SUPPLEMENTAL INFORMATION

Integrative genomic analyses identify lncRNA regulatory networks across pediatric leukemias and solid tumors

Apexa Modi[1,2], Gonzalo Lopez[1], Karina L. Conkrite[1], Chun Su[3], Tsz Ching Leung[1], Sathvik Ramanan[1], Elisabetta Manduchi[3], Matthew E. Johnson[3], Daphne Cheung[1], Samantha Gadd[4], Jinghui Zhang[5], Malcolm A. Smith[6], Jaime M. Guidry Auvil[7], Soheil Meshinchi[8], Elizabeth J. Perlman[4], Stephen P. Hunger[1,9,10], John M. Maris[1,9,10], Andrew D. Wells[3,11], Struan F.A. Grant[3,9,12,13], Sharon J. Diskin[1,9,10*]

[1] Division of Oncology and Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA.

[2] Genomics and Computational Biology Graduate Group, Biomedical Graduate Studies, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

[3] Center for Spatial and Functional Genomics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA.

[4] Department of Pathology and Laboratory Medicine, Ann & Robert H. Lurie Children's Hospital of Chicago, Robert H. Lurie Cancer Center, Northwestern University, Chicago, Illinois 60208, USA.

[5] Department of Computational Biology, St Jude Children's Research Hospital, Memphis, Tennessee 38105, USA.

[6] Cancer Therapy Evaluation Program, National Cancer Institute, Bethesda, Maryland 20892, USA.

[7] Office of Cancer Genomics, National Cancer Institute, Bethesda, Maryland 20892, USA.

[8] Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA.

[9] Department of Pediatrics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

[10] Abramson Family Cancer Research Institute, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

[11] Department of Pathology and Laboratory Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

[12] Department of Genetics, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

[13] Divisions of Human Genetics and Endocrinology & Diabetes, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, 19104, USA.

* Corresponding Author: diskin@email.chop.edu

Table of Contents

I. SUPPLEMENTARY MATERIALS AND METHODS

Gene/transcript mapping and quantification. To map reads to genes and quantify gene expression we ran StringTie 1.3.3 (RRID:SCR_016323) [1]. StringTie involves three steps, first quantifying expression of both known and novel gene transcripts using an annotation guided approach. We used the Gencode v19 gene annotation to guide gene detection.1) "stringtie bamfile -G gencode.v19.annotation_stringtie.gtf -B --rf -o out.gtf -A gene_abund.tab -C cov_refs.gtf -p 10. " In the second step, StringTie merges the gene annotation across all samples such that there is a uniform annotation for known and novel gene transcripts in one transcriptome gtf file. 2) "stringtie All_PanTARGET_PreMerge_StringTie_Files.txt --merge -G gencode.v19.annotation_stringtie.gtf -o StringTie_PanCancer_AllMergedTranscripts.gtf." Finally, StringTie is run again to quantify expression using the PanTarget transcriptome gtf file and de novo gene transcript detection is turned off. 3) "stringtie bamfile -G StringTie_PanCancer_AllMergedTranscripts.gtf -B -e --rf -o out.gtf -A gene_abund.tab -C cov_refs.gtf -p 10"

Comparison of pan-TARGET transcriptome with reference annotation. Novel transcripts were assigned as an isoform of a known gene based on exonic overlap (>50% by bp) with genes in either the GENCODE v19 (RRID:SCR_014966) or RefSeq v74 (RRID:SCR_003496) databases using custom Python scripts. Any remaining novel transcripts were assigned as novel genes (MSTRG_Merged.# or MSTRG.#) based on overlapping exon positions. Novel genes were further filtered based on read coverage, in that we required that at least one transcript for a novel gene have more than one exon with at least 5 reads in at least 20% of samples per cancer. High confidence novel genes were required to have at least 3 exons. Finally, for all transcripts (known and novel), to obtain gene level quantification, transcript FPKM and count values were summed to get a gene level value.

Prediction of novel gene coding potential and lncRNA gene annotation. We predicted coding potential of novel transcripts using the PLEK v1 (RRID:SCR_012132) algorithm tool [2]. PLEK uses a support vector machine (SVM) for a binary classification model to distinguish a lncRNA versus a coding mRNA. The features

used as input for the SVM are calibrated k-mer usage frequencies of a transcript's sequence. PLEK has previously been validated on RefSeq mRNAs and GENCODE lncRNAs (the main reference annotations used in our study) and has achieved >90% accuracy in predicting gene coding potential [2]. To further delineate lncRNAs, we removed any predicted novel non-coding transcripts that were < 200bp (sum of total exon length). We updated the gene type of GENCODE v19 genes with the gene type of genes that had matching gene names in GENCODE v29. Additionally we filtered out lncRNA genes that have been deprecated in Gencode v29. Finally, some lncRNA genes in Gencode v19, have both a lncRNA and small RNA transcript. For these 147 cases we did not include the small RNA transcript when summing gene transcripts to obtain gene level expression.

Assessing CNV impact on gene expression. To determine CNV impact on gene expression, we assessed differential expression of the gene in samples from the two groups (CNV yes or no) using Wilcoxon rank sum test ($p < 0.01$). Genes were considered to have evidence of differential expression due to copy number if the absolute value of the log2 fold change between the two groups was > 0.58 and $p < 0.05$.

Structural variant analysis. To obtain a comprehensive landscape of SVs we combined both the sequence junction and copy number read depth approaches to identify SVs, with co-localizing break points being orthogonally validated. Recurrence of SVs was considered based on overlap with genes from our pan-pediatric cancer transcriptome. Genomic overlap between SVs and genes was determined using the bedtools (RRID:SCR_006646) intersect tool (default parameters). Variants were assigned to genes based on if the sequence junction (left/right position) + 100 bp overlapped gene coordinates +/- 2.5kb. Genes were then ranked based on the number of unique samples per cancer with a SV breakpoint.

Gene signature analysis. We obtained a list of genes associated with the mesenchymal (MES) and adrenergic (ADRN) NBL cell types from GEO (GSE90805). We then used the GSVA (RRID:SCR_021058) R package[3] with the Poisson kernel (kcdf) parameter to assign a score per sample representing the total expression enrichment of genes associated with either the MES or ADRN cell types. We performed hierarchical clustering to divide NBL samples into three groups (MES, ADRN or mixed phenotype) based on expression of MES and

ADRN genes using the pheatmap R package and cutting the dendrogram at n=3. We correlated the MES and ADRN score with lncRNA expression across Stage 4 NBL TARGET cohort and GMKF cohort samples separately and identified lncRNAs as having significant correlation based on absolute value Spearman's rho > 0.6. These lncRNAs were then labeled as MES or ADRN based on significant correlation with either the MES or ADRN score. We next repeated score correlation with PCGs. We performed a guilt-by-association analysis assigning MES/ADRN PCGs and by association their correlated MES/ADRN lncRNAs (Spearman rho > 0.5) to pathways using Fisher exact test, FDR < 0.1 for gene sets in the gene ontology (GO) biological processes collection.

ChIP-seq data analysis. To determine which lncRNAs are regulated by transcription factors involved in the core regulatory circuitry (CRC) we utilized previously generated and analyzed histone and transcription factor ChIP-sequencing data for NBL. For NBL, we used peak files for our previously generated histone ChIP-seq data of: H3K27ac, H3K4me1, H3K4me3 for the BE(2)C cell line[4], available on GEO: GSE138315. We downloaded raw sequencing files for CRC transcription factor ChIP-seq data for MYCN, PHOX2B, HAND2, GATA3, TBX2, and ISL1 for the BE(2)C and KELLY cell lines from GEO: GSE94822 [5] and selected peaks with q-value < 0.001 for further analysis. We identified regions in the genome where at least 4/6 of the transcription factors overlapped. This was obtained using the homer (RRID:SCR_010881) mergePeaks tool: "mergePeaks -d 1000 -cobound 6 bed_file1… bed_file6" and the resulting coBoundBy4 output file.

Promoter-focused Capture C data analysis. Paired-end reads from each replicate were pre-processed using the HICUP (RRID:SCR_005569) pipeline (v0.5.9), with bowtie2 (RRID:SCR_016368) as aligner and hg19 as the reference genome. The unique ditags output from HiCUP were further processed by the chicagoTools bam2chicago.sh script before significant promoter interaction calling. Significant promoter interactions at 1-DpnII fragment resolution were called using CHiCAGO v1.1.8 (RRID:SCR_014941) with default parameters except for binsize set to 2500. Significant interactions at 4-DpnII fragment resolution were also called using CHiCAGO with artificial baitmap and rmap files in which DpnII fragments were concatenated *in silico* into 4 consecutive fragments using default parameters except for removeAdjacent set to False. Interactions with a CHiCAGO score > 5 in either 1-fragment or 4-fragment resolution were considered as significant interactions. The significant

interactions were finally converted to ibed format in which each line represents a physical interaction between fragments.

Identification of CRC transcription factor regulated genes. To identify genes regulated by the NBL CRC we considered CRC TF binding at both the gene's promoter and other regulatory region interacting with the gene's promoter. We first overlapped CRC regions using bedtools intersect with gene transcript promoter regions, which we defined as 3000bp upstream and downstream of the transcripts first exon. For NBL, we then utilized the promoter-focused Capture C data, inclusive of all interactions within 1Mb on the same chromosome, to identify genomic regions that were both bound by NBL CRC TFs and interacting with a gene's promoter. To determine this, we used bedtools intersect to determine overlap (minimum 1bp) between CRC bound loci with loci involved in chromatin interactions. From these regions, we determined which interacting regions corresponded with a lncRNA promoter region.

lncMod implementation: transcription factor target gene regulation. In the first part of the lncMod framework we had to first determine transcription factor target gene regulation specific to each cancer. Target genes here are defined as any protein coding or lncRNA gene and excludes pseudogenes and small RNAs. Given that ChIP-seq binding profiles for the majority of transcription factors were not available for tissues associated with each of these cancers we instead used transcription factor motif analysis as a proxy. We utilized motifs in the JASPAR database (RRID:SCR_003030) [6] and predictions of binding across the genome determined by FIMO (RRID:SCR_001783) and available in the UCSC genome database: http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2018/hg19/tsv/. For each transcript we determined potential regulatory transcription factors based on the presence of predicted binding motifs in the gene promoter region. Promoter regions were defined as regions 3000 bp upstream and downstream of the transcript's first exon. Next, we selected transcription factors based on their expression in each cancer and then performed linear regression considering the expression of the transcription factor and target gene specific to each cancer. We adjusted the false discovery rate due to multiple testing using the Benjamini-Hochberg method and selected TF-target gene pairs with significantly associated expression (adjusted p-value < 1e-5).

Identification of lncRNA modulators. To identify transcriptional perturbation, we first delineated genes (TF, target genes, or lncRNAs) that had differential expression defined by high expression variance (IQR > 1.5). For each differentially expressed lncRNA in each cancer we calculated the following, as has been done in previous studies [7-9]. For a given cancer and given lncRNA we sorted samples in the cancer based on the given lncRNAs expression (low to high). We then determined the correlation (Spearman's rho) between the expression of all transcription factor and target gene pairs previously identified in the given cancer. This correlation was calculated for the 25% of samples with the lowest lncRNA expression and separately for the 25% of samples with the highest expression for the given lncRNA. To ensure that we observed TF-target gene regulation we required that the correlation between the TF-target pair in either the low or high lncRNA expressing group was at least R>0.4. We only further evaluated the lncRNA TF-target gene triplet if the correlation difference between the low and high lncRNA expression group was R>0.45. To formally compare the correlation difference we first normalized the correlation using the Fisher r to z transformation. Then we calculated the rewiring score, z-statistic, as previously described [8], which is used to describe the degree of regulation change between the TF and target gene.

$$F(R) = \frac{1}{2}\ln\frac{1+R}{1-R}$$

$$rewire_{TF-gene} = P\left(|X| \leq \left|\frac{F(R_{high}) - F(R_{low})}{\sqrt{\frac{1.06}{n_{high}-3} + \frac{1.06}{n_{low}-3}}}\right|\right), X \sim N(0,1)$$

As a departure from what is described by Li et. al (lncMod method)[7], we used permutation analysis to robustly assess the significance of the rewire score in the context of multiple hypothesis testing as described by Sham et. al[10, 11]. We randomly shuffled target gene expression (TF-target gene pair labels) and calculated the rewire score P value across all TF-target gene pairs per given lncRNA. We kept the smallest observed P value and repeated the permutation 100 times. This empirical frequency distribution of the smallest P values was then compared to the P value in our real data to calculate an empirical adjusted P value (adj P value) as given by the

formula below, where r is the number of permutations where the smallest P value are less than our actual P value and n is the number of permutations.

$$adj\ Pvalue_j = \frac{1 + (\#\ permutations\ where\ q \leq p_j)}{1 + (\#\ permutations)}$$

The lncRNA-TF-target gene triplets, with adjusted p < 0.1 were considered significant. Datasets with smaller sample sizes had lower statistical power and thus fewer significant triplets. Triplets were then classified into three patterns based on correlation changes between the low and high expressing lncRNA group: increased correlation – enhanced, decreased correlation – attenuated, and inverted – positive to negative correlation and vice versa. We annotated lncRNA target genes as cancer genes based on if they were listed in the COSMIC database or a complied list from Chiu et. al[9].

NLF gene knockdown expression profiling. Total RNA was isolated from the NLF cell line 48 hours post treatment with siTBX2-AS1 and non-targeting control samples, siNTC, (three biological replicates per condition) and 1000 ng/sample was used as input for library preparation with the TruSeq Stranded mRNA Sample Prep Kit from Illumina (with Ribo-Zero treatment). RNA-seq libraries were sequenced on the Nextseq 500 at depth 10 million reads per sample minimum. Library prep and sequencing was performed by the Sidney Kimmel Cancer Center Genomics Facility of Thomas Jefferson University. Sample and read quality was assessed using FastQC and reads were aligned and mapped using the same methods as described above for TARGET cancer samples. Genes were retained if at least one sample had expression greater than 0 FPKM. To identify differentially expressed genes between siNTC and siTBX2-AS1 treated cells, we used the DESeq2 (RRID:SCR_015687) method with default parameters. Differentially expressed genes were annotated based on absolute value log fold change > 1.5 and Benjamini-Hochberg adjusted p-value < 0.1. Gene set enrichment analysis (GSEA, RRID:SCR_003199) was performed across samples using the MsigDB Hallmarks gene sets (RRID:SCR_016863) and significantly enriched gene sets with FDR q-val < 0.1 were retained. Up-stream co-regulators of differentially expressed genes were identified using default parameters from the iRegulon program part of the Cytoscape suite (RRID:SCR_003032). Raw and mapped data can be found through GEO at accession: GSE238166.

II. SUPPLEMENTARY TABLES

Supplementary Table S1: TARGET clinical sample and RNA-sequencing characteristics (xls file) Overview of RNA-seq samples selected for final cohort and information about the type of RNA-sequncing available per cancer.

Supplementary Table S2: Genomic loci for lncRNA and protein coding genes in this study (xls file) Genomic position, gene type, HUGO gene name, and chromosomal band for all lncRNA and protein coding genes considered in this study.

Supplementary Table S3: Number and types of genes expressed per cancer (xls file) The number of protein coding genes and known and novel lncRNAs expressed per cancer after filtering out lowly expressed genes.

Supplementary Table S4: Top 10 expressed lncRNAs across TARGET cancers and GTEx tissues (xls file) The top 10 expressed lncRNAs for TARGET cancers and GTEX tissues ranked based on highest proportion of expression over the total sum of all lncRNA expression (FPKM).

Supplementary Table S5: Tissue specificity index (tau score) annotation per gene (xls file) The tissue specifcity index, calculated as the tau score, per gene. The cancer with the highest expression of that gene is also listed.

Supplementary Table S6: Validation of tissue specific lncRNAs based on tau score analysis in alternate NBL datasets (xls file) The number of tissue specific lncRNAs per cancer with NBL pure data set, which includes samples that were 80-90% free of immune/stromal cell infiltration and number of tissue specific lncRNAs per cancer with GMKF NBL dataset.

Supplementary Table S7: Number of TARGET samples with WGS and SCNA events per cancer identified by GISTIC (xls file) The number of TARGET cancers with available WGS and the number of samples that had matched RNA-seq and the SCNA events identified by GISTIC per cancer including number of samples per event.

Supplementary Table S8: Differential expression of genes in samples with and without SCNA (xls file) List of all lncRNAs in SCNA regions and their log2 fold change and p-value (Wilcoxon rank sum test) between samples with and without SCNA.

Supplementary Table S9: Number of samples with SVs in lncRNAs (xls file)
List of all lncRNAs and number of samples of the specific cancer with a structural variant breakpoint in or near (+/- 2.5kb) the lncRNA and annotation of whether the lncRNA is located in an SCNA region of that cancer.

Supplementary Table S10: Statistics for input and output variables of lncMod analysis (xls file) Input parameters for lncMod analysis including the number of dysregulated lncRNAs, expressed transcription factors, and significant TF-target gene associations per cancer. The proportion and number of significantly dysregulated lncMod triplets compared to the number of all possible triplets per cancer.

Supplementary Table S11: Significantly dysregulated lncMod triplets (xls file) List of all significantly dysregulated lncMod triplets, which include a lncRNA modulator, transcription factor, and target gene.

Supplementary Table S12: lncRNA TF associations ranked by # target genes (xls file) The top 10 transcription factors, associated with a lncRNA modulator, ranked based on number of target genes associated with the transcription factor and lncRNA modulator per cancer.

Supplementary Table S13: lncRNAs associated with CRC of NBL (xls file) lncRNAs identified to be regulated by the CRC of NBL.

Supplementary Table S14: Differentially expressed lncRNAs between major subtypes in NBL (xls file) Differential expression analysis results comparing lncRNA expression between MYCN amplified and non-amplified NBL samples. Differential expression analysis results comparing lncRNA expression between the TAL1 subgroup and other T-ALL sample subgroups. lncRNAs that are differentially expressed in both NBL and T-ALL and annotation of whether the lncRNA is regulated by CRC transcription factors.

Supplementary Table S15: Data integration using multi-dimensional analysis to prioritize functional lncRNAs in each cancer (xls file) lncRNAs prioritized as likely functional based on association with a particular cancer through the results of various analyses used in this study.

Supplementary Table S16: Prioritized lncRNAs and predicted lncRNA target genes and pathways in NBL (xls file) lncRNAs prioritized as likely functional based on all analyses used in this study.

Supplementary Table S17: GSEA Analysis: MSigDB Hallmarks enriched across genes impacted by siTBX2-AS1 or siTBX2 treatment in NLF (xls file) Results from GSEA of genes significantly up- or down-regulated due to siTBX2-AS1 or siTBX2 treatment in the neuroblastoma cell line: NLF.

III. SUPPLEMENTARY FIGURES

Supplementary Fig. S1: Workflow for RNA-seq gene mapping and quantification (pdf file) (A) Workflow diagram showing how samples were processed using the StringTie program, which performs genes mapping and quantification. Custom scripts were then used for the following: Part I: Identified gene transcripts were assigned as a Gencode or Refseq gene, or as a novel gene. Part II: Gencode transcripts were further filtered including based on gene type. Part III: Novel genes were further filtered based on non-coding potential, length, number of transcripts and exons, and read coverage per exon. Part IV: Gene-level expression was considered as the sum of associated transcript expression. Table summarizing the number of transcripts and genes per gene type post filtering used in this study. **(B)** High confidence selected novel lncRNA genes are primarily intergenic or antisense. Sense-overlapping novel lncRNAs were considered low confidence and not-considered in this analysis. The majority of known lncRNAs in the Gencode database are either intergenic or antisense.

Supplementary Fig. S2: lncRNA expression varies across pediatric cancers (pdf file) (A) PCA of protein coding gene and lncRNA expression across pediatric cancers. **(B)** PCA showing unsupervised clustering of known lncRNA gene expression alone and **(C)** using novel lncRNA expressing. Each cancer is more closely clustered together individually than in the PCA including both protein coding genes and lncRNAs, suggesting that each cancer has distinct lncRNA expression, except for AML and B-ALL, which appear to have very similar lncRNA expression. Similar clustering was seen in PC3 and PC4 for all three gene type PCAs. **(D)** Average expression of protein coding genes, known, and novel lncRNAs in order of highest average expression. **(E)** Expression of a ubiquitously expressed lncRNA: *C17orf76-AS1* and its tau score: 0.296 is low. **(F)** Expression of the *MEG3* lncRNA is primarily in NBL and thus has a higher tau score of 0.986, (tau score >0.8 indicates tissue specificity). **(G)** The number of tissue specific lncRNAs across 12 adult cancers from TCGA. **(H)** Unsupervised clustering of the expression of the top 5 most TS lncRNAs (ranked by expression and tau score) in 12 adult cancers (total lncRNAs n=60).

Supplementary Fig. S3: Regions of somatic copy number aberration across cancers and genes dysregulated due to copy number (pdf file) (A) Plots of the frequency of copy number gain and loss across the genome for four pediatric cancer cohorts in this study. **(B)** lncRNA loci on chromosomes with copy number alterations across the pediatric cancers: NBL, WT, B-ALL, and AML. lncRNAs were evaluated to have differential expression due to copy number using the Wilcoxon rank sum test: highly differential: p-value < 0.05 and log |fold change| > 1.5 and moderately differential: p-value < 0.1 and log |fold change| > 1.0. Points are colored based on loci in an amplified or deleted region of the chromosome and if the lncRNA is highly or moderately differentially expressed.

Supplementary Fig. S4: Structural variants impact lncRNAs in pediatric cancers (pdf file) (A) The number and types of structural variants as annotated by Complete Genomics (CGI) per sample per cancer. **(B)** The number of samples with a structural variant in a lncRNA or protein coding gene (gene-sample pairs) vs number of lncRNA genes also found in copy number regions. **(C)** Ranking of genes with structural variants by the number of samples per cancer. **(D)** Expression of the top genes per cancer in samples with and without structural variant (NBL/WT: *MYCNOS*, **(E)** B-ALL: *CDKN2B-AS1*, *KIA00125*, and **(F)** AML: *MIR18A1HG*. **(G)** The number of genes with a structural variant found in 1-4 cancers.

Supplementary Fig. S5: Characteristics of lncMod analysis and results (pdf file) (A) Example of a dysregulated lncMod triplet in Wilm's Tumors (WT). Samples with high *RMST* expression also have higher expression correlation between *SOX2* and its target gene *SFRP2*. SOX2 regulation of *SFRP2* appears to be disrupted in samples with low *RMST* expression, as suggested by the low expression correlation of TF and target gene in these samples. **(B)** The number of modulator types (attenuate, enhance, or invert) associated with a lncRNA modulator based on its impact of a specific TF-target gene. **(C)** Number of target genes of the *H19*

lncRNA modulator in NBL that are enriched in MsigDB hallmarks gene sets. **(D)** Target genes of the *HOTAIRM1* lncRNA modulator in AML that are enriched in MsigDB hallmarks gene sets. **(E)** The top transcription factors, based on number of associated dysregulated target genes, impacted by the *GAS5* lncRNA modulator in T-ALL and the expression of *GAS5*, the transcription factor *E2F4* and its target genes in T-ALL samples with *GAS5* expression dysregulation. **(F)** The top transcription factors, based on number of associated dysregulated target genes, impacted by the *SNHG1* lncRNA modulator in T-ALL and the expression of *SNHG1*, the transcription factor *TP53* and its target genes in T-ALL samples with *SNHG1* expression dysregulation.

Supplementary Fig. S6: Defining mesenchymal vs adrenergic lncRNAs in two NBL cohorts (pdf file) (A) Heatmap of the expression of genes previously shown to be associated with either the MES or ADRN cell lineage in the TARGET NBL cohort. Samples were assigned into three groups using hierarchical clustering based on whether they had more expression of MES or ADRN genes. **(B)** *MYCN* expression in TARGET and GMKF samples predicted to have either ADRN or MES phenotype. *MYCN* expression is observed to be higher in ADRN samples. **(C)** Expression of lncRNAs with significant correlation (|r| >0.6) to the MES or ADRN score in the GMKF NBL cohort. lncRNAs were then correlated with protein coding genes on the same chromosome and subsequent gene set enrichment analysis was performed for MES and ADRN protein coding genes separately. Gene set enrichment results for each group are shown to the right of the heatmap. **(D)** Correlation of lncRNAs to the MES and ADRN score in the TARGET (x-axis) and GMKF (y-axis) NBL cohort respectively. Numbered points represent lncRNAs that had a significant correlation to the MES or ADRN score in both cohorts.

Supplementary Fig. S7: Validation of *TBX2-AS1* expression in NBL cell lines and impact on NBL cell growth (pdf file) (A) Expression of *TBX2-AS1* and *TBX2* across pediatric cancers. **(B)** Comparison of expression levels of *TBX2, TBX2-AS1, MYCN,* and *MYCNOS* in the NBL TARGET (un-stranded) and NBL GMKF (stranded) RNA-sequencing cohorts show high concordance. Pearson's correlation between the 14 common samples for these genes was r=0.979, 0.954, 0.983, and 0.877, respectively. **(C)** Correlation between *TBX2-AS1* and other genes previously shown to be regulated by TBX2. **(D)** Expression of *TBX2-AS1* and *TBX2* in 38 NBL cell lines. **(E)** RT-qPCR validation of *TBX2-AS1* and *TBX2* expression in 8 NBL cell lines. **(F)** RT-qPCR expression of *TBX2-AS1* and *TBX2* for NLF cell line treated with non-targeting control (siNTC) and four different siRNAs targeting *TBX2-AS1.* siTBX2-AS1-A is referred to as siTBX2-AS1 in the main figures. Three independent knockdown experiments are represented. **(G)** Western blot for TBX2 expression after independent treatment of multiple siRNAs targeting *TBX2-AS1* in NLF. **(H)** Representative image of NLF cell growth as measured by RT-ces assay following siRNA treatments. siPLK1 is a positive control. Cell index is normalized to time point when siRNA reagent is added at 24 hours post cell plating. All siTBX2-AS1 treatments resulted in significant growth inhibition. **(I)** RT-qPCR expression of *TBX2-AS1* and *TBX2* for SKNSH cell line treated with siNTC, siTBX2-AS1, and siTBX2. Three independent knockdown experiments are plotted, each plated in triplicate. **(J)** Representative Western blot for TBX2 expression after siTBX2 or siTBX2-AS1 treatment compared to NTC in SKNSH. Right panel: Protein quantification derived from ImageJ analysis of Western blots for three independent knockdown experiments. **(K)** Representative image of SKNSH cell growth as measured by RT-ces assay following siRNA treatments. Cell index is normalized to time point when siRNA reagent is added at 24 hours post cell plating. Both siTBX2-AS1 and siTBX2 show significant growth inhibition, consistent with results observed for the NLF cell line. **(L)** RT-qPCR expression of *TBX2-AS1* and *TBX2* for SKNSH cell line treated with siNTC and four different siRNAs targeting *TBX2-AS1.* Two independent knockdown experiments are represented. **(M)** Western blot for TBX2 expression after independent treatment of multiple siRNAs targeting *TBX2-AS1* in SKNSH. **(N)** Image of SKNSH cell growth as measured by RT-ces assay following independent treatments of four siRNAs targeting siTBX2-AS1. Cell index is normalized to time point when siRNA reagent is added at 24 hours post cell plating with index set to start at one.

Supplementary Fig. S8: RNA expression profiling following *TBX2-AS1* and *TBX2* knockdown in NLF (pdf file)

(A) Expression of *TBX2-AS1* and *TBX2* based on RNA sequencing analysis of NLF cells treated with either non-targeting control (siNTC) or siTBX2-AS1. *TBX2-AS1* expression is significantly reduced, but no significant change in *TBX2* expression is observed. **(B)** Expression of *TBX2-AS1* and *TBX2* based on RNA sequencing analysis of NLF cells treated with either non-targeting control (siNTC) or siTBX2. *TBX2* expression is significantly reduced, but no change in *TBX2-AS1* expression is observed. **(C)** Volcano plot showing genes with significant dysregulation (log fold change > 1.5 and -log10 p-value <0.1) observed in transcriptomic profiling of NLF cells treated with siTBX2-AS1 or **(D)** siTBX2. **(E)** The log2 fold change (FC) of significantly differentially expressed (DE) genes associated with siTBX2-AS1 or **(F)** siTBX2, plotted against their percentile rank and colored by gene expression quantile. **(G)** The number of DE genes overlapping between siTBX2-AS1 and siTBX2 and how correlated the log2 FC of DE genes are in each condition.

IV. REFERENCES

1. Pertea, M., et al., *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.* Nature biotechnology, 2015. **33**: p. 290-5.
2. Li, A., J. Zhang, and Z. Zhou, *PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme.* BMC Bioinformatics, 2014. **15**: p. 311.
3. Hanzelmann, S., R. Castelo, and J. Guinney, *GSVA: gene set variation analysis for microarray and RNA-seq data.* BMC Bioinformatics, 2013. **14**: p. 7.
4. Upton, K., et al., *Epigenomic profiling of neuroblastoma cell lines.* Sci Data, 2020. **7**(1): p. 116.
5. Durbin, A.D., et al., *Selective gene dependencies in MYCN-amplified neuroblastoma include the core transcriptional regulatory circuitry.* Nat Genet, 2018. **50**(9): p. 1240-1246.
6. Khan, A., et al., *JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework.* Nucleic Acids Res, 2018. **46**(D1): p. D260-D266.
7. Li, Y., et al., *Identification and characterization of lncRNA mediated transcriptional dysregulation dictates lncRNA roles in glioblastoma.* Oncotarget, 2016. **7**: p. 45027-45041.
8. Li, Y., et al., *LncMAP: Pan-cancer Atlas of long noncoding RNA-mediated transcriptional network perturbations.* Nucleic Acids Research, 2018. **46**: p. 1113-1123.
9. Chiu, H.S., et al., *Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context.* Cell Rep, 2018. **23**(1): p. 297-312 e12.
10. Sham, P.C. and S.M. Purcell, *Statistical power and significance testing in large-scale genetic studies.* Nat Rev Genet, 2014. **15**(5): p. 335-46.
11. Wagner, B.D., et al., *Permutation-based adjustments for the significance of partial regression coefficients in microarray data analysis.* Genet Epidemiol, 2008. **32**(1): p. 1-8.