6 h. Densitometry was performed and background-corrected on the LI-COR and values were normalized to β-actin.

Cell growth, cell cycle, and flow cytometry

For the competitive growth assay, Karpas 299 stably expressing GFP were mixed 1:1 with Karpas 299 cells stably transduced with empty pLEX vector, pLEX-HA-MSC^{wt}, or pLEX-HA-MSC^{E116K}. Cell mixtures were cultured for two weeks in the presence of puromycin, with or

without JQ1 at the indicated concentrations or DMSO. The percentage of GFP-positive cells was measured by flow cytometry at the start of the experiment and twice per week thereafter and normalized to control vector.

CD4$^+$ T cells transduced with pLVX-EF1α-IRES-ZsGreen1 empty vector or vectors containing *MSC*^{wt} or *MSC*^{E116K} were flow-sorted for GFP expression and re-activated with anti-CD3/CD28. MTS assays and cell cycle analyses were performed 3 days after re-activation. MTS assay (CellTiter 96® AQueous Non-Radioactive Cell Proliferation Assay, Promega) was performed as previously described.[21] Briefly, 3,000 cells were plated per well with or without drug, the MTS reagent was added, and colorimetric determination was performed at the indicated timepoints. For cell cycle analysis, cells were stained using the Click-iT™ EdU Alexa Fluor™ 647 Flow Cytometry Assay Kit (ThermoFisher, C10419) according to the manual with FxCycle Violet Stain (ThermoFisher, F10347) as the DNA stain. The cells were then analyzed on a BD FACSCanto™ II flow cytometer (BD Biosciences) and the data were analyzed using FlowJo v10 (Tree Star). GFP-positive cells were gated for cell cycle analysis.

For Karpas 299 cells, cell cycle analysis was performed 24 hours after co-transfection with GFP and vectors of interest. GFP-positive cells were gated for cell cycle analysis. Karpas 299 cells were stained 3 days after transfection for CD30 (BD Pharmingen 341635), c-MYC (R&D Systems, IC2585P), or HA (R&D Systems, IC6875R) using the BD Phosflow™ Protocol with Lyse/Fix Buffer (BD 558049) and Phosflow Perm Buffer III (BD 558050). Cells were analyzed on a BD FACSCanto™ II flow cytometer and data were analyzed with FlowJo v10. E2F2 expression was evaluated based on HA positivity.

RNA sequencing and gene expression profiling

Total RNA was extracted from transduced normal T cells and Karpas 299 cells using Qiagen

RNAEasy. RNA sequencing libraries were prepared using TruSeq mRNA v2 (Illumina) and

paired-end sequenced on an Illumina HiSeq as previously published.[21] RNAseq data were

analyzed using the MAP-RSeq pipeline.[22] Paired-end reads were aligned to the hg19 reference

genome using TopHat v2.1.0[23] and gene counts were estimated using HTSeq v0.5.3[24] based on

RefGene annotation. Gene counts were normalized to reads per kilobase per million mapped

reads (RPKM). Genes differentially expressed between MSC^{wt} and MSC^{E116K} were identified

using the edgeR package[25] with biological coefficient of variation (bcv) = 0.2, *P* value = 0.05,

and fold change = 3; raw gene counts were normalized using the trimmed mean of M-values

(TMM) option. For pathway analysis, a minimum read count threshold of 100 was applied, and

fold-change was used to rank genes for Gene Set Enrichment Analysis (GSEA).[26,27] Affymetrix

gene expression data from 31 ALCLs have been previously published and are publically

available (GEO accession, GSE118238).[16,28] Analysis of differences in gene expression between

mutated and unmutated ALCLs was carried out using multiple comparison (Bonferroni)

correction.

Chromatin immunoprecipitation sequencing (ChIP-seq)

ChIP-seq for HA-tagged musculin was performed on cells expressing empty vector, HA-MSC^{wt},

or HA-MSC^{E116K}. ChIP-seq for IRF4 was performed on normal $CD4^+$ T cells activated with anti-

CD3/CD28 beads and ALCL cell lines. Cells were cross-linked to 1% formaldehyde for 10 min

and quenched in 125 mM glycine for 5 min. ChIP-seq was performed as previously described[29]

with the following alterations: chromatin was digested with 1,000 gel units of MNase (NEB, Cat.

#M0247S) per 4×10^6 cells and subjected to 10 minutes of sonication (30 sec on / 30 sec off) in a

Bioruptor Twin (UCD-400; Diagenode). Immunoprecipitation was performed using anti-IRF4 (Santa Cruz; M-17) or anti-HA affinity matrix (Roche Diagnostics; Cat. #11815016001). ChIP-seq libraries were prepared from up to 10 ng ChIP and input DNA using the ThruPLEX® DNA-seq Kit V2 (Rubicon Genomics). Libraries were sequenced to 51 base pairs from both ends on an Illumina HiSeq 2000 instrument in the Mayo Clinic Center for Individualized Medicine Medical Genomics Facility.

ChIP-seq data were processed using the HiChIP pipeline.[30] Briefly, reads were aligned to the hg19 reference genome using the Burrows-Wheeler Alignment (BWA) tool v0.6.2[31] and only pairs with one or both ends being uniquely mapped were retained. Duplicates were removed using the Picard MarkDuplicates tool (http://broadinstitute.github.io/picard/). Peaks were identified using Model-based analysis of ChIP-Seq (MACS) v2.0.10 at a false discovery rate (FDR) of 1% and at least twofold enrichment over input.[32] Peaks were assigned to genes based on RefGenes if the transcriptional start site (TSS) was within 10 kb using the intersectBed command from BEDtools.[33] To identify sites showing differential musculin or IRF4 binding, peaks were merged if they showed overlap by at least 1 bp. For each merged peak, raw read counts were estimated from the unique alignments (after duplicate removal) using the intersectBed command and normalized to 10 million uniquely mapped reads. Fold-change was calculated using the normalized read counts and a fold-change cutoff of ≥ 2 was used to infer differential binding sites. To plot the musculin or IRF4 signals over RefGene TSS, we estimated the read density (RPM, reads per million) in each of the 100-bp non-overlapping bins over the TSS $\pm$ 5kb region using the ngs.plot tool v2.02.[34] To identify DNA binding motifs within musculin peaks, the 100-bp sequence centered on the peak center was extracted. Motifs were identified using MEME Suite v4.11.1[35], at the parameter settings of "-dna -mod zoops -nmotifs

10 -minw 7 -maxw 20 -revcomp". We used the FIMO tool[36] in MEME Suite to scan the peak center ± 50bp sequences for matches ($P \leq$ 1e-04) against the CAGCTG motif position weight matrix extracted from the MEME output.

ChIP was performed as above to determine binding of E2F2 to the human *TNFRSF8* (CD30) gene using an HA antibody on HA-E2F2 transfected Karpas 299 cells. Binding was assessed by quantitative PCR and analyzed with the ΔΔCt method using 1% of the input DNA as a normalization control. Two sets of primers were used targeting sequences 77 bp upstream (forward, 5'-GCGTCTCCTAGTGTGCCTTT-3'; reverse, 5'-ACACTTAGCTACAAGCAGCG-3') and 989 bp downstream (forward, 5'-ACTGAGCACTTACCAGTGGC-3'; reverse, 5'-AACACGCCTGGTTACTTTGC-3') of the *TNFRSF8* transcription start site.

Co-immunoprecipitation and tandem mass spectrometry (MS/MS)

Co-immunoprecipitation of HA-tagged musculin was performed on Karpas 299 cells transfected with pcDNA3.1, pcDNA3.1-HA-*MSC*[wt], or pcDNA3.1-HA-*MSC*[E116K] using the Pierce HA-Tag Magnetic IP Kit (ThermoFisher Scientific). Cells were lysed and proteins were co-immunoprecipitated following the manufacturer's instructions. Proteins were eluted by boiling in sample buffer. Eluates were then purified by electrophoresis and the gel was fixed and stained with BioSafe Coomassie stain (BioRad). The middle cross-section of each lane was excised as a single segment and digested with trypsin. Peptides were extracted and analyzed by nano-flow liquid chromatography electrospray tandem mass spectrometry (nanoLC-ESI-MS/MS) using a Thermo Scientific Q-Exactive Plus Mass Spectrometer (Thermo Fisher Scientific) coupled to a Thermo Ultimate 3000 RSLCnano HPLC system. Tandem mass spectra (MS/MS) were extracted using msConvert version 3.0.9134 (ProteoWizard[37]) and analyzed using MyriMatch[38]

database search engine (version 2.1.138). MyriMatch was configured to match the MS/MS

against SwissProt human protein sequence database (downloaded 6/2013) supplemented with

common sample handling contaminant proteins. Reversed protein sequences were appended to

the database for estimating false discovery rates (FDRs). IDPicker[39] software (version 3.0.504)

filtered the peptide identifications at 0.5% FDR and proteins with at least two unique peptides

were considered for further analysis. The total number of MS/MS matching a protein were

considered as a semi-quantitative measure of its abundance.[40] Protein spectral counts were

normalized and ratios were computed for each pair of samples (i.e. wild-type vs. control or

mutant vs. control or mutant vs. wild-type). Computed ratios were standardized and converted to

P values. Proteins that had no spectral counts in the control sample and had *P* values ≤0.01 and

absolute ratios ≥2.0 in at least one pair described above were considered robust candidates and

reported.

Electrophoretic mobility shift assay (EMSA) and DNA affinity assay

For EMSA, MSC^{wt}, MSC^{E116K}, TCF3^{wt} (E47), TCF3^{E555K} (E47), and HEB proteins were

synthesized in vitro from pcDNA3-*MSC*^{wt}, pcDNA3-*MSC*^{E116K}, pcDNA3-E47 (Addgene plasmid

#16059 from Robert Benezra[41]), pcDNA3-E47^{E555K} (generated using QuikChange II Site-

Directed Mutagenesis Kit, Agilent), and pCellFree-TCF12 (Addgene plasmid #67113 from Kirill

Alexandrov[42]) using the TnT Quick Coupled Transcription/Translation System (Promega).

EMSA was performed using the Odyssey EMSA Buffer Kit (LI-COR). IRDye 700-labeled

oligonucleotides were designed for previously described musculin and E2A binding motifs as

follows: 5'-CGGCCGAC<u>CAGCTG</u>GAGATCCT-3' and 5'-

CGGCCGAC<u>CAGGTG</u>GAGATCCT- 3' (Integrated DNA Technologies).[43,44] Probes were

incubated with protein combinations indicated using the Odyssey Infrared EMSA Kit (LI-COR). Resulting complexes were analyzed by electrophoresis on a 5% acrylamide gel and imaged using the Odyssey CL_X Imaging System (LI-COR).

For the DNA affinity assay, biotin-labeled oligonucleotides with the identical nucleotide sequences as above (Integrated DNA Technologies) were bound to streptavidin-conjugated magnetic beads (Dynabeads M-280 Streptavidin; Invitrogen) and incubated with nuclear lysates derived from cells transduced wth pLEX, pLEX-*MSC*^{wt} or pLEX-*MSC*^{E116K} for 2 hours. The magnetic beads were then isolated and washed and bound proteins were eluted for detection by Western blotting.

E2F luciferase reporter assay

The Cignal Reporter E2F luciferase assay (Qiagen) was used to assess E2F activity. Karpas 299 cells transduced with pLEX, pLEX-*MSC*^{wt} or pLEX-*MSC*^{E116K} were transfected with Cignal E2F Reporter and luciferase activity was measured at 24h post-transfection using the Dual-Luciferase Reporter Assay System (Promega).

Statistical analysis

Statistical analyses were performed either using JMP Pro 10 (SAS Institute) or in the SPSS or R statistical environment. Statistical tests were used as noted. *P*-values ≤0.05 were considered statistically significant except where indicated.

Data sharing statement

For original data, please contact feldman.andrew@mayo.edu.

SUPPLEMENTAL TABLES

Supplemental Table 1. Clinicopathologic features of frozen tissue exome discovery set

Case	Age/Sex	WHO Subtype	Paired Normal
T17	46/M	ALCL, ALK negative	No
T19	39/M	ALCL, ALK negative	No
T20	18/F	ALCL, ALK positive	No
T22	73/M	ALCL, ALK positive	No
T23	39/F	ALCL, ALK negative	No
T27	64/M	ALCL, primary cutaneous	Yes
T29	59/F	ALCL, primary cutaneous	Yes
T30	74/M	ALCL, ALK negative	No
T31	51/M	ALCL, ALK negative	No
T32	42/M	ALCL, ALK negative	No
T33	68/F	ALCL, primary cutaneous	No
T34	49/F	ALCL, ALK negative	No
T35	58/M	ALCL, ALK negative	No
T36	80/M	ALCL, ALK negative	No
T37	17/M	ALCL, ALK positive	No
T39	12/F	ALCL, primary cutaneous	No
T40	54/F	ALCL, ALK negative	No
T41	49/M	ALCL, ALK positive	No
T43	61/M	ALCL, primary cutaneous	No
T44	6/F	ALCL, ALK positive	No
T45	65/F	ALCL, ALK negative	No
T47	69/M	PTCL, NOS	Yes
T49	67/M	PTCL, NOS	Yes
T51	37/M	ALCL, ALK positive	Yes
T53	59/M	ALCL, ALK negative	Yes
T55	72/M	PTCL, NOS	Yes
T58	71/M	PTCL, NOS	Yes
T60	22/M	ALCL, ALK positive	No
T62	13/M	ALCL, ALK positive	Yes
T64	28/F	ALCL, ALK positive	No
T65	53/F	ENKTL	No
T66	47/M	ENKTL	No
T67	49/F	PTCL, NOS	No
T68	65/M	ENKTL	No

T70	64/F	AITL	Yes
T71	16/M	ALCL, ALK positive	No
T73	75/M	AITL	Yes
T75	58/F	AITL	Yes
T76	83/F	ALCL, ALK negative	No
T81	42/F	AITL	Yes
T85	42/M	AITL	Yes
T87	67/F	AITL	Yes
T90	76/M	ALCL, ALK negative	Yes
T91	74/M	ALCL, primary cutaneous	No
T95	61/F	ENKTL	No
T99	60/M	PTCL, NOS	No
T100	59/M	PTCL, NOS	No
T106	48/M	PTCL, NOS	No
T109	66/M	PTCL, NOS	No
T112	50/M	EATL	Yes
T114	60/F	MF	Yes
T117	55/M	PTCL, NOS	No
T118	64/F	PTCL, NOS	No
T120	78/M	MF	Yes
T122	62/M	PTCL, NOS	No
T123	77/F	ALCL, ALK negative	No
T124	74/M	PTCL, NOS	No
T125	48/F	ALCL, ALK positive	No
T127	48/M	ALCL, ALK negative	Yes
T129	50/M	ALCL, ALK negative	Yes
T133	67/F	ALCL, ALK negative	Yes
T190	44/M	ALCL, ALK positive	No

AITL, angioimmunoblastic T-cell lymphoma; ALCL, anaplastic large cell lymphoma; ALK, anaplastic lymphoma kinase; EATL, enteropathy-associated T-cell lymphoma; ENKTL, extranodal NK/T-cell lymphoma, nasal type; F, female; M, male; MF, mycosis fungoides; PTCL, NOS, peripheral T-cell lymphoma, not otherwise specified; WHO, World Health Classification.

Supplemental Table 2. COSMIC* mutations affecting the conserved glutamic acid (E) residue of the ERXR motif of bHLH proteins

Gene	Site	Amino acid alteration	Primary site	Primary histology	Mutation ID	Mutation CDS	Additional reference
ARNT	98	p.E98Q	Kidney	Carcinoma	COSM1660634	c.292G>C	[45]
ARNT2	72	p.E72K	Stomach; large intestine	Carcinoma	COSM4057237	c.214G>A	[46]
ATOH7	49	p.E49K	Urinary tract	Carcinoma	COSM3790891	c.145G>A	
FERD3L	110	p.E110K	Skin	Carcinoma	COSM4542968	c.328G>A	[47]
FERD3L	110	p.E110Q	Esophagus	Carcinoma	COSM5430632	c.328G>C	
MAX	32	p.E32*	Lung	Carcinoma	COSM371744	c.94G>T	[48]
MITF	213	p.E213Q	Urinary tract	Carcinoma	COSM3775349	c.637G>C	
MYF6	102	p.E102K	Skin	Melanoma	COSM3465839	c.304G>A	
MYOD1	118	p.E118Q	Urinary tract	Carcinoma	COSM1297893	c.352G>C	
MYOD1	118	p.E118*	Lung	Carcinoma	COSM394903	c.352G>T	[48]
NEUROD1	110	p.E110G	Lung	Carcinoma	COSM402132	c.329A>G	[48]
NEUROD1	110	p.E110K	Skin	Melanoma	COSM4400389	c.328G>A	
NEUROD4	96	p.E96K	Lung	Carcinoma	COSM5246904	c.286G>A	
NEUROD6	103	p.E103K	Large intestine; ovary; skin	Carcinoma; melanoma	COSM1329868	c.307G>A	[49]
NEUROG3	92	p.E92K	Skin	Melanoma	COSM3440003	c.274G>A	
OLIG3	92	p.E92K	Endometrium	Carcinoma	COSM1073792	c.274G>A	
SOHLH1	62	p.E62K	Esophagus	Carcinoma	COSM1173467	c.184G>A	[50]
TCF12	610	p.E610K	CNS	Glioma	COSM4746332	c.1828G>A	[51]
TCF4	573	p.E573K	Hematopoietic; CNS	Hematopoietic; Glioma	COSM1736660	c.1717G>A	[52,53]
TCF4	573	p.E573Q	Pancreas	Carcinoma	COSM5000304	c.1717G>C	[54]
TCF4	573	p.E573*	Lung	Carcinoma	COSM563718	c.1717G>T	
TFAP4	57	p.E57K	Skin	Melanoma	COSM3509614	c.169G>A	

*https://cancer.sanger.ac.uk/cosmic.

bHLH, basic helix-loop-helix; CNS, central nervous system.

Supplemental Table 3. T-NHL samples studied for MSC^{E116K}

WHO Diagnosis	Genetic Subtype					
	ALK	**DUSP22**	**-/-/-**	**TP63**	**N/A***	**Total**
ALCL, ALK positive	46	0	0	0	-	46
ALCL, ALK negative	0	28	46	8	5	87
ALCL, cutaneous	0	12	19	1	-	32
AITL	-	-	-	-	22	22
PTCL, NOS	-	-	-	-	40	40
Other	-	-	-	-	11	11
Total	46	40	65	9	78	238
Age, mean						56 years
Age, range						6 – 96 years
M:F ratio						1.7

*N/A, not applicable. Includes subtypes other than ALCL and 5 cases of ALCL for which genetic subtype could not be determined.

AITL, angioimmunoblastic T-cell lymphoma; ALCL, anaplastic large cell lymphoma; PTCL, NOS, peripheral T-cell lymphoma, not otherwise specified; T-NHL, T-cell non-Hodgkin lymphoma; WHO, World Health Organization.

Supplemental Table 4. T-NHL samples studied musculin expression by immunohistochemistry

WHO Diagnosis	Genetic Subtype					
	ALK	**DUSP22**	**-/-/-**	**TP63**	**N/A***	**Total**
ALCL, ALK positive	25	0	0	0	-	25
ALCL, ALK negative	0	7	27	3	2	39
ALCL, cutaneous	0	7	10	0	4	21
AITL	-	-	-	-	22	22
PTCL, NOS	-	-	-	-	32	32
Other	-	-	-	-	8	8
Total	25	14	37	3	68	147
Age, mean						58 years
Age, range						6 – 90 years
M:F ratio						1.9

*N/A, not applicable. Includes subtypes other than ALCL and 6 cases of ALCL for which genetic subtype could not be determined.

AITL, angioimmunoblastic T-cell lymphoma; ALCL, anaplastic large cell lymphoma; PTCL, NOS, peripheral T-cell lymphoma, not otherwise specified; T-NHL, T-cell non-Hodgkin lymphoma; WHO, World Health Organization.

Supplemental Table 5. Proteins co-immunoprecipitated with wild-type and/or mutant musculin

Gene Symbol	Distinct Peptides	Distinct Matches	Filtered Spectra	Description	Normalized Arithmetic Average Counts*			Log$_2$ Ratio			*P* value		
					M	W	C	M vs C	W vs C	M vs W	M vs C	W vs C	M vs W
HK1	24	26	57	hexokinase 1	1	5	0	1.00	2.58	-1.58	0.161	0.001	0.039
GSTP1	17	25	83	glutathione S-transferase pi 1	2	7	1	0.58	2.00	-1.42	0.395	0.014	0.073
RUNX1	16	19	36	runt related transcription factor 1	0	5	0	0.00	2.58	-2.58	0.647	0.001	0.000
UBTF	12	12	28	upstream binding transcription factor	0	2	5	-2.58	-1.00	-1.58	0.000	0.157	0.039
TCF4	24	30	87	transcription factor 4	5	9	0	2.58	3.32	-0.74	0.000	0.000	0.406
AP1B1	28	30	49	adaptor related protein complex 1 subunit beta 1	5	2	0	2.58	1.58	1.00	0.000	0.060	0.113
UBP1	11	13	30	upstream binding protein 1	0	5	0	0.00	2.58	-2.58	0.647	0.001	0.000
TCF3	37	47	122	transcription factor 3	2	15	0	1.58	4.00	-2.42	0.021	0.000	0.001
PDCD6IP	33	43	93	programmed cell death 6 interacting protein	7	6	0	3.00	2.81	0.19	0.000	0.000	0.564
BCAR1	31	43	118	BCAR1, Cas family scaffold protein	2	27	8	-1.58	1.64	-3.22	0.027	0.051	0.000
TCF4	19	25	73	transcription factor 4	4	7	0	2.32	3.00	-0.68	0.000	0.000	0.445
SH2D3C	15	15	36	SH2 domain containing 3C	0	5	1	-1.00	1.58	-2.58	0.186	0.060	0.000
JUNB	12	16	52	JunB proto-oncogene, AP-1 transcription factor subunit	1	5	8	-2.17	-0.58	-1.58	0.002	0.346	0.039
SPTAN1	59	63	115	spectrin alpha, non-erythrocytic 1	2	7	1	0.58	2.00	-1.42	0.395	0.014	0.073
TCF3	40	51	143	transcription factor 3	4	17	0	2.32	4.17	-1.85	0.000	0.000	0.013
TCF12	48	67	229	transcription factor 12	15	23	0	4.00	4.58	-0.58	0.000	0.000	0.503
TPR	48	52	94	translocated promoter region, nuclear basket protein	0	6	3	-2.00	0.81	-2.81	0.004	0.349	0.000
UBA1	39	55	120	ubiquitin like modifier activating enzyme 1	2	5	0	1.58	2.58	-1.00	0.021	0.001	0.243
FASN	81	90	202	fatty acid synthase	8	16	3	1.17	2.09	-0.92	0.098	0.010	0.291
GZMB	22	48	212	granzyme B	31	38	5	2.42	2.70	-0.29	0.000	0.001	0.638
CLTC	61	75	164	clathrin heavy chain	7	12	1	2.00	2.70	-0.70	0.003	0.001	0.430

DOCK2	54	63	135	dedicator of cytokinesis 2	1	12	5	-1.58	1.12	-2.70	0.027	0.202	0.000
MARS	26	30	64	methionyl-tRNA synthetase	7	1	5	0.42	-1.58	2.00	0.500	0.028	0.001
MSC	16	32	116	musculin	6	17	0	2.81	4.17	-1.36	0.000	0.000	0.088
MVP	42	51	118	major vault protein	2	13	4	-0.74	1.49	-2.22	0.333	0.081	0.002
ACTR3	26	44	134	ARP3 actin related protein 3 homolog	6	20	4	0.49	2.07	-1.58	0.458	0.011	0.039
ACTR2	20	30	67	ARP2 actin related protein 2 homolog	2	7	1	0.58	2.00	-1.42	0.395	0.014	0.073
PLIN3	25	43	121	perilipin 3	8	2	1	2.17	0.58	1.58	0.001	0.458	0.012
MCM4	38	51	102	minichromosome maintenance complex component 4	3	8	1	1.00	2.17	-1.17	0.161	0.007	0.157
SND1	37	45	98	staphylococcal nuclease and tudor domain containing 1	2	9	1	0.58	2.32	-1.74	0.395	0.004	0.021
DHX30	39	42	98	DExH-box helicase 30	1	8	3	-1.00	1.17	-2.17	0.186	0.180	0.003
PRRC2C	39	40	88	proline rich coiled-coil 2C	5	1	9	-0.74	-2.32	1.58	0.333	0.001	0.012
EIF5B	28	31	59	eukaryotic translation initiation factor 5B	0	2	5	-2.58	-1.00	-1.58	0.000	0.157	0.039
DDX47	7	8	25	DEAD-box helicase 47	0	2	5	-2.58	-1.00	-1.58	0.000	0.157	0.039
DDX28	17	17	41	DEAD-box helicase 28	0	1	6	-2.81	-1.81	-1.00	0.000	0.012	0.243
PLEC	218	249	608	plectin	17	8	39	-1.15	-2.15	1.00	0.122	0.003	0.113

C, control; M, mutant (E116K); W, wild-type.

*Based on the results of 3 independent experiments.

Supplemental Table 6. Top GSEA pathways altered in CD3/CD28-stimulated normal CD4$^+$ T cells: MSC^{wt} vs. vector only

| Gene Set	Size	ES	NES	Nominal *P* Value	FDR Q Value	FWER *P* Value
HALLMARK_MYC_TARGETS_V2	57	-0.80122	-2.92922	0	0	0
HALLMARK_MYC_TARGETS_V1	199	-0.64298	-2.9243	0	0	0
HALLMARK_MTORC1_SIGNALING	189	-0.54864	-2.47751	0	0	0
HALLMARK_E2F_TARGETS	197	-0.50034	-2.25411	0	0	0
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	100	-0.51779	-2.1134	0	0	0
HALLMARK_OXIDATIVE_PHOSPHORYLATION	195	-0.46664	-2.11053	0	0	0
HALLMARK_ESTROGEN_RESPONSE_LATE	99	-0.44014	-1.81324	0	0.001719	0.013
HALLMARK_GLYCOLYSIS	130	-0.42485	-1.8019	0	0.001815	0.016
HALLMARK_G2M_CHECKPOINT	189	-0.40258	-1.80152	0	0.001614	0.016

ES, enrichment score; FDR, false discovery rate; FWER, familywise error rate; GSEA, gene set enrichment analysis; NES, normalized enrichment score.

Supplemental Table 7. Top GSEA pathways altered in CD3/CD28-stimulated normal CD4$^+$ T cells: MSC^{E116K} vs. MSC^{wt}

Gene Set	Size	ES	NES	Nominal *P* Value	FDR Q Value	FWER *P* Value
HALLMARK_MYC_TARGETS_V1	199	0.651127	2.785856	0	0	0
HALLMARK_MYC_TARGETS_V2	57	0.769985	2.732203	0	0	0
HALLMARK_MTORC1_SIGNALING	189	0.565114	2.396883	0	0	0
HALLMARK_E2F_TARGETS	197	0.533282	2.285488	0	0	0
HALLMARK_IL2_STAT5_SIGNALING	143	0.54192	2.212824	0	0	0
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	100	0.508958	1.984764	0	5.16E-04	0.002
HALLMARK_OXIDATIVE_PHOSPHORYLATION	195	0.462097	1.972647	0	4.43E-04	0.002
HALLMARK_G2M_CHECKPOINT	189	0.451521	1.93144	0	5.33E-04	0.003
HALLMARK_ESTROGEN_RESPONSE_LATE	99	0.490707	1.908234	0	8.87E-04	0.006
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	42	0.562034	1.834938	0.001736	0.001803	0.014
HALLMARK_KRAS_SIGNALING_UP	77	0.475999	1.758647	0.001818	0.003282	0.028
HALLMARK_XENOBIOTIC_METABOLISM	97	0.443442	1.74612	0	0.003552	0.033
HALLMARK_GLYCOLYSIS	130	0.425628	1.722956	0	0.004779	0.046

ES, enrichment score; FDR, false discovery rate; FWER, familywise error rate; GSEA, gene set enrichment analysis; NES, normalized enrichment score.

Supplemental Table 8. T-NHL samples studied for MYC expression by

immunohistochemistry

| WHO Diagnosis | MSC^{E116K} | | |
	Yes	No	Total
ALCL, ALK positive	0	17	17
ALCL, ALK negative	7	23	30
ALCL, cutaneous	1	13	14
AITL	0	8	8
PTCL, NOS	0	11	11
ENKTL	0	2	2
MF	0	1	1
Total	8	75	83
Age, mean			54 years
Age, range			6 – 87 years
M:F ratio			1.8

AITL, angioimmunoblastic T-cell lymphoma; ALCL, anaplastic large cell lymphoma; ENKTL, extranodal NK/T-cell lymphoma, nasal type; FFPE, formalin-fixed, paraffin-embedded; MF, mycosis fungoides; PTCL, NOS, peripheral T-cell lymphoma, not otherwise specified; T-NHL, T-cell non-Hodgkin lymphoma; WHO, World Health Organization.

SUPPLEMENTAL FIGURES
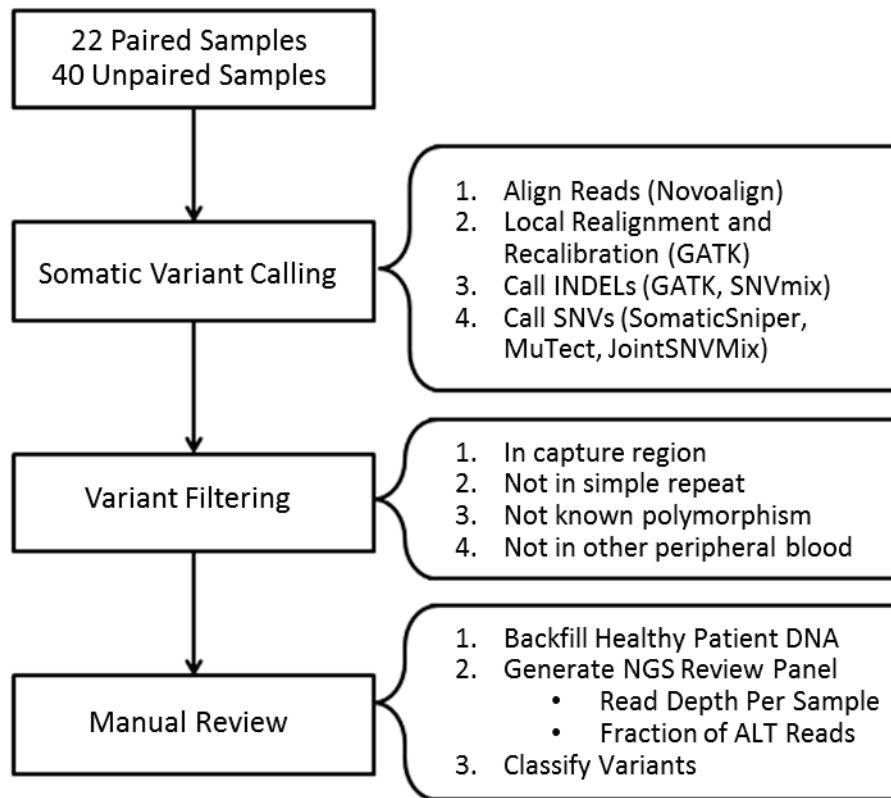
Supplemental Figure 1. Workflow used for variant calling and prioritization.
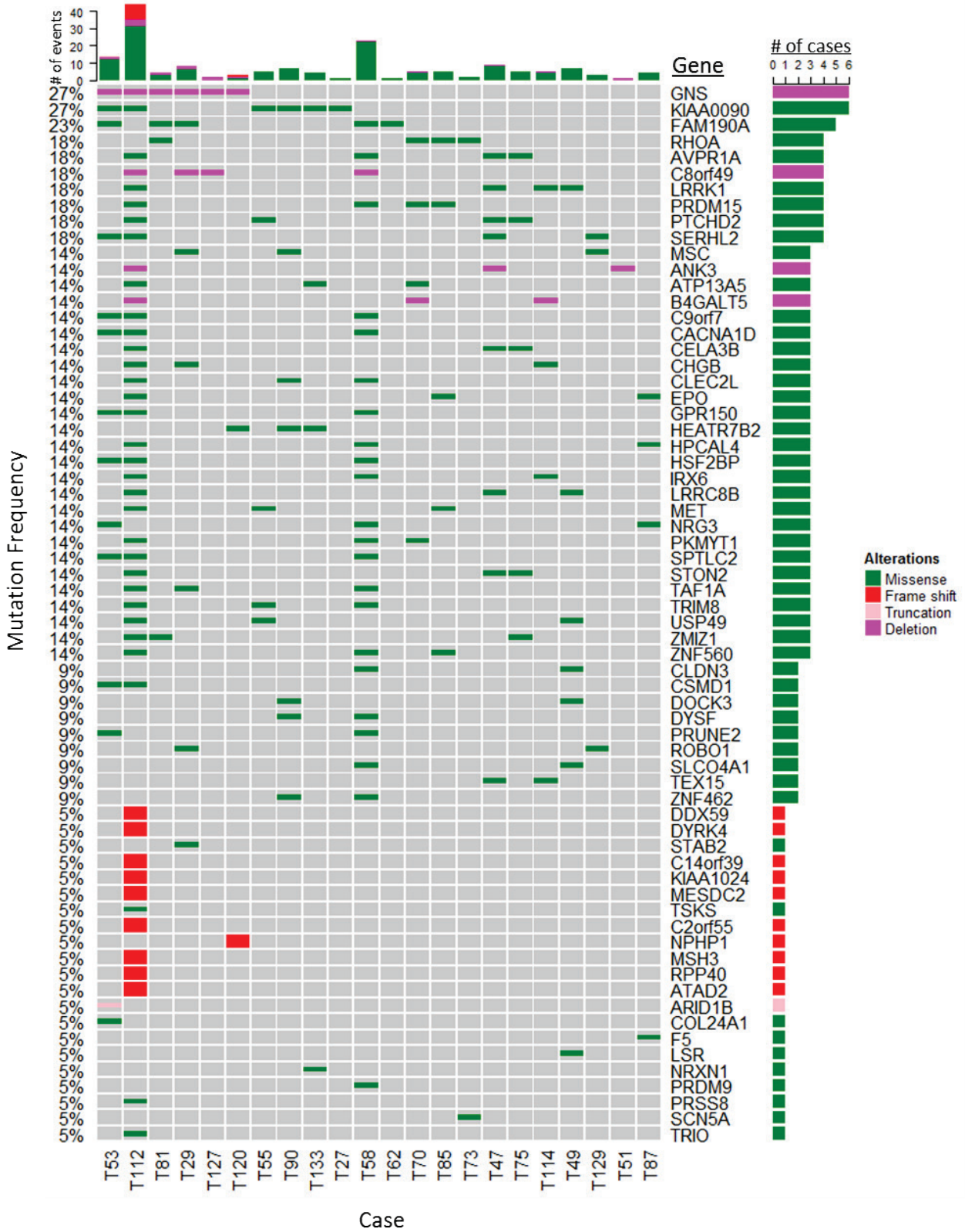
Supplemental Figure 2. Gene-level analysis of somatic variants in paired exome data.
Somatic variants were called in 22 paired tumor-normal samples. Clinicopathologic features are

shown in Supplemental Table 1. In addition to the *RHOA* and *MSC* mutations discussed in the main text, other genes previously reported to be mutated in T-cell lymphomas are noted, such as the chromatin-modifying gene *ARID1B*.[55] Interestingly, the case with the highest mutation burden, T112, an enteropathy-associated T-cell lymphoma, had a frameshift mutation of *MSH3*, c.1148delA, an event also identified in an additional unpaired sample (Table 1). *MSH3* (MutS Homolog 3) encodes a component of the post-replicative DNA mismatch repair machinery. c.1148delA is a known somatic cancer mutation resulting in a premature stop codon and microsatellite instability,[56] but previously unreported in lymphoma. Of note, mutations leading to mismatch repair deficiency are associated with an increased somatic mutation burden and may predict response to immunotherapeutic approaches such as immune checkpoint blockade.[57]

A

MSC	109	RNAANAR**ER**A**R**MRVLSKAFSRLKTSLPWVPPDTKLSKLDTLRLASSYIAHL	159
HAND1	96	KGSGPKK**ER**R**R**TESINSAFAELRECIPNVPADTKLSKIKTLRLATSYIAYL	146
MYOD1	111	RKAATMR**ER**R**R**LSKVNEAFETLKRCT-SSNPNQRLPKVEILRNAIRYIEGL	160
TCF3/E47	548	RMANNAR**ER**V**R**VRDINEAFREL	569
TCF12	603	RMANNAR**ER**L**R**VRDINEAFKEL	624
TCF4	570	RMANNAR**ER**L**R**VRDINEAFKEL	591

B

TCF3 E555K; Chr19, 1612356 C>T

TCF3

Human	549	MANNAR**E**RVRVRDINEAFRELGRMCQMHLKSDKAQTKLLILQQAVQVILGLEQQ	602
Rat	547	VANNAR**E**RLRVRDINEAFKELGRMCQLHLSTEKPQTKLLILHQAVAVILSLEQQ	600
Mouse	551	VANNAR**E**RLRVRDINEAFKELGRMCQLHLSSEKPQTKLLILHQAVAVILSLEQQ	604
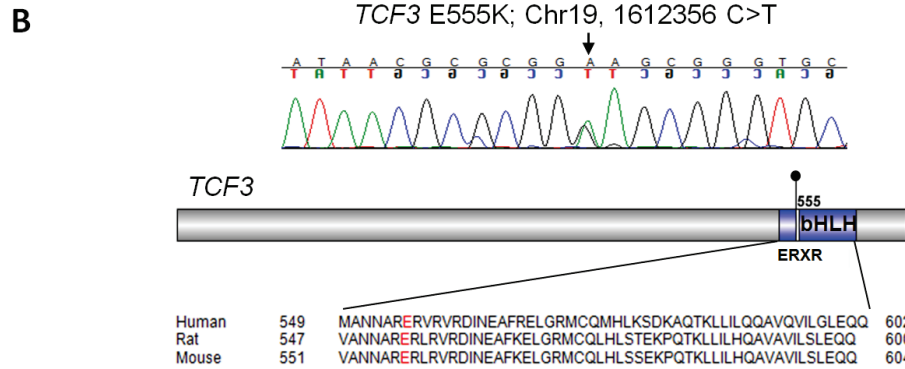
Supplemental Figure 3. Conservation of the ERXR motif in the basic domain of bHLH proteins. (A) Amino acid sequences around the ERXR motif of several bHLH proteins are shown with positions of the first and last residues. Conserved glutamic acid (E) and arginine (R) residues are shown in red. (B) *TCF3*[E555K] mutation affecting the glutamic acid residue of the ERXR motif of the bHLH protein TCF3 (E2A) identified in a case of ALK-negative ALCL. Sanger validation is shown.
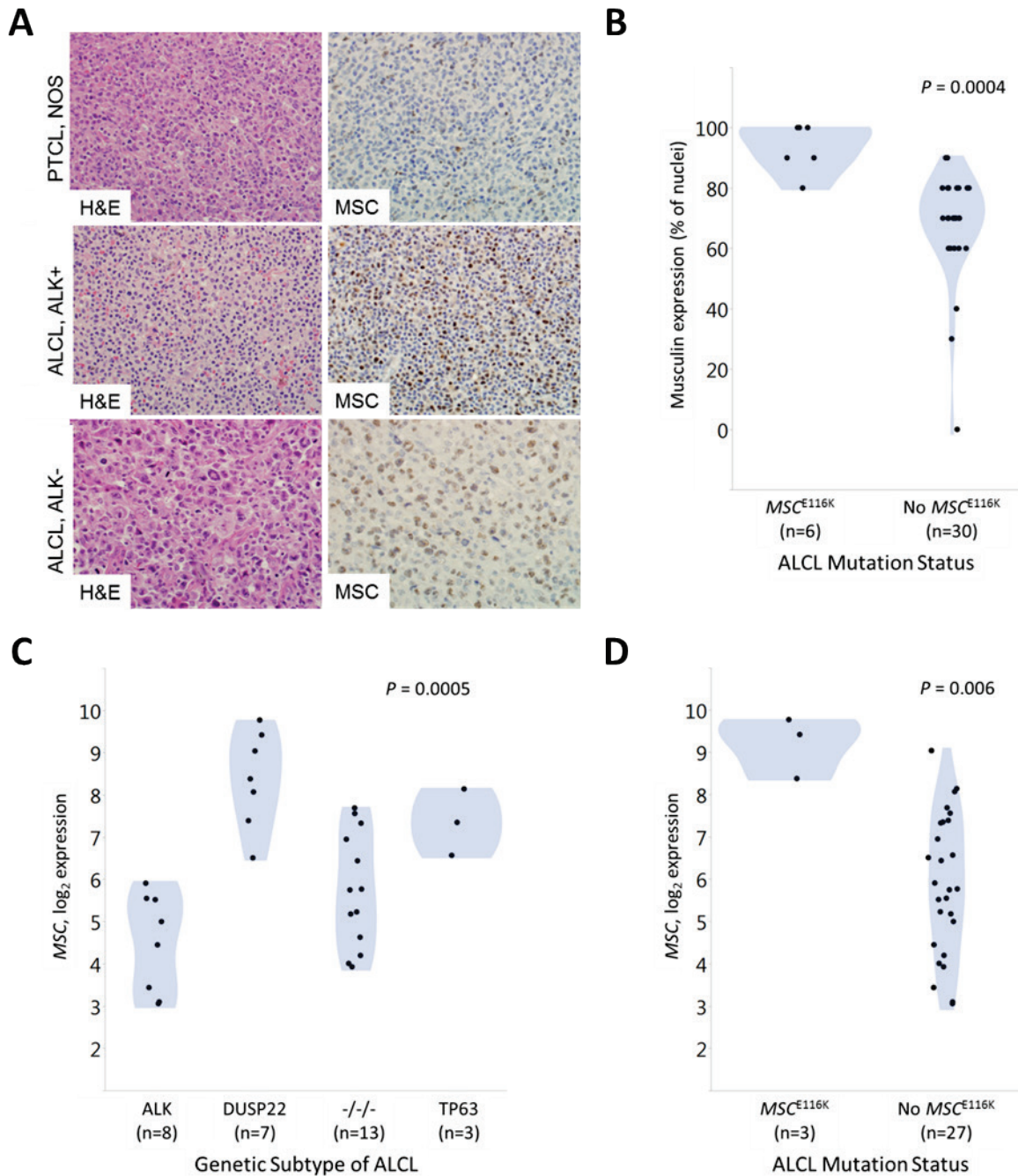
Supplemental Figure 4. Expression of musculin protein and the *MSC* gene in ALCL. (A) Immunohistochemical staining for musculin in T-NHLs without MSC^{E116K}. All images: original magnification, $\times 400$. (B) ALCLs with MSC^{E116K} consistently express musculin. (C) Gene expression data (N=31)[16,28] also show high levels of *MSC* expression in ALK-negative ALCLs with *DUSP22* rearrangement. (D) ALCLs with MSC^{E116K} consistently express high levels of *MSC*. N=30; *MSC* mutation status was unavailable in 1 case.

A

B
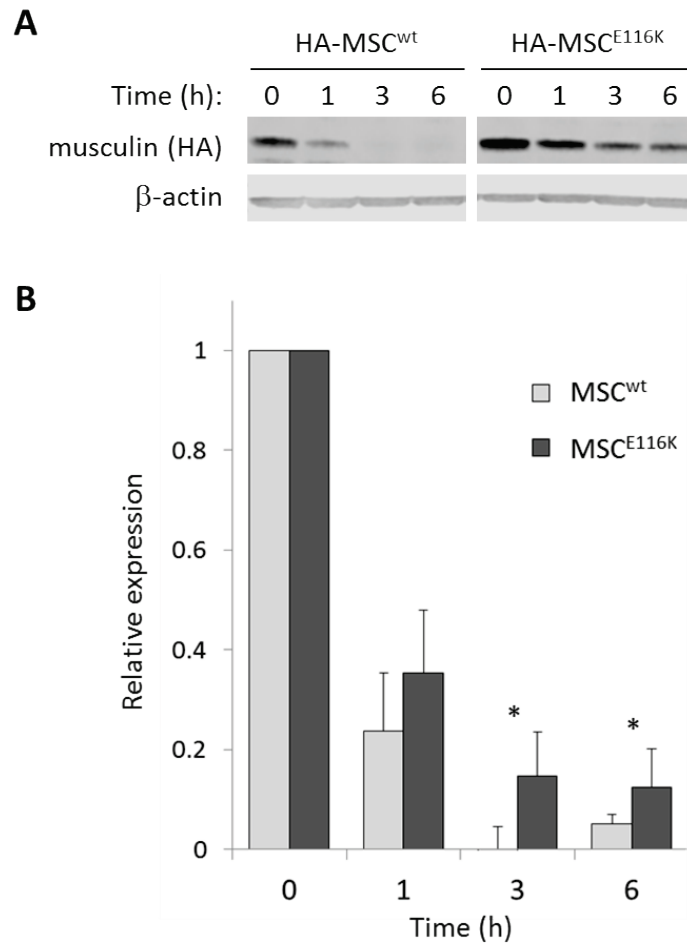
Supplemental Figure 5. Protein stability of MSC^wt and MSC^E116K. (A) Representative Western blot of lysates from Karpas 299 cells overexpressing HA-tagged MSC^wt and MSC^E116K at the timepoints indicated after treatment with cycloheximide and the caspase inhibitor Q-VD-OPh. (B) Background-corrected densitometry values normalized to β-actin and expressed relative to time 0. Data reflect 3 independent experiments. *, $P < 0.05$.
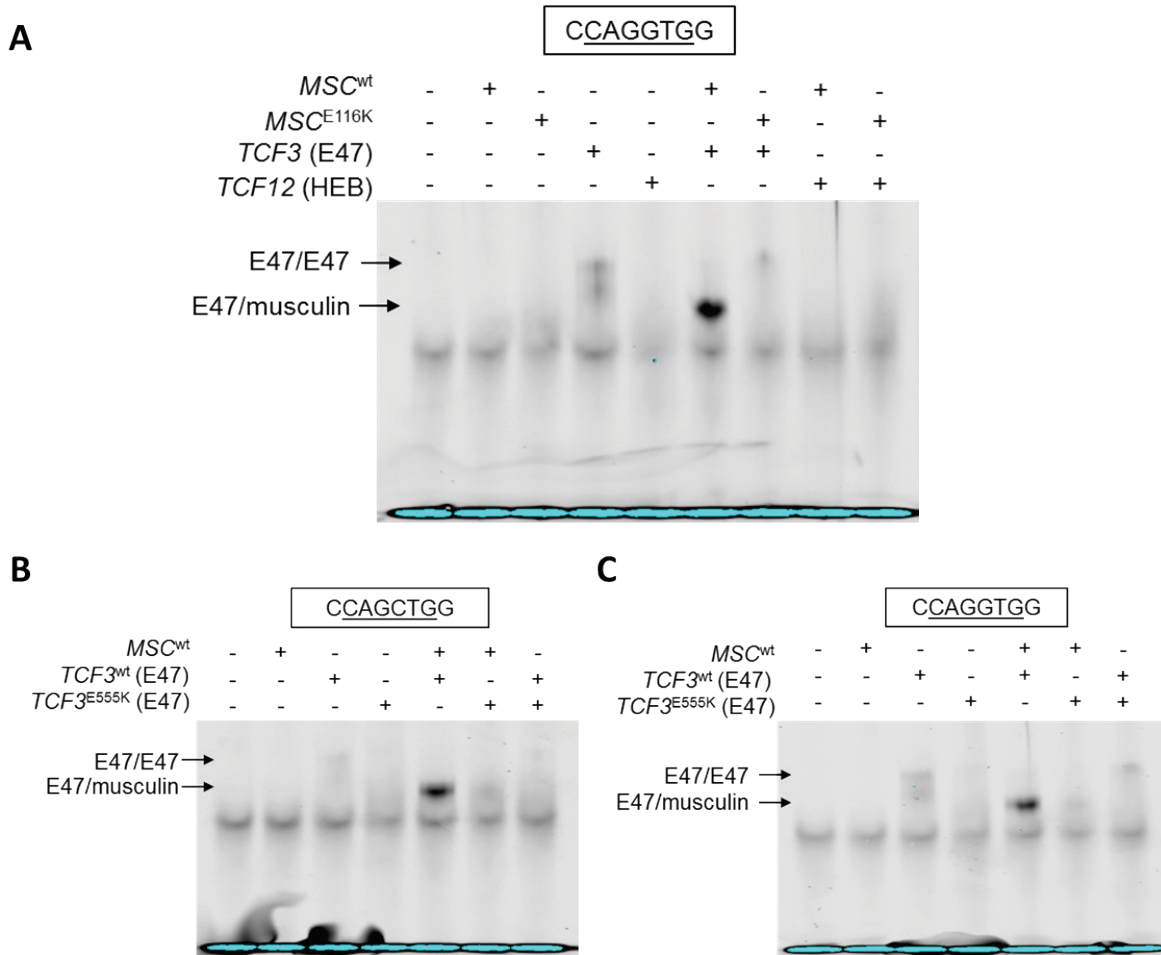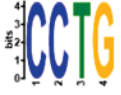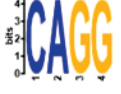
Supplemental Figure 6. DNA binding of wild-type and mutant bHLH proteins. (A) EMSA of proteins generated by *in vitro* translation showed binding of E47/musculin heterodimers to the E2A-binding E-box motif CAGGTG in the presence of MSC^{wt} but not MSC^{E116K}. E47 homodimer binding to CAGGTG was abolished in the presence of either MSC^{wt} or MSC^{E116K}. (B) *TCF3*^{E555K} diminished E47/musculin heterodimer binding to the musculin-binding motif CAGCTG. (C) *TCF3*^{E555K} diminished E47/musculin heterodimer binding to the E2A-binding motif CAGGTG.
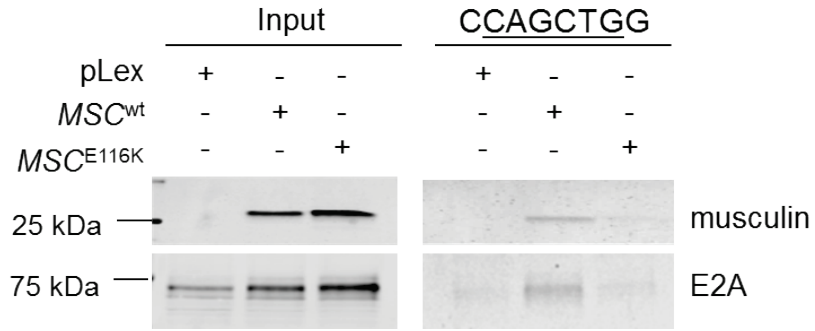
Supplemental Figure 7. DNA affinity assay in Karpas 299 cells. An oligonucleotide probe containing the musculin-binding motif CAGCTG pulled down both musculin and E2A in Karpas 299 cells transduced with *MSC*^{wt} but not *MSC*^{E116K}.
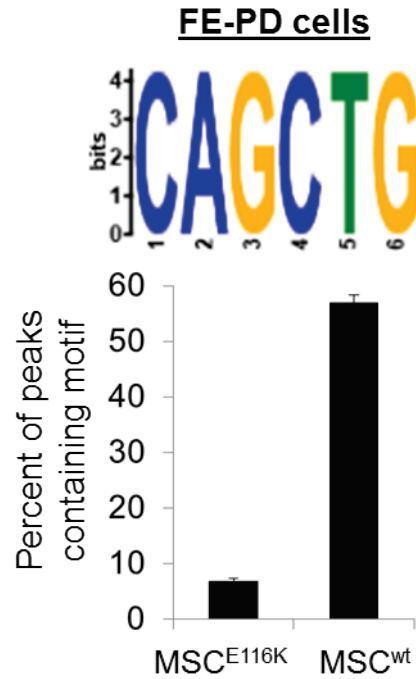
Supplemental Figure 8. ChIP-seq for HA-tagged MSC^{wt} and MSC^{E116K} in ALK-negative ALCL cells. ChIP-seq for HA-tagged MSC^{wt} and MSC^{E116K} in lentivirally transduced FE-PD cells demonstrated enrichment of the CAGCTG E-box motif for MSC^{wt} (E-value = 5.50E-109) but not for MSC^{E116K}. Data represent results of two independent experiments and are shown as means ± S.D.

Motif	E-value	Distribution	TF prediction
Karpas 299 MSC^wt			
CAGCTG	1.40E-176	Centered	musculin
TGACTCA	3.80E-12	Not centrally enriched	AP-1
TTCCT	1.10E-05	Not centrally enriched	BCL-6
Karpas 299 MSC^E116K			
TGACTCA	7.00E-34	Not centrally enriched	AP-1
CCTG	2.80E-15	Centered	TCF3/4
CAGG	2.20E-13	Centered	TCF3/4

**Supplemental Figure 9. ChIP-seq motif analysis for HA-tagged MSC^wt and MSC^E116K in Karpas 299 cells.**
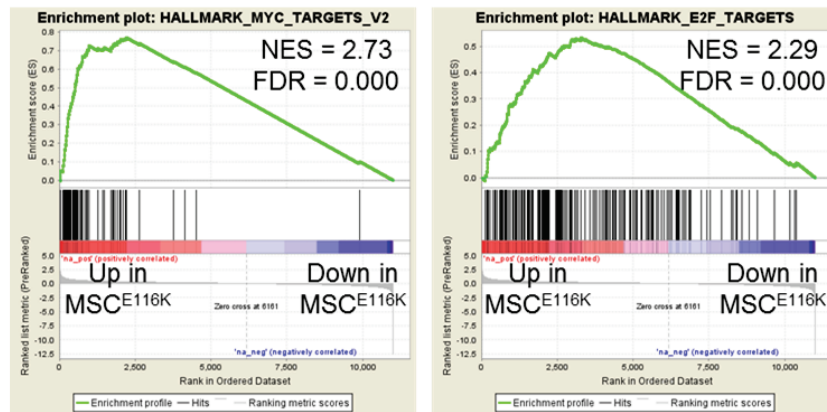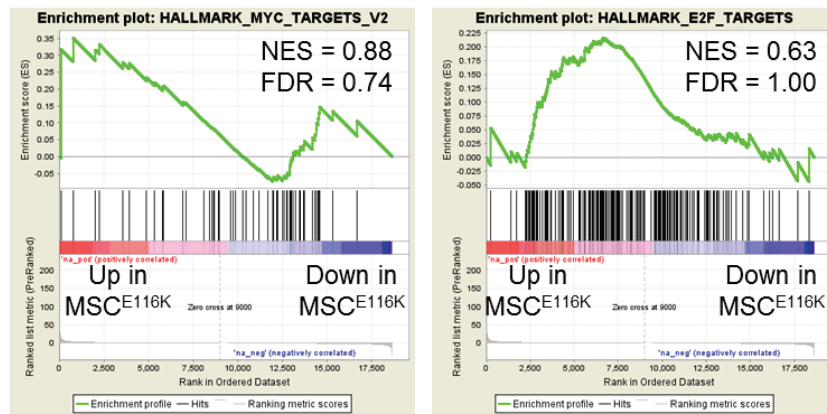
Supplemental Figure 10. Gene Set Enrichment Analysis of MSC^E116K^-transduced T cells.
(A) Gene Set Enrichment Analysis (GSEA) of genes differentially expressed in CD4+ T cells overexpressing *MSC*^E116K^ versus cells overexpressing *MSC*^wt^ showed enrichment of genes encoding MYC and E2F targets (the same pathways depleted in *MSC*^wt^- versus control-transduced cells; main text, Figure 3E). (B) GSEA of genes differentially expressed in *MSC*^E116K^- versus control-transduced cells showed neither enrichment nor depletion of these pathways.
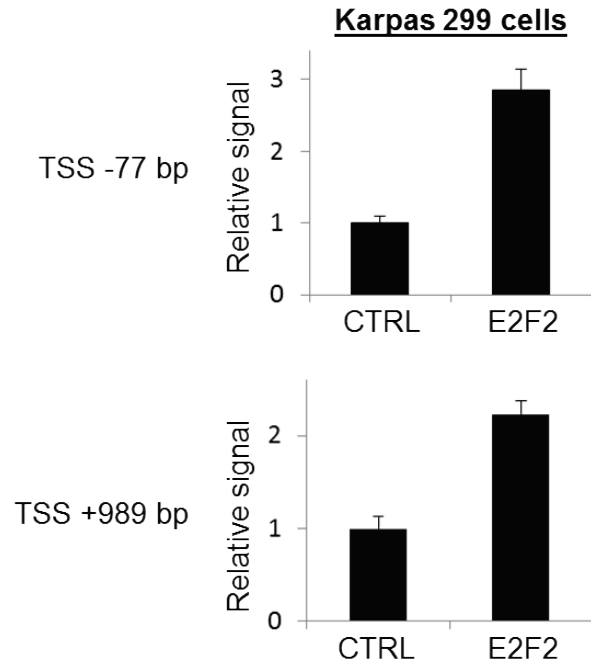
Karpas 299 cells

Supplemental Figure 11. Binding of E2F2 to *TNFRSF8* (CD30) DNA in human ALCL cells. Chromatin immunoprecipitation was performed to determine binding of HA-tagged E2F2 to *TNFRSF8* in Karpas 299 cells. Data reflect quantitative PCR results for 2 sites 77 bp upstream and 989 bp downstream of the *TNFRSF8* transcription start site (TSS), respectively.
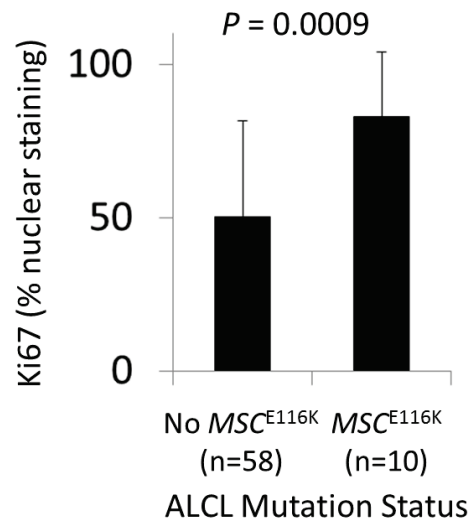
Supplemental Figure 12. Proliferative rate in ALCLs with or without MSC^{E116K}. Data reflect the percentage of tumor cells with nuclear staining detected by immunohistochemistry for Ki67 in ALCL tissue samples. Data are shown as means ± S.D.
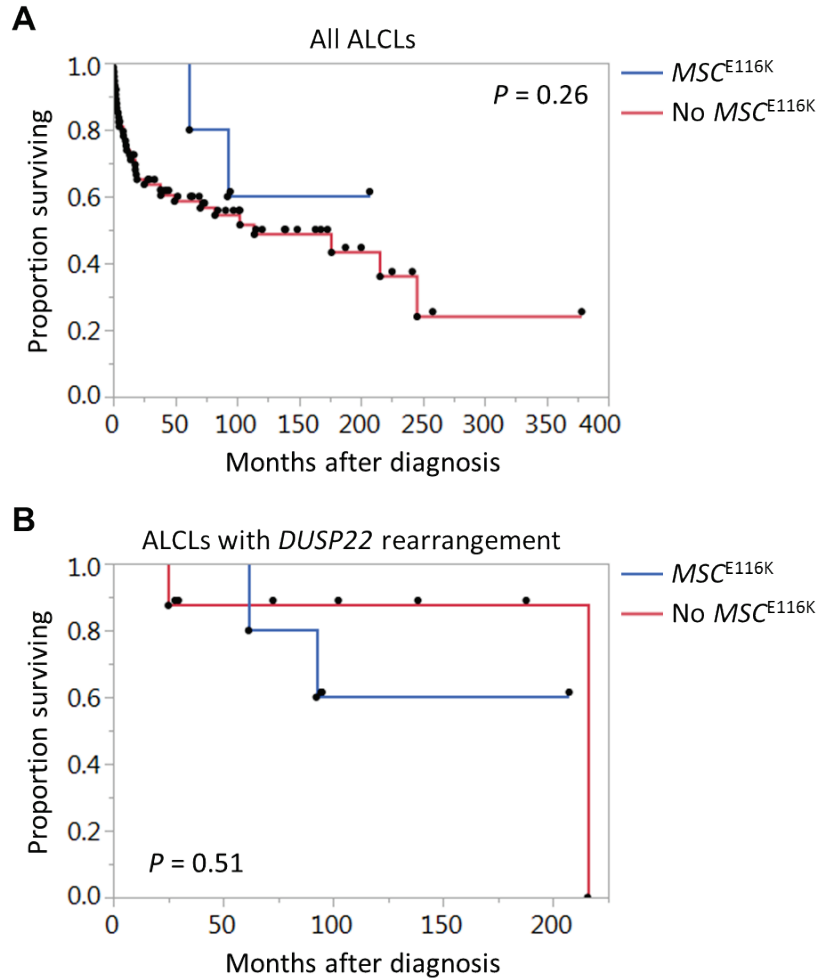
Supplemental Figure 13. Outcome analysis of ALCLs with and without MSC^{E116K}. (A) Among all patients with ALCL and available outcome data, no significant difference in overall survival was observed between those with MSC^{E116K} (n=9) and those without MSC^{E116K} (n=77; *P*=0.26, log-rank test). (B) Among patients with ALCL with *DUSP22* rearrangement and available outcome data, no significant difference in overall survival was observed between those with MSC^{E116K} (n=9) and those without MSC^{E116K} (n=10; *P*=0.51).

SUPPLEMENTAL REFERENCES

1.   Feldman AL, Dogan A, Smith DI, et al. Discovery of recurrent t(6;7)(p25.3;q32.3) translocations in ALK-negative anaplastic large cell lymphomas by massively parallel genomic sequencing. *Blood*. 2011;117(3):915-919.

2.   Cerhan JR, Link BK, Habermann TM, et al. Cohort Profile: The Lymphoma Specialized Program of Research Excellence (SPORE) Molecular Epidemiology Resource (MER) Cohort Study. *Int J Epidemiol*. 2017;46(6):1753-1754i.

3.   Olson JE, Ryu E, Johnson KJ, et al. The Mayo Clinic Biobank: a building block for individualized medicine. *Mayo Clin Proc*. 2013;88(9):952-962.

4.   Hercus C. Novoalign; 2009.

5.   McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303.

6.   DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-498.

7.   Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43:11 10 11-33.

8.   Goya R, Sun MG, Morin RD, et al. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*. 2010;26(6):730-736.

9.   Roth A, Ding J, Morin R, et al. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*. 2012;28(7):907-913.

10. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*. 2012;28(3):311-317.

11. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213-219.

12. Exome Aggregation Consortium (ExAC). Cambridge, MA (URL: http://exac.broadinstitute.org). Jan 2015.

13. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. 2009;Chapter 4:Unit 4 10.

14. Wang C, Evans JM, Bhagwate AV, et al. PatternCNV: a versatile tool for detecting copy number changes from exome sequencing data. *Bioinformatics*. 2014;30(18):2678-2680.

15. Feldman AL, Law ME, Inwards DJ, Dogan A, McClure RF, Macon WR. PAX5-positive T-cell anaplastic large cell lymphomas associated with extra copies of the PAX5 gene locus. *Mod Pathol*. 2010;23(4):593-602.

16. Luchtel RA, Dasari S, Oishi N, et al. Molecular profiling reveals immunogenic cues in anaplastic large cell lymphomas with DUSP22 rearrangements. *Blood*. 2018;132(13):1386-1398.

17. Parrilla Castellar ER, Jaffe ES, Said JW, et al. ALK-negative anaplastic large cell lymphoma is a genetically heterogeneous disease with widely disparate clinical outcomes. *Blood*. 2014;124(9):1473-1480.

18. Davis TH, Morton CC, Miller-Cassman R, Balk SP, Kadin ME. Hodgkin's disease, lymphomatoid papulosis, and cutaneous T-cell lymphoma derived from a common T-cell clone. *N Engl J Med*. 1992;326(17):1115-1122.

19.   Boddicker RL, Kip NS, Xing X, et al. The oncogenic transcription factor IRF4 is regulated by a novel CD30/NF-kappaB positive feedback loop in peripheral T-cell lymphoma. *Blood*. 2015;125(20):3118-3127.

20.   Boddicker RL, Razidlo GL, Dasari S, et al. Integrated mate-pair and RNA sequencing identifies novel, targetable gene fusions in peripheral T-cell lymphoma. *Blood*. 2016;128(9):1234-1245.

21.   Wang X, Dasari S, Nowakowski GS, et al. Retinoic acid receptor alpha drives cell cycle progression and is associated with increased sensitivity to retinoids in T-cell lymphoma. *Oncotarget*. 2017;8(16):26245-26255.

22.   Kalari KR, Nair AA, Bhavsar JD, et al. MAP-RSeq: Mayo Analysis Pipeline for RNA sequencing. *BMC Bioinformatics*. 2014;15:224.

23.   Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14(4):R36.

24.   Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166-169.

25.   Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26(1):139-140.

26.   Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550.

27.   Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34(3):267-273.

28.   Hu G, Dasari S, Asmann YW, et al. Targetable fusions of the FRK tyrosine kinase in ALK-negative anaplastic large cell lymphoma. *Leukemia*. 2018;32(2):565-569.

29.   Zhong J, Ye Z, Lenz SW, et al. Purification of nanogram-range immunoprecipitated DNA in ChIP-seq application. *BMC Genomics*. 2017;18(1):985.

30.   Yan H, Evans J, Kalmbach M, et al. HiChIP: a high-throughput pipeline for integrative analysis of ChIP-Seq data. *BMC Bioinformatics*. 2014;15:280.

31.   Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25(14):1754-1760.

32.   Zhang Y, Liu T, Meyer CA, et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 2008;9(9):R137.

33.   Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842.

34.   Shen L, Shao N, Liu X, Nestler E. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*. 2014;15:284.

35.   Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. 2009;37(Web Server issue):W202-208.

36.   Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. 2011;27(7):1017-1018.

37.   Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics*. 2008;24(21):2534-2536.

38.   Tabb DL, Fernando CG, Chambers MC. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res*. 2007;6(2):654-661.

39.   Ma ZQ, Dasari S, Chambers MC, et al. IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res*. 2009;8(8):3872-3881.

40.   Sadygov RG, Liu H, Yates JR. Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal Chem*. 2004;76(6):1664-1671.

41.   Jen Y, Weintraub H, Benezra R. Overexpression of Id protein inhibits the muscle differentiation program: in vivo association of Id with E2A proteins. *Genes Dev*. 1992;6(8):1466-1479.

42.   Gagoski D, Polinkovsky ME, Mureev S, et al. Performance benchmarking of four cell-free protein expression systems. *Biotechnol Bioeng*. 2016;113(2):292-300.

43.   Wendt H, Thomas RM, Ellenberger T. DNA-mediated folding and assembly of MyoD-E47 heterodimers. *J Biol Chem*. 1998;273(10):5735-5743.

44.   Macquarrie KL, Yao Z, Fong AP, Tapscott SJ. Genome-wide binding of the basic helix-loop-helix myogenic inhibitor musculin has substantial overlap with MyoD: implications for buffering activity. *Skelet Muscle*. 2013;3(1):26.

45.   Sato Y, Yoshizato T, Shiraishi Y, et al. Integrated molecular analysis of clear-cell renal cell carcinoma. *Nat Genet*. 2013;45(8):860-867.

46.   Giannakis M, Hodis E, Jasmine Mu X, et al. RNF43 is frequently mutated in colorectal and endometrial cancers. *Nat Genet*. 2014;46(12):1264-1266.

47.   Pickering CR, Zhou JH, Lee JJ, et al. Mutational landscape of aggressive cutaneous squamous cell carcinoma. *Clin Cancer Res*. 2014;20(24):6582-6592.

48.   Imielinski M, Berger AH, Hammerman PS, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*. 2012;150(6):1107-1120.

49.   Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330-337.

50.   Agrawal N, Jiao Y, Bettegowda C, et al. Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. *Cancer Discov*. 2012;2(10):899-905.

51.   Taylor KR, Mackay A, Truffaux N, et al. Recurrent activating ACVR1 mutations in diffuse intrinsic pontine glioma. *Nat Genet*. 2014;46(5):457-461.

52.   Nangalia J, Massie CE, Baxter EJ, et al. Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *N Engl J Med*. 2013;369(25):2391-2405.

53.   Killela PJ, Pirozzi CJ, Reitman ZJ, et al. The genetic landscape of anaplastic astrocytoma. *Oncotarget*. 2014;5(6):1452-1457.

54.   Witkiewicz AK, McMillan EA, Balaji U, et al. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat Commun*. 2015;6:6744.

55.   McKinney M, Moffitt AB, Gaulard P, et al. The Genetic Basis of Hepatosplenic T-cell Lymphoma. *Cancer Discov*. 2017;7(4):369-379.

56.   Risinger JI, Umar A, Boyd J, Berchuck A, Kunkel TA, Barrett JC. Mutation of MSH3 in endometrial cancer and evidence for its functional role in heteroduplex repair. *Nat Genet*. 1996;14(1):102-105.

57.   Le DT, Durham JN, Smith KN, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science*. 2017;357(6349):409-413.