Integrative genomics identifies lncRNA regulatory networks across 1,044 pediatric leukemias and solid tumors

Apexa Modi[1,2], Gonzalo Lopez[1,3], Karina L. Conkrite[1], Tsz Ching Leung[1], Sathvik Ramanan[1], Daphne Cheung[1], Chun Su[4], Elisabetta Manduchi[4], Matthew E. Johnson[4], Samantha Gadd[5], Jinghui Zhang[6], Malcolm A. Smith[7], Jaime M. Guidry Auvil[8], Daniela S. Gerhard[7], Soheil Meshinchi[9], Elizabeth J. Perlman[5], Stephen P. Hunger[1,10], John M. Maris[1,10,11], Andrew D. Wells[4,12], Struan F.A. Grant[4,13,14], Sharon J. Diskin[1,10,11*]

[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] Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York 10029, USA.

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

[5] 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.

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

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

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

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

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

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

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

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

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

* Corresponding Author: diskin@email.chop.edu

Abstract

Long non-coding RNAs (lncRNAs) play an important role in gene regulation and can contribute to tumorigenesis. While pan-cancer studies of lncRNA expression have been performed for adult malignancies, the lncRNA landscape across pediatric cancers remains largely uncharted. Here, we curate RNA sequencing data for 1,044 pediatric leukemia and solid tumors and integrate paired tumor whole genome sequencing and epigenetic data in relevant cell line models to explore lncRNA expression, regulation, and association with cancer. We report a total of 2,657 robustly expressed lncRNAs across six pediatric cancers, including 1,142 lncRNAs exhibiting histotype-specific expression. DNA copy number alterations contributed to lncRNA dysregulation at a proportion comparable to protein coding genes. Analysis of upstream regulation via tissue-specific, oncogenic transcription factors further implicated 608 distinct histotype-associated lncRNAs. Application of a multi-dimensional framework to identify and prioritize lncRNAs impacting entire gene networks revealed that lncRNAs dysregulated in pediatric cancer are associated with proliferation, metabolism, and DNA damage pathways. Silencing *TBX2-AS1*, the top-prioritized neuroblastoma-specific lncRNA, resulted in significant growth inhibition of neuroblastoma cells. Taken together, these data provide a comprehensive characterization of lncRNA regulation and function in pediatric cancers and pave the way for hypothesis-driven mechanistic studies.

Long non-coding RNAs (lncRNAs) are transcribed RNA molecules greater than 200 nucleotides in length that do not code for proteins. These molecules account for 70% of the expressed human transcriptome and influence key aspects of gene regulation [1-4]. Compared to protein coding genes (PCGs), lncRNAs typically have fewer exons, weaker conservation, and lower abundance [3]. Despite this, lncRNAs have been shown to play significant roles in both transcriptional and post-transcriptional gene regulation [5]. LncRNAs perform these roles by physically interacting with a variety of substrates, including proteins (transcription co-factors), RNAs (microRNA sponges), and DNA (chromatin interaction scaffolds) [1, 2, 6, 7]. While the mechanisms and function for the majority of lncRNAs remain unknown [3, 8], those that have been experimentally characterized are involved in a variety of cellular processes [6] including gene silencing (*ANRIL*) [9], modulation of chromatin architecture (*Xist*) [10], and pre-mRNA processing (*MALAT1*) [11]. LncRNAs are also important in development [12]. For example, the *H19* lncRNA is involved in imprinting [13], while the well-conserved *TUNA* lncRNA controls stem cell pluripotency and lineage differentiation [14].

Dysregulation of lncRNA expression has been widely observed in cancer [3, 15, 16] and studies have shown that lncRNAs play important roles in tumor initiation and progression [17]. LncRNAs can function as tumor suppressors, such as the *PANDA* lncRNA which regulates DNA damage response in Diffuse Large B-cell lymphoma [18]; however, many more lncRNAs appear to be oncogenes. Examples include the *HOTAIR* and *PVT1* lncRNAs which promote proliferation in various cancers through tissue specific mechanisms [19, 20]. Pan-cancer analyses of lncRNA expression in adult malignancies have uncovered many cancer-associated lncRNAs [3, 15-17, 21, 22]. Identification of functional lncRNAs amongst the large set of cancer-associated lncRNAs, however, remains challenging [15, 23]. Current methods to identify putative functional lncRNAs involve identifying lncRNA-specific genetic aberrations [15, 16, 24] or using lncRNA expression to predict overall patient survival [16]. To address how lncRNAs may actually drive cancer, recent computational methods seek to assign function to these molecules based on predicted target genes and regulatory network models. These methods have been applied to adult malignancies and allow for more focused hypotheses to be tested [21, 22].

LncRNA studies and evidence of related function in pediatric cancers have been primarily limited to neuroblastoma (NBL) [25-30], T-lymphoblastic leukemia (T-ALL) [31, 32], and more recently glioblastoma [33]. *CASC15* and *NBAT-1* are a sense-antisense lncRNA pair that map to a NBL susceptibility locus identified by genome-wide association study [26, 34]. Both lncRNAs are downregulated in high-risk NBL tumors and have been shown to be involved in cell proliferation and differentiation [25, 26]. In pediatric T-ALL, the NOTCH- regulated lncRNA, *LUNAR1*, promotes T-ALL cell growth by sustaining IGF1 signaling [32]. To date, it is unknown whether lncRNAs function as common drivers across multiple pediatric cancers, or if instead, the majority of lncRNAs influence oncogenesis in a histotype-specific manner.

Here, we perform a pan-pediatric cancer study of lncRNAs across 1,044 pediatric leukemias and solid tumors [35, 36]. We present the landscape of lncRNA expression across these childhood cancers and perform integrative multi-omic analyses to assess tissue specificity, regulation, and putative function. To validate our approach, we show that silencing of the top-prioritized NBL-specific lncRNA, *TBX2-AS1*, impairs NBL cell growth in human-derived NBL cell line models.

Results

The lncRNA landscape of pediatric cancers

To define the repertoire of lncRNAs expressed in childhood cancers, we analyzed RNA-sequencing data for six distinct pediatric cancer histotypes profiled through the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) project (https://ocg.cancer.gov/programs/target/data-matrix) (**Online Methods; Supplementary Table 1).** This curated set of 1,044 leukemia and solid tumor samples includes 280 acute myeloid leukemia (AML), 190 B-lymphoblastic leukemias (B-ALL), 244 T-lymphoblastic leukemias (T-ALL), 121 Wilms tumors (WT), 48 rhabdoid tumors (RT), and 161 neuroblastomas (NBL) (**Fig. 1a**). Since one of our goals was to identify novel cancer-associated lncRNAs, we performed guided *de novo* transcriptome assembly using StringTie v1.3.3 [37] with the GENCODE v19 database [38] as a gene annotation reference (**Supplementary Fig. 1**). Expressed gene sequences that did not match exons and transcript structures of any known gene in the GENCODE v19 or RefSeq v74 databases were considered putative novel genes (**Supplementary Fig. 1, Online Methods**). Of these novel genes, we identified candidate lncRNAs by using the PLEK v1 algorithm [39] to assess non-coding potential, and then additionally filtered hits by transcript length, exon read coverage, and genomic location (**Fig. 1a, Online Methods, Supplementary Fig. 1**). As validation of our lncRNA discovery pipeline, we observed that 36% (87 of 242) of identified novel lncRNAs not annotated in Gencode v19 (hg19) were indeed annotated in the more recent Gencode v29 (hg38) genome build (**Supplementary Table 2**). To ensure that we considered robustly expressed genes in the setting of cancer heterogeneity and sequencing variability, we selected a conservative expression cutoff of Fragments Per Kilobase of transcript per Million mapped reads (FPKM) >1 in at least 20% of samples for each cancer. Across all cancers there were 15,588 PCGs, 2,512 known lncRNAs, and 145 novel lncRNAs expressed, though the total number of expressed genes varied per cancer (**Fig 1b, Supplementary Table 3**). Collectively, this pan-pediatric cancer catalog of robustly expressed lncRNA and protein-coding genes serves as the basis for downstream integrative lncRNA analyses focused on tissue specificity, predicted function, and gene regulation **(Fig 1b).**

Overall, lncRNAs had lower average expression compared to PCGs resulting in fewer highly expressed lncRNAs (**Supplementary Fig. 2a**). Between 10-100 (3.7%) lncRNAs accounted for 50% of the total sum of lncRNA expression (**Fig. 1c**). In contrast, between 100-1000 (6.4%) PCGs accounted for 50% of the total sum of PCG expression (**Fig. 1d**). We examined the union of the top five most highly expressed lncRNAs across pediatric cancers (total 11 lncRNAs). Some of these lncRNAs had higher expression in the blood cancers (*MALAT1* and *RP11-386I14.4*), in the solid cancers (*H19)*, or in only one cancer, such as *MEG3* and *RP11-386G11.10* in NBL (**Fig. 1e**). Five of these lncRNAs were among the top 10 lncRNAs expressed across normal tissues in the Genotype-Tissue Expression (GTEx) project [40]. Specifically, *C17orf76-AS1 (LRRC75A-AS1), MALAT1, GAS5, SNHG6, SNHG8* were expressed ubiquitously in 30 of the 49 GTEx tissues (**Supplementary Table 4**).

Principal component analysis (PCA) of lncRNA gene expression showed that blood (AML, B-ALL, T-ALL) and solid (NBL, WT, RT) cancers form two distinct groups. Moreover, individual cancer histotypes clustered more closely using lncRNA expression than PCG expression alone (**Supplementary Fig. 2b-c**), consistent with the known tissue specific nature of lncRNA expression and function [3].

Tissue specific lncRNA expression distinguishes pediatric cancers

To evaluate more formally the tissue specific expression of lncRNAs, we annotated all genes with a tissue specificity index (tau score) [41, 42] (**Online Methods**). The established tau score ranges from 0 (ubiquitous expression) to 1 (tissue-specific). As an example, the highly expressed lncRNA *C17orf76-AS1* yielded a tau score of 0.296 in this study, indicating ubiquitous expression (**Supplementary Fig. 2d**). In contrast, the highly expressed *MEG3* lncRNA, which is known to have tissue-specific expression in NBL [30, 43], yielded a tau score of 0.986 (**Supplementary Fig. 2e**). Overall, we observed that lncRNAs yielded a higher tau score range and mean, and thus greater tissue specific expression than PCGs (t-test $p=1.62x10^{-42}$). Novel lncRNAs had the greatest tissue specific expression (t-test: vs proteins- $p=1.62x10^{-42}$, vs known lncRNAs- $p = 3.39x10^{-13}$) (**Fig. 2a**). A tau score threshold of 0.8 has been suggested to distinguish tissue specific genes [42], and using this cutoff we identified 1,142 (43%) tissue specific (TS) lncRNAs (**Fig. 2b, Supplementary Table 5**). To assess how well TS lncRNAs distinguish cancers, we performed clustering based on the top five highest expressed TS lncRNAs per cancer (30 total). The expression of just these lncRNAs was sufficient to cluster samples of the same cancer type (**Fig. 2c**). Furthermore, the blood and solid cancers separately clustered together with little expression overlap observed between the two groups across the 30 genes (**Fig. 2c**). Finally, we identified a similar proportion of TS lncRNAs (38%, n = 1624) across 12 adult cancers from The Cancer Genome Atlas (TCGA) (**Online Methods**) and observed that adult cancer tissue types were also well distinguished based on the expression of the top 5 most TS lncRNAs (**Supplementary Fig. 2f-g**).

Notably, NBL tumors expressed 2.5x more TS lncRNAs (n=522) than the cancer with the next highest: WT (TS lncRNAs: n=211), and 10x more than AML, which had with the least number of TS lncRNAs (n=49) (**Fig. 2b**). To validate NBL's striking quantity of TS lncRNAs, we first assessed whether the known immune and stromal cell infiltration within our NBL tumor samples [36] could be contributing to the variety of lncRNAs expressed. We ran the ESTIMATE algorithm as previously described [36]

(**Online Methods**), to determine levels of immune and stromal cell presence in each tumor sample using expression data. Using these purity estimates, we re-calculated each cancer's tau score and restricted our analysis to NBL samples with either 80% or 90% purity. In both cases, we found that NBL still had the greatest number of TS lncRNAs (n =588 – NBL 90% purity) compared to other cancers (**Supplementary Table 6**). Finally, given that the TARGET NBL RNA-seq dataset is un-stranded, we validated our findings using stranded RNA-seq data in an independent NBL cohort generated through the Gabriela Miller Kids First (GMKF) program (n=223). We observed that 48% of expressed lncRNAs were tissue specific in the GMKF cohort, an increase from the 31% observed in the TARGET cohort (**Supplementary Table 6**). These results confirm NBL's lncRNA abundance and demonstrate that the tau score robustly identifies TS lncRNAs across varying datasets.

Somatic DNA copy number alterations impact lncRNA expression

Many pediatric cancers are marked by a lower single nucleotide variant (SNV) and insertion-deletion (indel) burden than observed in adult cancers [36]. Instead, large chromosomal events, such as somatic copy number aberrations (SCNAs) and other structural variants (SVs) have been shown to dysregulate protein coding driver genes [36, 44]. However, the extent to which large chromosomal alterations impact lncRNAs in pediatric cancers remains unknown. We thus sought to identify SCNAs and SVs using whole genome sequencing (WGS) data from the TARGET project available for NBL (n=146), B-ALL (n=302), AML (n=297), and WT (n=81) (**Online Methods**). The GISTIC v2 algorithm [45] was applied to detect regions of recurrent SCNA (q-value < 0.25) (**Supplementary Fig. 3a**). We identified 673 expressed lncRNAs overlapping 176 significant SCNA regions across the cancers (**Supplementary Table 7**). WGS samples with matched RNA-sequencing were then used to compare lncRNA expression in samples with or without an SCNA event and determine significant differential expression (DE) (**Online Methods**, **Supplementary Table 8**). Across all cancers, between 10-30% of expressed genes overlapping SCNA regions showed significant differential expression based on SCNA, a proportion that was similar for both PCGs and lncRNAs (**Fig 3a**). Altogether, there were 198 (29%) unique lncRNAs with significant DE due to SCNA (**Supplementary Fig 3b**). The majority of the significantly dysregulated lncRNAs were identified in the two cancers with the greatest overall number of expressed lncRNAs, NBL and WT, and mapped to regions with highly recurrent SCNAs in those cancers (chromosomes 1, 7, 11, and 17) (**Fig 3b**).

While SCNAs can cause the dysregulation of lncRNA expression based on gene dosage, structural variant (SV) breakpoints within a lncRNA could cause loss or gain of function [36, 44]. We utilized WGS data to identify lncRNAs disrupted by SV breakpoints using a previously described combination approach involving copy number read-depth and discordant junction approach [44] (**Online Methods**). There were 650 unique expressed lncRNA genes disrupted by SVs, 89% of which were found in only one sample (**Supplementary Fig. 4a)**. We observed 212 SV-impacted lncRNA genes located at SCNA regions (**Fig. 3c**), and 65% of lncRNAs genes disrupted by SV breakpoints in at least five samples were located at SCNA regions (**Supplementary Fig. 4b, Supplementary Table 9**). Indeed, the top-ranked SV-impacted lncRNA *MYCNOS,* in both NBL and WT, associates with the disease driving chr2p24 amplification [46, 47] (**Supplementary Fig. 4c-d)**. In B-ALL, the SV-impacted lncRNAs: *KIAA0125* and *CDKN2B-AS1 (ANRIL)* associate with the well-studied *IGH* translocation and *CDKN2A*/*B* deletion locus (**Supplementary Fig. 4e**) [48]. The top-ranked SV-impacted lncRNA in AML, *MIR181A1HG (MONC)*, associates with a recurrent SCNA deletion on 1q (**Fig 3a**) and is mildly up-regulated in the AML dataset (p = 0.061, **Supplementary Fig. 4f**). *MIR181A1HG* (*MONC*) was described previously as an oncogene in acute megakaryoblastic leukemia [49, 50]. Finally, we observed 30 lncRNAs with pan-cancer (n>3) expression and SV breakpoints(**Supplementary Fig. 4h**). The most number of breakpoints across unique samples was observed in *LINC00910,* which was shown previously to be essential for cell growth in the K562 cell line [51].

Identification of potential cancer driver lncRNAs via integration of epigenetic data

Given the low SNV and indel mutational burden of pediatric cancers, we next sought to determine whether upstream regulation of lncRNAs by known cancer driver transcription factors (TFs) could be used to further implicate cancer-associated lncRNAs. The NBL and T-ALL datasets were chosen for this analysis given that their cancer driver TFs, defined as their core transcriptional circuitry (CRC), have been well-studied [52-55]. The CRC involves a set of co-bound and auto-regulated TFs that drive a cell's transcriptional state and cell identity [56]. CRC-bound regulatory loci were identified from publicly available ChIP-seq data for two MYCN-amplified NBL cell lines, SKNBE(2)C and KELLY, for the CRC TF's: MYCN, PHOX2B, HAND2, GATA3, ISL1, and TBX2 [53, 57] (**Fig. 4a, Online Methods**). Similarly, we used available ChIP-seq data for the TAL1 mutated T-ALL cell lines, Jurkat and CCRF-CEM, to identify loci bound by the T-ALL CRC TF's (TAL1, MYB, GATA3, and RUNX1) [55]. CRC gene regulation occurs both by direct promoter binding (**Fig. 4a-1**) and by distal binding to either promoter (**Fig. 4a-2**) or enhancer regions (**Fig. 4a-3**) which then regulate the gene of interest via long-range chromatin interactions [52-55]. To comprehensively identify both short- and long- range CRC gene regulation, we generated high-resolution (i.e. using 4-cutter restriction enzyme DpnII) genome-wide promoter-focused Capture C [58] in the NBL cell line NB1643 and used publicly available SMC1 (cohesin) ChIA-PET data for the T-ALL Jurkat cell line [54]. After pinpointing gene promoters interacting with CRC TF bound regulatory loci (promoters or enhancers)(**Fig. 4a, Online Methods**), we identified 547 lncRNA genes associated with the NBL CRC and 71 lncRNA genes associated with that of T-ALL's (**Fig 4b, Supplementary Table 10**). Notably, only 249 (NBL) and 22 (T-ALL) of these respective lncRNA genes were bound by CRC TFs within their promoter regions.

Given that co-regulated lncRNA and PCGs could share functional pathways [15, 22], we assessed globally the correlation between CRC-regulated lncRNAs and PCGs on the same chromosome (**Online Methods**). We identified 295 (NBL) and 21 (T-ALL) lncRNAs with significant expression correlation (Pearson's $r > 0.4$ and $FDR < 0.1$) to a CRC-regulated PCG (**Fig 4b**, **Supplementary Table 11**). Since co-regulation of lncRNAs and PCGs may reveal shared functional pathways [15, 22], we performed gene set enrichment analysis of correlated PCGs using the MsigDB's Hallmark Gene Sets (HMS) [59] (Fisher exact test $FDR < 0.1$, **Online Methods**). Results showed enrichment of proliferation and immune related hallmarks in NBL, while signaling hallmarks dominated in T-ALL (**Fig 4c-d, Supplementary Table 11**).

The CRC TFs MYCN and TAL1 are known to be mutated in NBL and T-ALL, respectively. [53, 54]. We therefore investigated CRC-regulated lncRNAs that are associated with the MYCN-amplified and TAL1 subtypes of NBL and T-ALL, respectively. We observed 384 differentially expressed (DE) lncRNAs

(23%) associated with MYCN amplification in NBL and 98 DE-lncRNAs (11%) associated with the previously defined TAL1 subgroup in T-ALL [60], which includes samples with either TAL1 mutation or TAL1-associated gene expression signature (**Supplementary Table 12, Supplementary Fig. 5**). We prioritized 72 (NBL) and 7 (T-ALL) CRC-regulated lncRNAs with differential expression and significant correlation to CRC-regulated protein-coding gene (**Fig. 4b**). One of the strongest correlations identified in NBL was between the DE-lncRNAs, *NR2F1-AS1,* and PCG: *NR2F1* (Pearson's r =0.74, FDR < 0.1) (**Fig. 4e**). The two genes share a CRC-regulated promoter (**Fig. 4f**). *NR2F1* encodes a transcription factor known to bind regulatory elements of neural crest cells, the precursor cells of NBL tumors [61], though the role of *NR2F1-AS1* is unknown. In T-ALL, we observed that the DE-lncRNA, *PRKCQ-AS1,* was highly correlated with *PRKCQ* (Pearson's r =0.66) (**Fig. 4g**). The shared promoter of *PRKCQ-AS1* and *PRKCQ* appeared to be interacting with a CRC-bound enhancer region within the *PRKCQ* gene (**Fig. 4h**). *PRKCQ* is a known T-cell activator and is suggested to have a role in the initiation of leukemia [62], though the role of *PRKCQ-AS1* in leukemia is unknown. Taken together, this novel data integration nominates multiple lncRNAs with previously unknown function for further study as potential driver genes in these respective cancers.

Characterization of transcriptional network perturbation mediated by dysregulated lncRNAs

To determine how lncRNAs may drive pediatric cancers, we examined the downstream impact of lncRNAs on gene regulation. We focused on identifying lncRNAs that mediate transcriptional regulation by modulating TF activity (lncRNA modulators) [63-66]. We wrote custom scripts to implement the lncMod computational framework [67] (**Online Methods**) to first identify DE-lncRNAs and then to assess their impact on correlated expression between a TF and its target genes [21, 67] (**Fig. 5a, Supplementary Fig. 6a, Online Methods**). Across all cancers studied, we identified 313,370 unique, dysregulated lncMod triplets (lncRNA-TF-target gene), representing 0.02-0.2% of possible triplets, which have significant correlation differences between a TF and target gene upon lncRNA expression dysregulation (**Supplementary Table 13-14**). This proportion was consistent with previous findings from the lncMap study in adult cancers [21], although more triplets were identified in datasets with greater sample size

(**Supplementary Table 13-14**). LncRNA modulators were assigned to one of three categories based on their impact on TF-target gene correlation; either the correlation was enhanced, attenuated, or inverted (**Fig 5a-b**). LncRNA modulators have context specific function such that for different TF-target gene pairs they could exert different types of regulation (**Supplementary Fig. 6b**). The majority of lncRNA modulators also appeared to be active in only one cancer, with only 15% (138 of 923 lncRNAs) having pan-cancer activity (n>3) (**Fig. 5c**).

To determine the biological impact of lncRNA modulators, we identified lncRNAs whose target genes were enriched in MsigDB's Hallmark Gene Sets (HMS) [59] (Fisher exact test FDR < 0.1, **Online Methods**). Across the majority of cancers, lncRNA modulator target genes had significant enrichment in the proliferation, metabolism, and DNA damage hallmark categories (FDR range: 0.1 to 2.24×10^{-36}; **Fig. 5d**). Overall, the top-enriched hallmark pathways closely mirrored those found for lncRNA modulators in adult cancers [22]. Consistent with its role in development and as an oncogene in certain cancers [23], the top-enriched hallmarks for the *H19* lncRNA, dysregulated in NBL, were EMT (development) and G2M-checkpoint (proliferation) (**Supplementary Fig. 6c**). The blood cancers exhibited strong enrichment of lncRNA modulators regulating MYC targets, which has a well-established role in leukemias[68]. Furthermore, in AML, we observed that gene targets of the myeloid-specific lncRNA, *HOTAIRM1*, were most enriched for proliferation hallmarks (**Supplementary Fig. 6d**), consistent with this lncRNA's known role in proliferation as an oncogene in adult AML [69].

Finally, we sought to determine potential lncRNA mechanism by identifying recurring patterns of regulation amongst lncMod triplets. To this end, we nominated candidate lncRNA-TF associations by ranking TF's based on the number of target genes regulated by each given TF (**Supplementary Table 15**). As proof-of-concept, we were able to detect known lncRNA-TF associations such as *GAS5* with E2F4 [70] (RNA-protein), and *SNHG1* with *TP53* [71] (RNA-RNA) amongst lncMod triplets in our study (**Supplementary Fig. 6e-f**). A notable example from the hundreds of novel associations identified is between the B-ALL specific lncRNA, *BLACE* (B-cell acute lymphoblastic leukemia expressed, tau score: 0.999) and its top associated TF, XBP1, which has known roles in pre-B-ALL cell proliferation and tumorigenesis [72] (**Fig 5e-f)**. These predictions of lncRNA transcriptional networks provide focused avenues to elucidate the mechanisms through which lncRNAs can drive pediatric cancers.

Integrative multi-omic analysis prioritizes *TBX2-AS1* as a candidate functional lncRNA in NBL

To obtain a comprehensive prioritization of candidate functional lncRNAs, we annotated lncRNA modulators with information on (1) tissue specific expression, (2) dysregulation due to DNA copy number aberration, and (3) regulation by CRC TFs (**Supplementary Table 16**). Here, we focus on the NBL cohort since this cancer has data available for all of the prioritization steps (**Supplementary Table 17**). The top ranked lncRNA in NBL was *MEG3*, which has a known role in both NBL and other cancers [43]. The next prioritized lncRNA, *TBX2-AS1,* has unknown function, yet shares a promoter and is highly correlated (Pearson's r=0.77) with the CRC transcription factor *TBX2* (**Fig. 6a**). TBX2 has been shown to drive NBL proliferation via the *FOXM1*/*E2F1* gene regulatory network [57]. We observed that *TBX2-AS1*, like *TBX2*, was up-regulated in NBLs harboring 17q gain (**Fig. 6b**). In addition, both genes exhibit NBL-specific expression (tau score: *TBX2*- 0.807, *TBX2-AS1*- 0.86; **Supplementary Fig. 7a**). Predictions from our lncMod analysis indicate that *TBX2-AS1* impacts E2F targets and G2M checkpoint genes (**Fig. 6c**), the same pathways observed to be impacted upon knockdown of the TBX2 protein in a previous study [57]. Furthermore, the TFs primarily impacted by TBX2 knockdown [57], MYBL2 and E2F1, were found to have the most target genes predicted to be regulated by *TBX2-AS1* (**Fig 6d-e**). Evidence for this association was further supported by the correlation (Spearman's rho > 0.4) between *TBX2-AS1* and *TBX2*'s target TFs: *FOXM1*, *E2F1*, and *MYBL2* (**Supplementary Fig. 7b**). While the strong correlation between *TBX2-AS1* and *TBX2* may confound our predictions, a previous study showed positionally conserved lncRNAs, such as *TBX2-AS1*, which share promoter regulation with developmental TFs (TBX2), can play roles in genome organization, development, and cancer [69]. Based on the promising *in silico* evidence, we prioritized *TBX2-AS1* for experimental study.

Silencing of *TBX2-AS1* inhibits cell growth and alters morphology of neuroblastoma cells

We assessed the role of *TBX2-AS1* using human-derived NBL cell line models. First, we evaluated *TBX2-AS1* expression across 38 NBL cell lines using RNA-seq [73] (**Supplementary Fig 8a**). Expression of *TBX2* and *TBX2-AS1* were subsequently validated in eight cell lines using RT-qPCR (**Supplementary Fig. 8b**). We selected NLF and SKNSH models for further study based on their high *TBX2-AS1* expression and differing levels of *TBX2* expression. Silencing of *TBX2-AS1* using small interfering RNA (siRNA) achieved 63 - 95% reduction of *TBX2-AS1* expression (**Fig. 6f, Supplementary Fig. 8c**). We monitored cell growth via the real-time cell electronic sensing (RT-CES) system and observed that si*TBX2-AS1* treated cells exhibited an average of 46.6% decreased cell growth in NLF (SD =0.02, t-test: p = 8.1 x 10^{-4}) and 42% in SKNSH as compared to non-targeting control (NTC) (SD =0.06, t-test: p = 3.87 x 10^{-3}) (**Fig. 6g**). In addition, live cell analysis using the IncuCyte revealed changes in cell morphology for si*TBX2-AS1* treated cells, featuring an appearance of disrupted cell to cell adhesion and elongated cell body (**Supplementary Fig. 8c**). Taken together, these data demonstrate the utility of our integrative lncRNA characterization and prioritization approach and suggest that *TBX2-AS1* is a newly identified functional lncRNA in NBL.

Discussion

LncRNAs have emerged as important regulators of gene expression and their dysregulation can impact key cancer pathways and drive tumorigenesis [1-4]. Despite this, relatively few lncRNAs have been experimentally characterized and the landscape of lncRNA expression across pediatric cancers remained unknown. In this study, we explored lncRNA expression, cancer association, and regulatory networks across 1,044 pediatric leukemias and solid tumors, representing six different cancer types. The breadth of samples and cancer types included allowed for robust identification of novel and cancer-specific lncRNAs, and facilitated identification of expression patterns for both up- and downstream lncRNA gene regulation. We provide multi-dimensional insight into the predicted biological and functional relevance of lncRNAs by integrating WGS, ChIP-seq, chromatin capture, and predictions of transcriptional networks.

Analysis of the lncRNA landscape across pediatric cancers revealed the histotype and context-specific nature of lncRNAs. We report a total of 2,657 robustly expressed lncRNAs across the six cancer types studied. This number is notably smaller than reports from pan-cancer studies of adult malignancies [15, 17], likely due to the smaller number of cancer types studied here and conservative expression threshold applied. However, similar to our findings in adult cancers, 43% (1,142/ 2,657) of expressed lncRNAs exhibited tissue-specific (TS) expression across pediatric cancers. Indeed, lncRNAs had significantly greater tissue specificity than protein coding genes, making them more ideal candidates as biomarkers. Currently there is one lncRNA, *PCA3*, that is FDA-approved as a biomarker for prostate cancer [74] and multiple trials investigating ncRNAs in cancer prognostics are underway [75]. In this study, the top five most TS lncRNAs per cancer were sufficient to differentiate each cancer histotype, suggesting there is potential for a small number of lncRNAs to be used as highly sensitive markers in childhood cancers.

Typically, investigation of lncRNA dysregulation involves comparing lncRNA expression between cancer and normal control samples and is an analysis that amply yields adult-cancer associated lncRNAs [15]. However, the lack of normal expression controls for the majority of pediatric cancers [36] is a major complication in defining pediatric cancer-associated lncRNAs. To overcome this and the previously described low mutation burden [46, 47, 60], we integrated ChIP-sequencing of CRC transcription factors with our expression data to identify cancer-associated lncRNAs. CRC TFs bind to cell-type-specific enhancers and regulate the expression of cell-type-specific genes [76]. By taking advantage of this information we were able to prioritize lncRNAs likely to be important for cancer cell identity based on CRC TF regulation. CRC TFs have been well defined for NBL and T-ALL [53, 55]; however, the fact that they largely bind enhancer regions necessitated that we also use chromatin interaction data to accurately determine regulated genes. Incorporation of these datasets allowed us to identify 2-fold more CRC regulated lncRNAs in NBL and 3-fold in T-ALL as compared to using just ChIP-seq data alone, which restricts lncRNA identification to those with CRC TFs bound at their promoter. Notably, there were ten common CRC-regulated lncRNAs between NBL and T-ALL, and an important next step for further identification of pan-pediatric cancer associated lncRNAs is application of this novel analysis to a broader set of pediatric cancers.

While upstream regulation can help nominate cancer-associated lncRNAs, determining the mechanism through which dysregulated lncRNAs impact downstream target genes is also crucial. However, prediction of lncRNA function is limited given that very few lncRNA mechanisms have been fully established and lncRNAs lack conserved sequence and structure [77]. Many studies instead use correlated protein coding gene expression as a proxy to define lncRNA pathways, but this approach often results in many false positives and does not provide mechanistic insight [77]. To address this, we used the lncMod method [21, 67] to model the functional mechanism of dysregulated lncRNAs by examining correlated changes in transcription factor to target gene regulation. We used motif presence and regression analysis to identify TF-target gene relationships, though future studies will be strengthened by incorporating TF ChIP-seq data, when it becomes more widely available for pediatric cancers. Nevertheless, we were able to successfully associate lncRNAs to TFs with known interactions, such as SNHG1 with TP53, while also providing a prioritized list of novel associations that serve as a starting point for future experimental studies such as RIP/MS [78] and ChiRP-seq [79]. Finally, while our lncMod analysis was focused on transcriptional regulation, the addition of microRNA binding and RNA-binding protein data, as utilized in adult cancers [22], is an important next step in understanding how lncRNAs impact post-transcriptional regulation in pediatric cancers.

Our study delineated high confidence lncRNA expression across pediatric cancers within the restrictions set by the sequencing depth and RNA-seq type available per cancer dataset. We required RNA-seq samples included in our study to have at least 10 million reads and read length of at least 75 bp; and with the exception of the T-ALL samples, all samples were poly-A selected. Future studies involving total RNA-seq, greater sequencing depth, and longer read sizes could capture a larger diversity and more accurate set of expressed lncRNAs by accounting for non-polyadenylated genes and identifying scarcer or temporally expressed lncRNAs. Nevertheless, our high confidence set of lncRNAs are very likely to be functional given that low or rare expression can be an indicator of transcriptional noise [80]. In addition to having a limited number of RNA matched WGS samples, the Complete Genomics short read technology limits the detection of structural variants based on size as previously described [36, 44]. The use of long-read sequencing and greater sequencing depth in future studies will enable more accurate copy number and structure variant detection in pediatric cancers.

Finally, multi-dimensional integration of our computational predictions resulted in the nomination of functionally relevant lncRNAs in each pediatric cancer. We annotated tissue specificity, copy number, pathway, and likely targets for these lncRNAs, providing a solid foundation for mechanistic studies. As proof-of-principle, we demonstrate that the top-prioritized tissue-specific and copy number dysregulated lncRNA, *TBX2-AS1,* impacts NBL cell growth, validating our approach and corroborating the pathway analysis results. Overall, this study provides a comprehensive characterization of lncRNAs across pediatric cancers and serves as a rich resource for future mechanistic studies; these data may also aid in the selection of cancer biomarkers and candidate therapeutic targets.

Online Methods

RNA-seq data processing. A comprehensive RNA-seq analysis pipeline was used on all samples (Supp. Table 1, Supp. Fig 1). First FASTQC was run on all samples and any samples that had a Phred score < 30 for more than 25% of read bases were removed. Samples were then aligned using STAR_2.4.2a [81] with the following parameters: "STAR --runMode alignReads --runThreadN 10 --twopassMode Basic --twopass1readsN -1 --chimSegmentMin 15 --chimOutType WithinBAM –genomeDir X--genomeFastaFiles ucsc.hg19.fa --readFilesIn fasta1 fasta2 --readFilesCommand zcat --outSAMtype BAM SortedByCoordinate --outFileNamePrefix X --outSAMstrandField intronMotif --quantMode TranscriptomeSAM GeneCounts --sjdbGTFfile gencode.v19.annotation.gtf --sjdbOverhang X." To assess the quality of the aligned RNA-seq data we ran MultiQC [82], and removed samples with < 70% uniquely mapped reads and < 10 million mapped reads.

Gene/transcript mapping and quantification. To map reads to genes and quantify gene expression we ran StringTie 1.3.3 [37]. 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 pan-TARGET 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 or RefSeq v74 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 algorithm tool [39]. 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 [39]. 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.

Tissue specific gene expression. The tau score, a measure of the tissue specific expression of a gene was calculated as described by Yadai et. al [35]. The formula for the score is listed below. x_i is defined as the mean expression of a gene in a particular cancer and n is the total number of cancers considered, in this case n = 6. Since the total RNA-sequencing method was used for the T-ALL dataset, we removed genes that have been previously confirmed [83] to lack a polyA tail (146 protein coding genes, 4 lncRNAs), given that the tau score for these genes would not be interpretable.

$$\tau = \frac{\sum_{i=1}^{n}(1-\hat{x_i})}{n-1};\hat{x_i} = \frac{x_i}{\max\limits_{1\leq x\leq n}\ (x_i)}$$

CNV detection, processing, and impact on gene expression. Copy number calls were made by Complete Genomics (CGI) from WGS for NBL, WT, AML, and B-ALL. We used CGI files: "somaticCnvDetailsDiploidBeta" containing ploidy estimates and tumor/blood coverage along 2kb bins across the genome. To create segmentation files, we used custom scripts to reformat CGI coverage data to meet requirements of the "copynumber" R bioconductor package as previously described [44]. We used the winsorize function in this package, which performs data smoothing and segmentation via a piecewise constant segmentation (pcf) algorithm (kmin =2 and gamma= 1000). Segmentation files were visualized using the R package svpluscnv (https://github.com/ccbiolab/svpluscnv) https://doi.org/10.1093/bioinformatics/btaa878. We then ran GISTIC2.0, using segmentation data as inputs and parameters: "GISTIC2 -v 30 -refgene hg19 -genegistic 1 -smallmem 1 -broad 1 -twoside 1 -brlen 0.98 -conf 0.90 -armpeel 1 -savegene 1 -gcm extreme -js 2 -rx 0". To determine which genes are impacted by copy number, we intersected CNV regions listed in the "all_lesions.conf_90.txt" file from GISTIC output with gene positions. We used section 1 from the "all_lesions.conf_90.txt" file to assign a binary descriptor to each gene as either being not amplified or deleted (CNV-no) if the sample had actual copy gain 0 for the region containing the gene. We assigned CNV-yes if the region containing the gene was amplified or deleted, which included samples with actual copy gain 1 or 2, where 1 indicates low level copy number aberration (exceeds low threshold of copy number: 1: 0.1<t< 0.9) and 2 indicates a high level of copy number aberration, CNV exceeds high threshold (t>0.9) according to GISTIC. 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 detection and filtering. Structural variants were identified from WGS via identification of incongruously aligned read mate-pairs. SVs involve sequence junctions spanning two breakpoints in the genome (SJ-BP). Additionally, breakpoints for copy number events can also be identified using read-depth tumor-blood ratios (RD-BP) converted to segmentation profiles. This method can provide breakpoint resolution at 2kb, unlike sequence junction break points where location is known at a 1bp resolution. Furthermore, with copy number read depths we can only know dosage information (amplification vs deletion), while with sequence junctions we can determine the size and type of variant (inversion (>30bp), translocation, deletion (>500bp), duplications (>40bp), and tandem-duplication). Nevertheless, these two distinct approaches to identify SVs can provide orthogonal validation for some events.

Somatic sequence junctions that were completely absent in the normal genome are reported by CGI in the somaticAllJunctionsBeta file. To obtain a high confidence set of junctions, where there is a likely true physical connection between the left and right sections of a junction, the following filtering was applied by CGI to obtain the highConfidenceSomaticAllJunctionsBeta.

1) DiscordantMatePairAlignments ≥ 10 (10 or more discordant mate pairs in cluster
2) JunctionSequenceResolve = Y (local de novo assembly is successful)
3) Exclude interchromosomal junction if present in any genomes in baseline samples (FrequencyInBaseline > 0)
4) Exclude the junction if overlap with known underrepresented repeats (KnownUnderrepresentedRepeat = Y): ALR/Alpha, GAATGn, HSATII, LSU_rRNA_Hsa, and RSU_rRNA_Hsa
5) Exclude the junction if the length of either of the side sections is less than 70 base pairs.

Further filtering of these high confidence structural variants included removing rare/common germline variants that passed the CGI filters. We used the Database of Genomic Variants (DGV v. 2016-05-15, GRCh37) in order to remove SVs that had at least 50% reciprocal overlap with DGV annotated common events and were type matched.

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 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.

ChIP-seq data analysis. To determine which lncRNAs are regulated by transcription factors involved in the core regulatory circuitry (CRC) we utilized previously generated/ analyzed histone and transcription factor ChIP-sequencing data for NBL and T-ALL. 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 [84], 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 [53] 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 mergePeaks tool: “mergePeaks -d 1000 -cobound 6 bed_file1… bed_file6” and the resulting coBoundBy4 output file. For the T-ALL CRC we obtained overlapping CRC transcription factor loci for TAL1, GATA3, and RUNX1 from the study by Sanda et. al [55], GEO: GSE29181 for both the Jurkat and CCRF-CEM cell lines and integrated ChIP-seq data for the MYB transcription factor from GEO: GSE59657 [54], only available in the Jurkat line. We selected loci for further analysis if they were bound by TAL1, GATA3, and RUNX1 as previously annotated by Sanda et. al.

Identification of CRC transcription factor regulated genes. To identify genes regulated by the NBL or T-ALL 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. We performed a similar analysis in T-ALL, however we utilized publicly available SMC1 (cohesin) ChIA-PET data available on the ENCODE project to consider chromatin interactions.

Promoter-focused Capture C data generation. High resolution promoter-focused Capture C was performed in the neuroblastoma cell line, NB1643, (untreated) in triplicate. Cell fixation, 3C library generation, capture C, and sequencing was performed as described by Chesi et. al (2019) and Su et al (2020). For each replicate, 10^7 fixed cells were centrifuged to cell pellets and split to 6 tubes for a pre-digestion incubation with 0.3%SDS, 1x NEB DpnII restriction buffer, and dH2O for 1hr at 37ºC shaking at 1,000rpm. A 1.7% solution of Triton X-100 was added to each tube and shaking was continued for another hour.10 ul of DpnII (NEB, 50 U/µL) was added to each sample tube and continued shaking for 2 days. 100uL Digestion reaction was then removed and set aside for digestion efficiency QC.The remaining samples were heat inactivated incubated at 1000 rpm in a MultiTherm for 20 min, at 65°C to inactivate the DpnII, and cooled on ice for 20 additional minutes. Digested samples were ligated with 8 uL of T4 DNA ligase (HC ThermoFisher, 30 U/µL) and 1X ligase buffer at 1,000 rpm overnight at 16°C .The ligated samples were then de-crosslinked overnight at 65°C with Proteinase K (20 mg/mL, Denville Scientific) along with pre-digestion and digestion control. Both controls and ligated samples were incubated for 30 min at 37°C with RNase A (Millipore), followed by phenol/chloroform extraction, ethanol precipitation at -20°C, then the 3C libraries were centrifuged at 3000 rpm for 45 min at 4°C to pellet the samples. The pellets of 3C libraries and controls were resuspended in 300uL and 20µL dH2O, respectively, and stored at −20°C. Sample concentrations were measured by Qubit. Digestion and ligation efficiencies were assessed by gel electrophoresis on a 0.9% agarose gel and also by quantitative PCR (SYBR green, Thermo Fisher).

Isolated DNA from 3C libraries was quantified using a Qubit fluorometer (Life technologies), and 10 μg of each library was sheared in dH2O using a QSonica Q800R to an average fragment size of 350bp.QSonica settings used were 60% amplitude, 30s on, 30s off, 2 min intervals, for a total of 5 intervals at 4 °C. After shearing, DNA was purified using AMPureXP beads (Agencourt). DNA size was assessed on a Bioanalyzer 2100 using a DNA 1000 Chip (Agilent) and DNA concentration was checked via Qubit. SureSelect XT library prep kits (Agilent) were used to repair DNA ends and for adaptor ligation following the manufacturer protocol. Excess adaptors were removed using AMPureXP beads. Size and concentration were checked again by Bioanalyzer 2100 using a DNA 1000 Chip and by Qubit fluorometer before hybridization. One microgram of adaptor-ligated library was used as input for the SureSelect XT capture kit using manufacturer protocol and custom-designed 41K promoter Capture-C probe set. The quantity and quality of the captured libraries were assessed by Bioanalyzer using a high sensitivity DNA Chip and by Qubit fluorometer. SureSelect XT libraries were then paired-end sequenced on Illumina NovaSeq 6000 platform (51bp read length) at the Center for Spatial and Functional Genomics at CHOP.

Promoter-focused Capture C data analysis. Paired-end reads from each replicated were pre-processed using the HICUP pipeline (v0.5.9), with bowtie2 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) 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.

lncRNA pathway association. Each CRC regulated lncRNA was correlated (Pearson's r) with the expression of all CRC regulated PCGs in NBL and T-ALL respectively. To account for multiple testing, we applied the Benjamini-Hoschberg method and selected correlated lncRNA-PCG pairs with adjusted p-value. < 0.1. We then assigned these PCGs and by association their correlated lncRNAs to pathways using Fisher exact test, FDR < 0.1 for gene sets in the MsigDB Hallmarks Gene Set Collection. We similarly assigned pathways to lncRNA modulators identified from our lncMod analysis based on the pathways assigned to their target PCGs.

Differential gene expression analysis for NBL/T-ALL subtypes. We identified differentially expressed genes using the DESeq2 tool. We compared gene expression between MYCN Amplified NBL and Not-Amplified samples as annotated in the TARGET clinical file (https://ocg.cancer.gov/programs/target/data-matrix). We also elucidated expression differences between the TAL1 subgroup of T-ALL samples as compared to other T-ALL subgroups. The TAL1 subtype was defined previously by Liu, et al [60] based on samples with either TAL1 mutation or TAL1-associated gene expression signature. We ran DESeq2 using default parameters and considered genes as significantly differentially expressed if their absolute value of the log2 fold change was > 0.58 and their Benjamini-Hoschberg adjusted-p value was < 0.01.

lncMod implementation: transcription factor target gene regulation. We developed custom Python scripts to implement the general framework of the lncMod method. The first part of this framework involved determining transcription factor target gene regulation specific to each cancer. Target genes here are defined as any protein coding or lncRNA gene and excludes pseduogenes 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 [85] and predictions of binding across the genome determined by FIMO 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 perturbations, we first delineated genes (TF, target genes, or lncRNAs) that had high expression variance (IQR > 1.5). We evaluated each differentially expressed lncRNA in each cancer in a manner similar to previous studies [21, 22, 67]. Specifically, 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 difference in correlation we first normalized the correlation using the Fisher r to z transformation. Then we calculated the rewiring score, z-statistic, as previously described [21], which is used to describe the degree of regulation change between the TF and target gene.

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

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

As a departure from what is described by Li et. al (lncMod method) [67], 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 [86, 87]. 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.

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

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 [22].

Cell lines and reagents. NBL cell lines were obtained from the American Type Tissue Culture Collection (ATCC) and grown in RPM1-1640 with HEPES, L-glutamine and phenol red, supplemented with 10% FBS, 1% L-glutamine in an incubator at 37°C with 5% CO_2. Cell line identity was confirmed biennially through genotyping and confirmation of STR (short tandem repeat) profiles, while routine testing for Mycoplasma contamination was confirmed to be negative.

siRNA and growth assays. The NBL cell lines, NLF and SKNSH, were plated in a 96-well RTCES microelectronic sensor array (ACEA Biosciences, San Diego, CA, USA). Cell density measurements were made every hour and were normalized to 24 hours post-plating (at transfection time). We used siRNAs to knockdown the expression of genes in NLF and SKNSH. The siRNAs utilized were either a non-targeting negative control siRNA (Silencer^{TM} Select Negative Control siRNA, cat #4390843), TBX2-AS1 Silencer^{TM} Select siRNA (cat # n514841), and SMARTpool: ON-TARGETplus PLK1 siRNA (cat # L-003290-00-0010). Transfection of cells was done using the DharmaFECT 1 transfection reagent (cat # T-2001-02). siRNA at a concentration of 50nM and 2% DharmaFECT was added to RPMI medium without 10% FBS or any antibiotic separately and then incubated at room temperature for 5 minutes. The siRNA medium was then added to the DharmaFECT and incubated for another 20 minutes to form a complex. This solution was then mixed with our normal growth media and applied to cells 24 hours after they had been initially plated. All experiments were repeated in triplicate, with technical replicates (n=3) being averaged per biological replicate.

Real time quantitative PCR. Total RNA was extracted from NBL cells using miRNeasy kit (Qiagen) and the provided protocol for animal cells. The concentration of RNA was determined with the Nanodrop (Thermo Scientific). cDNA synthesis was performed using the SuperScript^{TM} First-Strand Synthesis System for RT-PCR using the SuperScript^{TM} reverse transcriptase (Invitrogen). 5-20ng of cDNA were mixed with the TaqMan Universal PCR Master Mix (Thermo Fisher Scientific) and TaqMan probes/primers for either TBX2-AS1 (Hs00417285_m1) or the house keeping gene, *HPRT1* (Hs02800695_m1). Gene expression from these reactions were measured using RT-qPCR and *TBX2-AS1* expression was normalized to *HPRT1* expression.

Data Availability

All TARGET RNA and DNA-sequencing data analyzed in this study are available through the database of Genotypes and Phenotypes (dbGaP; https://www.ncbi.nlm.nih.gov/gap/) under study-id phs000218 and accession number phs000467. Neuroblastoma cell line RNA-sequencing data analyzed in this study are available through GEO at accessions GSE89413. NBL histone ChIP-seq and transcription factor ChIP-seq data used in this study are both available through GEO at accessions GSE138315 and GSE94822, respectively. T-ALL transcription factor ChIP-seq data and SMC1 ChIA-PET data are available through GEO at accessions GSE29181, GSE59657, and GSE68977.

Acknowledgements

This work was supported in part by NIH grants R01-CA124709 (S.J.D.), R03 CA230366 (S.J.D.), X01 HL136997 (J.M.M.), and T32-HG46-18 (A.M.). This project was also funded in part by a supplement to the Children's Oncology Group Chair's grant CA098543 and with federal funds from the National Cancer Institute, National Institutes of Health, under Contract No. HHSN261200800001E to S.J.D and Complete Genomics. Promoter Capture C studies were funded by the Center for Spatial and Functional Genomics (A.D.W and S.FA.G) at CHOP. S.FA.G. was supported by the Daniel B. Burke Endowed and Chair for Diabetes Research and R01 HG010067.

Author Contributions

A.M. and S.J.D. conceived and designed the study. M.A.S., J.M.G.A, D.S.G., E.J.P, S.M., S.P.H., S.J.D. and J.M.M. generated the TARGET data. K.L.C., M.E.J., S.J.D., A.D.W. and S.F.A.G. generated promoter-focused capture C data. E.M. C.S, and A.M. analyzed promoter-focused capture C data. A.M., G.L., S.R. analyzed TARGET data. A.M., K.L.C., T.C.L. and D.C. performed TBX2-AS1 experiments. S.J.D. supervised the study. A.M. and S.J.D drafted the manuscript. All authors edited and approved the manuscript.

References

1. Rinn, J.L. and H.Y. Chang, *Genome regulation by long noncoding RNAs.* Annu Rev Biochem, 2012. **81**: p. 145-66.
2. Bonasio, R. and R. Shiekhattar, *Regulation of transcription by long noncoding RNAs.* Annual review of genetics, 2014. **48**: p. 433-55.
3. Iyer, M.K., et al., *The landscape of long noncoding RNAs in the human transcriptome.* Nature genetics, 2015. **47**: p. 199-208.
4. Gil, N. and I. Ulitsky, *Regulation of gene expression by cis-acting long non-coding RNAs.* Nat Rev Genet, 2020. **21**(2): p. 102-117.
5. Dykes, I.M. and C. Emanueli, *Transcriptional and Post-transcriptional Gene Regulation by Long Non-coding RNA.* Genomics Proteomics Bioinformatics, 2017. **15**(3): p. 177-186.
6. Marchese, F.P., I. Raimondi, and M. Huarte, *The multidimensional mechanisms of long noncoding RNA function.* Genome Biol, 2017. **18**(1): p. 206.
7. Villegas, V.E. and P.G. Zaphiropoulos, *Neighboring gene regulation by antisense long non-coding RNAs.* International journal of molecular sciences, 2015. **16**: p. 3251-66.
8. Kopp, F. and J.T. Mendell, *Functional Classification and Experimental Dissection of Long Noncoding RNAs.* Cell, 2018. **172**(3): p. 393-407.
9. Kotake, Y., et al., *Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene.* Oncogene, 2011. **30**(16): p. 1956-62.
10. Engreitz, J.M., et al., *The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome.* Science, 2013. **341**(6147): p. 1237973.
11. Tripathi, V., et al., *The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation.* Mol Cell, 2010. **39**(6): p. 925-38.
12. Perry, R.B. and I. Ulitsky, *The functions of long noncoding RNAs in development and stem cells.* Development, 2016. **143**(21): p. 3882-3894.
13. Monnier, P., et al., *H19 lncRNA controls gene expression of the Imprinted Gene Network by recruiting MBD1.* Proc Natl Acad Sci U S A, 2013. **110**(51): p. 20693-8.
14. Lin, N., et al., *An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment.* Mol Cell, 2014. **53**(6): p. 1005-19.
15. Yan, X., et al., *Comprehensive Genomic Characterization of Long Non-coding RNAs across Human Cancers.* Cancer Cell, 2015. **28**(4): p. 529-540.
16. Du, Z., et al., *Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer.* Nat Struct Mol Biol, 2013. **20**(7): p. 908-13.
17. Lanzós, A., et al., *Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features.* Scientific Reports, 2017. **7**: p. 1-16.
18. Wang, Y., et al., *Discovery and validation of the tumor-suppressive function of long noncoding RNA PANDA in human diffuse large B-cell lymphoma through the inactivation of MAPK/ERK signaling pathway.* Oncotarget, 2017. **8**(42): p. 72182-72196.
19. Hajjari, M. and A. Salavaty, *HOTAIR: an oncogenic long non-coding RNA in different cancers.* Cancer Biol Med, 2015. **12**(1): p. 1-9.
20. Onagoruwa, O.T., et al., *Oncogenic Role of PVT1 and Therapeutic Implications.* Front Oncol, 2020. **10**: p. 17.
21. Li, Y., et al., *LncMAP: Pan-cancer Atlas of long noncoding RNA-mediated transcriptional network perturbations.* Nucleic Acids Research, 2018. **46**: p. 1113-1123.

22. 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.
23. Huarte, M., *The emerging role of lncRNAs in cancer.* Nature medicine, 2015. **21**: p. 1253-61.
24. Xu, Y., et al., *Identification and comprehensive characterization of lncRNAs with copy number variations and their driving transcriptional perturbed subpathways reveal functional significance for cancer.* Brief Bioinform, 2019.
25. Mondal, T., et al., *Sense-Antisense lncRNA Pair Encoded by Locus 6p22.3 Determines Neuroblastoma Susceptibility via the USP36-CHD7-SOX9 Regulatory Axis.* Cancer Cell, 2018. **33**: p. 417-434.e7.
26. Russell, M.R., et al., *CASC15-S is a tumor suppressor lncRNA at the 6p22 neuroblastoma susceptibility locus.* Cancer Res, 2016. **75**: p. 3155-3166.
27. Pandey, G.K., et al., *The Risk-Associated Long Noncoding RNA NBAT-1 Controls Neuroblastoma Progression by Regulating Cell Proliferation and Neuronal Differentiation.* Cancer Cell, 2014. **26**: p. 722-737.
28. Sahu, D., et al., *Co-expression analysis identifies long noncoding RNA SNHG1 as a novel predictor for event-free survival in neuroblastoma.* Oncotarget, 2016. **7**: p. 58022-58037.
29. Mazar, J., et al., *The long non-coding RNA GAS5 differentially regulates cell cycle arrest and apoptosis through activation of BRCA1 and p53 in human neuroblastoma.* Oncotarget, 2016. **5**: p. 6589-6607.
30. Rombaut, D., et al., *Integrative analysis identifies lincRNAs up- and downstream of neuroblastoma driver genes.* Sci Rep, 2019. **9**(1): p. 5685.
31. Ngoc, P.C.T., et al., *Identification of novel lncRNAs regulated by the TAL1 complex in T-cell acute lymphoblastic leukemia.* Leukemia, 2018. **32**(10): p. 2138-2151.
32. Trimarchi, T., et al., *Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia.* Cell, 2014. **158**(3): p. 593-606.
33. Liu, Y., H. Liu, and D. Zhang, *Identification of novel long non-coding RNA in diffuse intrinsic pontine gliomas by expression profile analysis.* Oncol Lett, 2018. **16**(5): p. 6401-6406.
34. McDaniel, L.D., et al., *Common variants upstream of MLF1 at 3q25 and within CPZ at 4p16 associated with neuroblastoma.* PLoS Genet, 2017. **13**(5): p. e1006787.
35. Downing, J.R., et al., *The Pediatric Cancer Genome Project.* Nat Genet, 2012. **44**(6): p. 619-22.
36. Ma, X., et al., *Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours.* Nature, 2018. **555**(7696): p. 371-376.
37. Pertea, M., et al., *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.* Nature biotechnology, 2015. **33**: p. 290-5.
38. Frankish, A., et al., *GENCODE reference annotation for the human and mouse genomes.* Nucleic Acids Res, 2019. **47**(D1): p. D766-D773.
39. 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.
40. Consortium, G.T., *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.
41. Yanai, I., et al., *Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.* Bioinformatics, 2005. **21**(5): p. 650-9.
42. Kryuchkova-Mostacci, N. and M. Robinson-Rechavi, *A benchmark of gene expression tissue-specificity metrics.* Brief Bioinform, 2017. **18**(2): p. 205-214.

43. Dong, K., W. Tang, and R. Dong, *MEG3, HCN3 and linc01105 influence proliferation and apoptosis of neuroblastoma cells via HIF-1 alpha and p53 pathway.* Pediatric Blood and Cancer, 2016. **63**: p. S194.
44. Lopez, G., et al., *Somatic structural variation targets neurodevelopmental genes and identifies SHANK2 as a tumor suppressor in neuroblastoma.* Genome Res, 2020. **30**(9): p. 1228-1242.
45. Mermel, C.H., et al., *GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.* Genome Biol, 2011. **12**(4): p. R41.
46. Pugh, T.J., et al., *The genetic landscape of high-risk neuroblastoma.* Nat Genet, 2013. **45**(3): p. 279-84.
47. Gadd, S., et al., *A Children's Oncology Group and TARGET initiative exploring the genetic landscape of Wilms tumor.* Nat Genet, 2017. **49**(10): p. 1487-1494.
48. Harvey, R.C., et al., *Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling: correlation with genome-wide DNA copy number alterations, clinical characteristics, and outcome.* Blood, 2010. **116**(23): p. 4874-84.
49. Emmrich, S., et al., *LincRNAs MONC and MIR100HG act as oncogenes in acute megakaryoblastic leukemia.* Mol Cancer, 2014. **13**: p. 171.
50. Gruber, T.A. and J.R. Downing, *The biology of pediatric acute megakaryoblastic leukemia.* Blood, 2015. **126**(8): p. 943-9.
51. Liu, Y., et al., *Genome-wide screening for functional long noncoding RNAs in human cells by Cas9 targeting of splice sites.* Nat Biotechnol, 2018.
52. Boeva, V., et al., *Heterogeneity of neuroblastoma cell identity defined by transcriptional circuitries.* Nat Genet, 2017. **49**(9): p. 1408-1413.
53. 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.
54. Mansour, M.R., et al., *An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element.* Science, 2014. **346**: p. 1373-1377.
55. Sanda, T., et al., *Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia.* Cancer Cell, 2012. **22**(2): p. 209-21.
56. Saint-Andre, V., et al., *Models of human core transcriptional regulatory circuitries.* Genome Res, 2016. **26**(3): p. 385-96.
57. Decaesteker, B., et al., *TBX2 is a neuroblastoma core regulatory circuitry component enhancing MYCN/FOXM1 reactivation of DREAM targets.* Nat Commun, 2018. **9**(1): p. 4866.
58. Chesi, A., et al., *Genome-scale Capture C promoter interactions implicate effector genes at GWAS loci for bone mineral density.* Nat Commun, 2019. **10**(1): p. 1260.
59. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.
60. Liu, Y., et al., *The genomic landscape of pediatric and young adult T-lineage acute lymphoblastic leukemia.* Nat Genet, 2017. **49**(8): p. 1211-1218.
61. Rada-Iglesias, A., et al., *Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest.* Cell Stem Cell, 2012. **11**(5): p. 633-48.
62. Brezar, V., W.J. Tu, and N. Seddiki, *PKC-Theta in Regulatory and Effector T-cell Functions.* Front Immunol, 2015. **6**: p. 530.

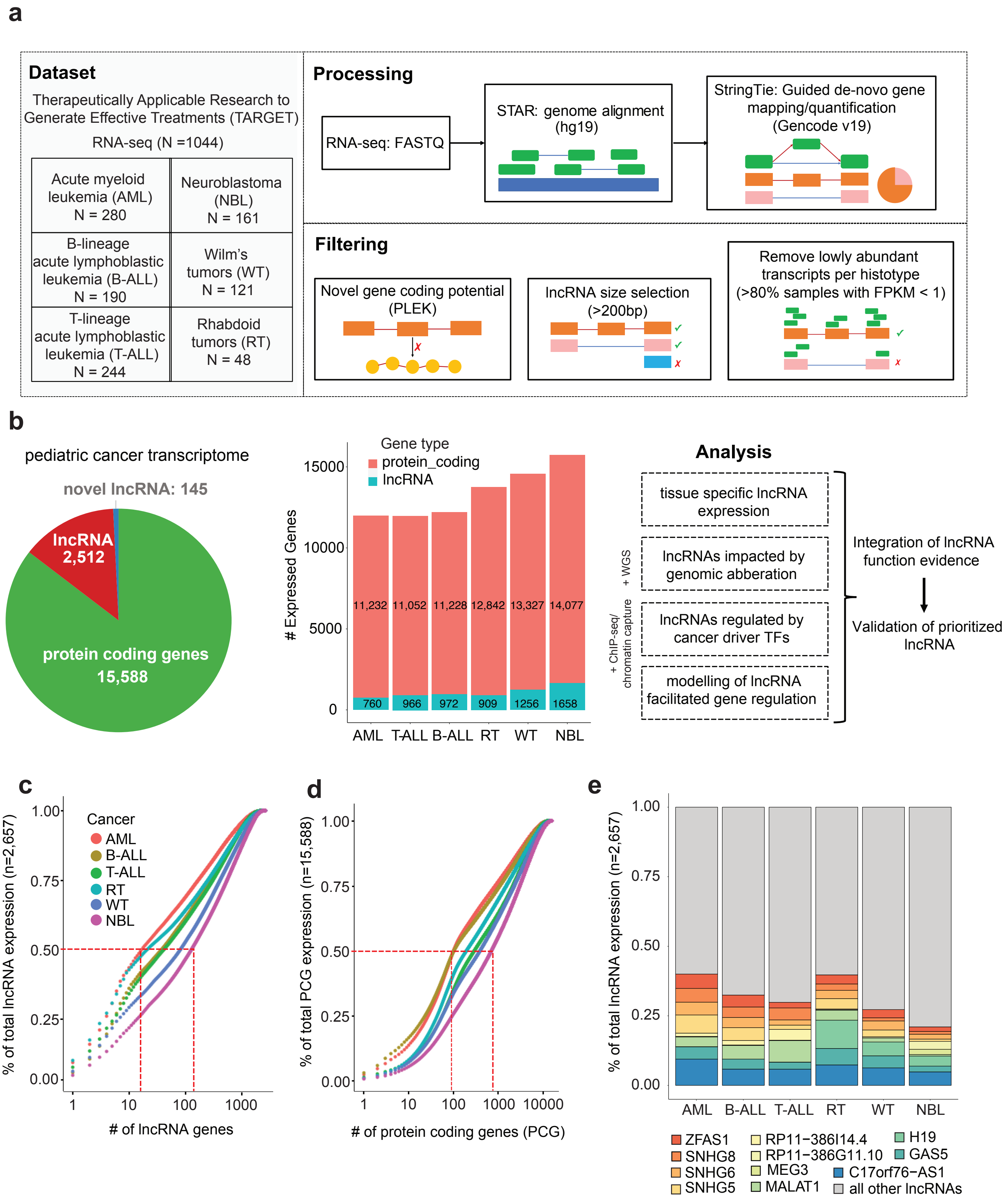
63. Zhao, X., et al., *CTCF cooperates with noncoding RNA MYCNOS to promote neuroblastoma progression through facilitating MYCN expression.* Oncogene, 2015: p. 1-12.
64. Ng, S.Y., et al., *The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis.* Molecular Cell, 2013. **51**: p. 349-359.
65. Tseng, Y.Y., et al., *PVT1 dependence in cancer with MYC copy-number increase.* Nature, 2014. **512**(7512): p. 82-6.
66. Jeon, Y. and J.T. Lee, *YY1 tethers Xist RNA to the inactive X nucleation center.* Cell, 2011. **146**(1): p. 119-33.
67. Li, Y., et al., *Identification and characterization of lncRNA mediated transcriptional dysregulation dictates lncRNA roles in glioblastoma.* Oncotarget, 2016. **7**: p. 45027-45041.
68. Delgado, M.D. and J. Leon, *Myc roles in hematopoiesis and leukemia.* Genes Cancer, 2010. **1**(6): p. 605-16.
69. Amaral, P.P., et al., *Genomic positional conservation identifies topological anchor point RNAs linked to developmental loci.* Genome Biol, 2018. **19**(1): p. 32.
70. Wang, M., et al., *Long noncoding RNA GAS5 promotes bladder cancer cells apoptosis through inhibiting EZH2 transcription.* Cell Death Dis, 2018. **9**(2): p. 238.
71. Zhao, Y., et al., *Long non-coding RNA (lncRNA) small nucleolar RNA host gene 1 (SNHG1) promote cell proliferation in colorectal cancer by affecting P53.* Eur Rev Med Pharmacol Sci, 2018. **22**(4): p. 976-984.
72. Kharabi Masouleh, B., et al., *Mechanistic rationale for targeting the unfolded protein response in pre-B acute lymphoblastic leukemia.* Proc Natl Acad Sci U S A, 2014. **111**(21): p. E2219-28.
73. Harenza, J.L., et al., *Transcriptomic profiling of 39 commonly-used neuroblastoma cell lines.* Sci Data, 2017. **4**: p. 170033.
74. Bourdoumis, A., et al., *The novel prostate cancer antigen 3 (PCA3) biomarker.* Int Braz J Urol, 2010. **36**(6): p. 665-8; discussion 669.
75. Slack, F.J. and A.M. Chinnaiyan, *The Role of Non-coding RNAs in Oncology.* Cell, 2019. **179**(5): p. 1033-1055.
76. Chen, Y., et al., *Core transcriptional regulatory circuitries in cancer.* Oncogene, 2020.
77. Zhang, X. and T.T. Ho, *Computational Analysis of lncRNA Function in Cancer.* Methods Mol Biol, 2019. **1878**: p. 139-155.
78. Scheibe, M., et al., *Quantitative mass spectrometry and PAR-CLIP to identify RNA-protein interactions.* Nucleic Acids Res, 2012. **40**(19): p. 9897-902.
79. Chu, C., et al., *Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions.* Mol Cell, 2011. **44**(4): p. 667-78.
80. Hon, C.C., et al., *An atlas of human long non-coding RNAs with accurate 5' ends.* Nature, 2017. **543**: p. 199-204.
81. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
82. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report.* Bioinformatics, 2016. **32**(19): p. 3047-8.
83. Zhang, X.O., et al., *Gene expression profiling of non-polyadenylated RNA-seq across species.* Genom Data, 2014. **2**: p. 237-41.
84. Upton, K., et al., *Epigenomic profiling of neuroblastoma cell lines.* Sci Data, 2020. **7**(1): p. 116.

85. 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.

86. 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.

87. 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.

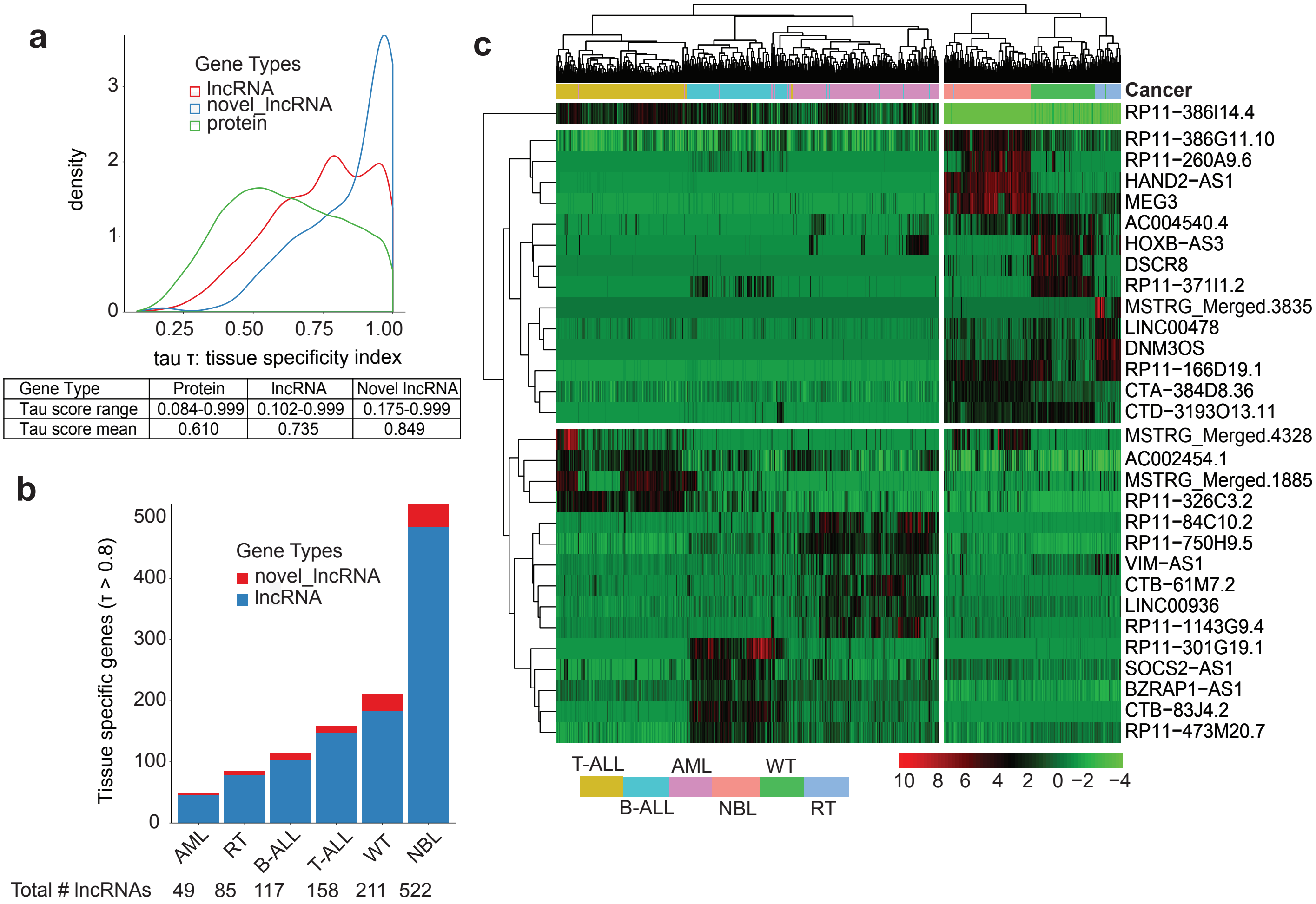
Figure 1



Acute myeloid leukemia (AML) N = 280	Neuroblastoma (NBL) N = 161
B-lineage acute lymphoblastic leukemia (B-ALL) N = 190	Wilm's tumors (WT) N = 121
T-lineage acute lymphoblastic leukemia (T-ALL) N = 244	Rhabdoid tumors (RT) N = 48

Fig 1: Pan-pediatric transcriptome characterization.
a. Overview of RNA-seq sample sizes for each histotype and schematic of data processing and filtering. Reads from RNA-seq fastq files were aligned using the STAR algorithm and then gene transcripts were mapped in a guided de-novo manner and quantified via the StringTie algorithm. Genes were considered novel if they did not have transcript exon structures matching genes in the Gencode v19 or RefSeq v74 databases. Novel genes were assigned as lncRNAs based on length >200bp and non-coding potential calculated using the PLEK algorithm. Transcripts with low expression (FPKM <1 in >80% samples per histotype) were not considered for further analysis. **b.** Pie graph showing the quantity of robustly expressed protein coding genes, Gencode/RefSeq annotated lncRNAs, and novel lncRNAs. Bar graph showing number of protein coding genes and lncRNAs expressed per cancer. Adjoining schematic gives overview of additional data types that were integrated with transcriptome data: WGS, ChIP-seq, and chromatin capture. Listed are the analyses used to elucidate which lncRNAs are likely to play functional roles in pediatric cancer. **c.** Cumulative expression plots comparing the number of lncRNAs and **d**. protein coding genes, respectively, that constitute the total sum of gene expression (FPKM) per pediatric cancer. **e.** Percentage of total lncRNA expression (FPKM) accounted for by the union of top 5 expressed lncRNAs per cancer (total 11 lncRNAs).

Figure 2



Gene Type	Protein	lncRNA	Novel lncRNA
Tau score range	0.084-0.999	0.102-0.999	0.175-0.999
Tau score mean	0.610	0.735	0.849

Fig 2: lncRNAs exhibit tissue specific expression that can distinguish cancers.
a. Tissue specificity index (tau score) which ranges from 0 (ubiquitously expressed) to 1 (tissue specific) is plotted for genes across three gene types: protein coding genes, lncRNAs, and novel lncRNAs. Table shows the tau score range and mean per gene type. **b.** Number of tissue specific known and novel lncRNAs in each cancer as defined by tissue specific gene threshold: tau score > 0.8. **c.** Heatmap showing the hierarchically clustered gene expression for the top 5 most highly expressed tissue specific lncRNAs per cancer. Samples from each cancer cluster together based on expression of these genes alone.

Figure 3

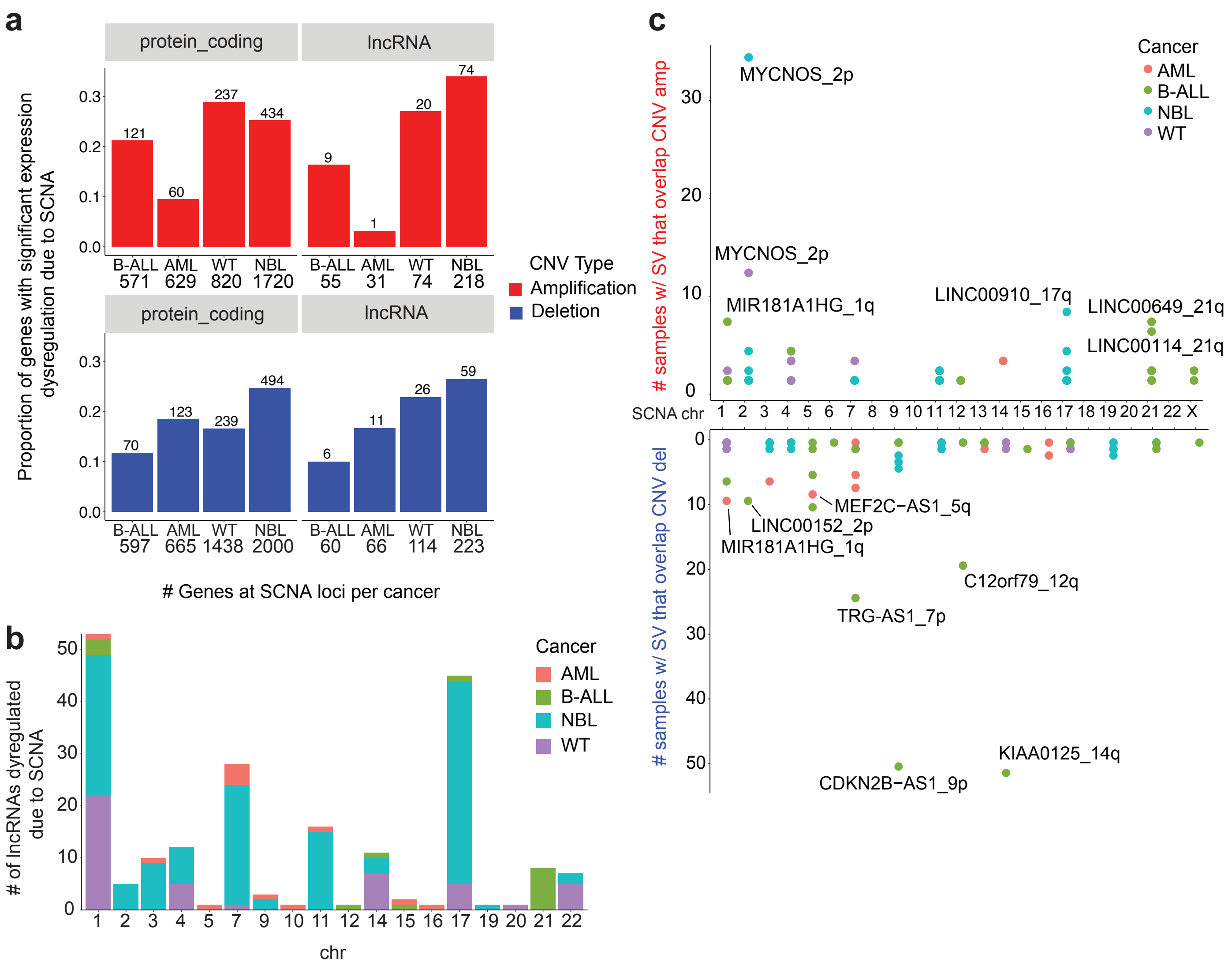


Fig 3: A similar proportion of lncRNAs and protein coding genes are dysregulated due to SCNA. **a.** The proportion of protein coding and lncRNA genes that have significant differential expression due SCNA, separated by copy number type (amplification or deletion). The number of genes found in SCNA loci is shown per cancer. Genes were evaluated to have differential expression due to copy number using the Wilcoxon rank sum test (p-value < 0.05) and log2 |fold change| > 1.5), comparing samples with no SCNA to samples with low/high SCNA as defined by GISTIC scores. **b.** The number of differentially expressed lncRNAs per chromosome and per cancer, distinguished by color. Chromosome 1 and 17 had the most dysregulated lncRNAs associating with the greater frequency of SCNA on these chromosomes across cancers. **c.** Number of samples with structural variant breakpoints in or near (+/- 2.5kb) lncRNAs and that are also located in copy number regions, stratified by amplification or deletion status of the locus.

Figure 4

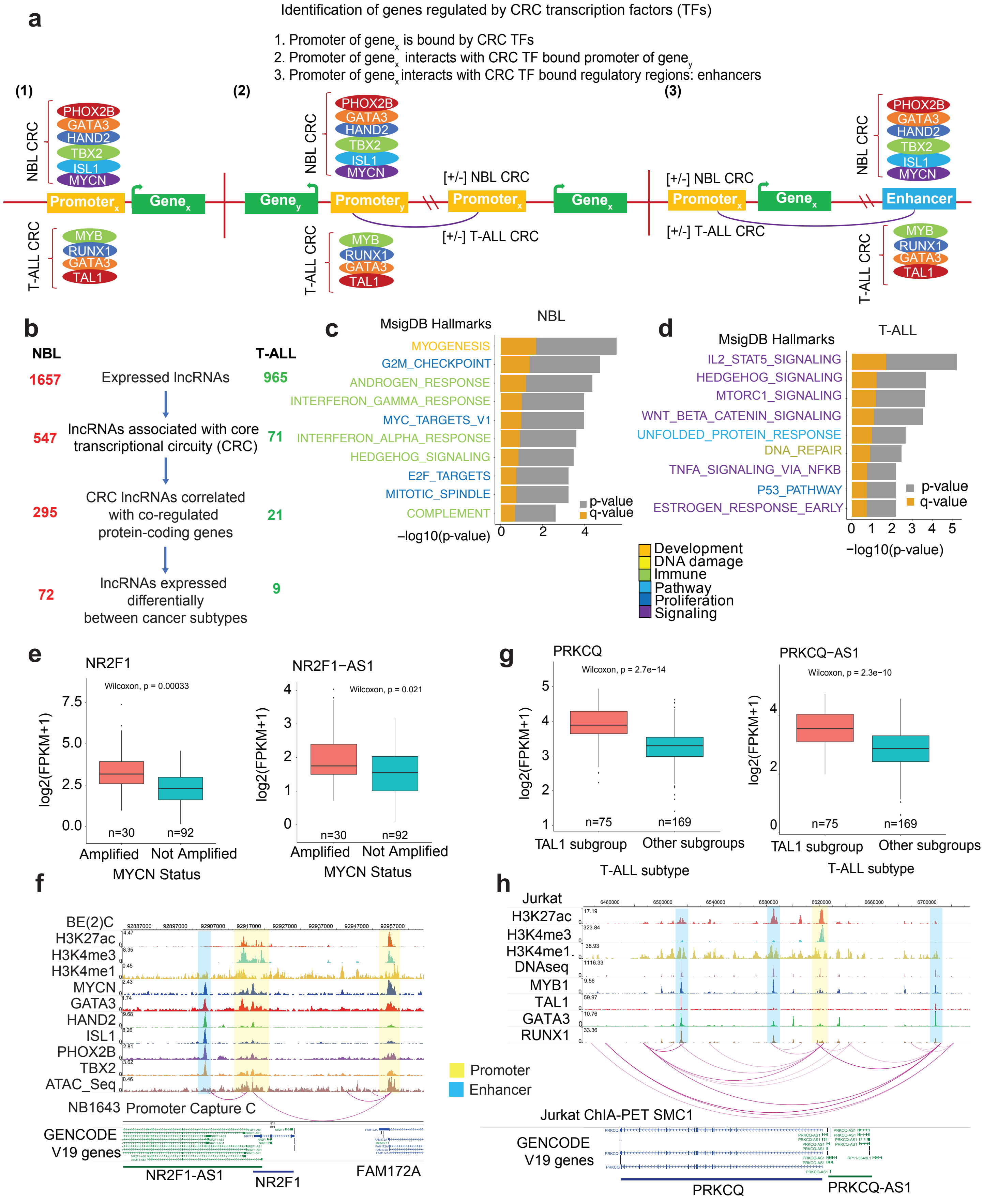

a
Identification of genes regulated by CRC transcription factors (TFs)
1. Promoter of gene$_x$ is bound by CRC TFs
2. Promoter of gene$_x$ interacts with CRC TF bound promoter of gene$_y$
3. Promoter of gene$_x$ interacts with CRC TF bound regulatory regions: enhancers
(1)
(2)
(3)
NBL CRC
PHOX2B
GATA3
HAND2
TBX2
ISL1
MYCN
Promoter$_x$
Gene$_x$
T-ALL CRC
MYB
RUNX1
GATA3
TAL1
Gene$_y$
Promoter$_y$
[+/-] NBL CRC
[+/-] T-ALL CRC
Enhancer
b
NBL
T-ALL
1657
Expressed lncRNAs
965
547
lncRNAs associated with core transcriptional circuity (CRC)
71
295
CRC lncRNAs correlated with co-regulated protein-coding genes
21
72
lncRNAs expressed differentially between cancer subtypes
9
c
MsigDB Hallmarks
NBL
MYOGENESIS
G2M_CHECKPOINT
ANDROGEN_RESPONSE
INTERFERON_GAMMA_RESPONSE
MYC_TARGETS_V1
INTERFERON_ALPHA_RESPONSE
HEDGEHOG_SIGNALING
E2F_TARGETS
MITOTIC_SPINDLE
COMPLEMENT
p-value
q-value
0
2
4
−log10(p-value)
d
MsigDB Hallmarks
T-ALL
IL2_STAT5_SIGNALING
HEDGEHOG_SIGNALING
MTORC1_SIGNALING
WNT_BETA_CATENIN_SIGNALING
UNFOLDED_PROTEIN_RESPONSE
DNA_REPAIR
TNFA_SIGNALING_VIA_NFKB
P53_PATHWAY
ESTROGEN_RESPONSE_EARLY
p-value
q-value
0 1 2 3 4 5
−log10(p-value)
Development
DNA damage
Immune
Pathway
Proliferation
Signaling
e
NR2F1
Wilcoxon, p = 0.00033
log2(FPKM+1)
7.5
5.0
2.5
0.0
n=30
n=92
Amplified
Not Amplified
MYCN Status
NR2F1−AS1
Wilcoxon, p = 0.021
4
3
2
1
0
n=30
n=92
Amplified
Not Amplified
MYCN Status
g
PRKCQ
Wilcoxon, p = 2.7e−14
log2(FPKM+1)
5
4
3
2
1
n=75
n=169
TAL1 subgroup
Other subgroups
T-ALL subtype
PRKCQ−AS1
Wilcoxon, p = 2.3e−10
4
2
0
n=75
n=169
TAL1 subgroup
Other subgroups
T-ALL subtype
f
BE(2)C
H3K27ac
H3K4me3
H3K4me1
MYCN
GATA3
HAND2
ISL1
PHOX2B
TBX2
ATAC_Seq
92887000
92897000
92907000
92917000
92927000
92937000
92947000
92957000
NB1643 Promoter Capture C
GENCODE V19 genes
NR2F1-AS1
NR2F1
FAM172A
Promoter
Enhancer
h
Jurkat
H3K27ac
H3K4me3
H3K4me1.
DNAseq
MYB1
TAL1
GATA3
RUNX1
6460000
6500000
6540000
6580000
6620000
6660000
6700000
Jurkat ChIA-PET SMC1
GENCODE V19 genes
PRKCQ
PRKCQ-AS1

Fig 4: Identification of cancer associated lncRNAs regulated by core transcription factors in NBL and T-ALL

a. Schematic of how CRC regulated genes are identified using ChIP-seq and chromatin interaction data. We identified lncRNAs based on three types of regulation. 1) CRC transcription factors binding directly at the promoter of the lncRNA. 2) CRC TFs bind an enhancer region that interacts with a lncRNA promoter. 3) CRC TFs bind the promoter of a different gene and this promoter interacts with a lncRNA promoter. CRC TF binding was identified from ChIP-seq data, while enhancer-promoter and promoter promoter interactions were identified from chromatin capture data. **b.** Filtering of lncRNAs expressed in either NBL or T-ALL based on CRC TF regulation, co-regulation with a CRC associated protein-coding gene, and differential expression based on cancer subtypes. In NBL, differentially expressed lncRNAs are between MYCN-amplified vs non-amplified samples. In T-ALL, differentially expressed lncRNAs are between the TAL1 subgroup vs other T-ALL sample subtypes. Co-regulation of CRC regulated lncRNAs and protein coding genes was determined by correlation analysis. **c-d.** Gene set enrichment analysis results for protein coding genes significantly correlated with CRC regulated lncRNAs (Pearson's r > 0.4 and FDR < 0.1) in NBL and T-ALL, respectively. **e.** Expression of *NR2F1* and *NR2F1-AS1* stratified by NBL sample MYCN amplification status **f.** ChIP-seq tracks for histone marks and CRC transcription factors in the NBL cell line: BE(2)C, and promoter capture C chromatin interactions in NBL cell line: NB1643, at the *NR2F1/NR2F1-AS1* locus. **g.** Expression of *PRKCQ* and *PRKCQ-AS1* stratified based on the TAL1 subgroup of T-ALL samples. **h**. ChIP-seq tracks for histone marks and CRC transcription factors and ChIA-PET chromatin interactions in the T-ALL cell line: Jurkat, at the *PRKCQ/ PRKCQ-AS1* locus.

Figure 5

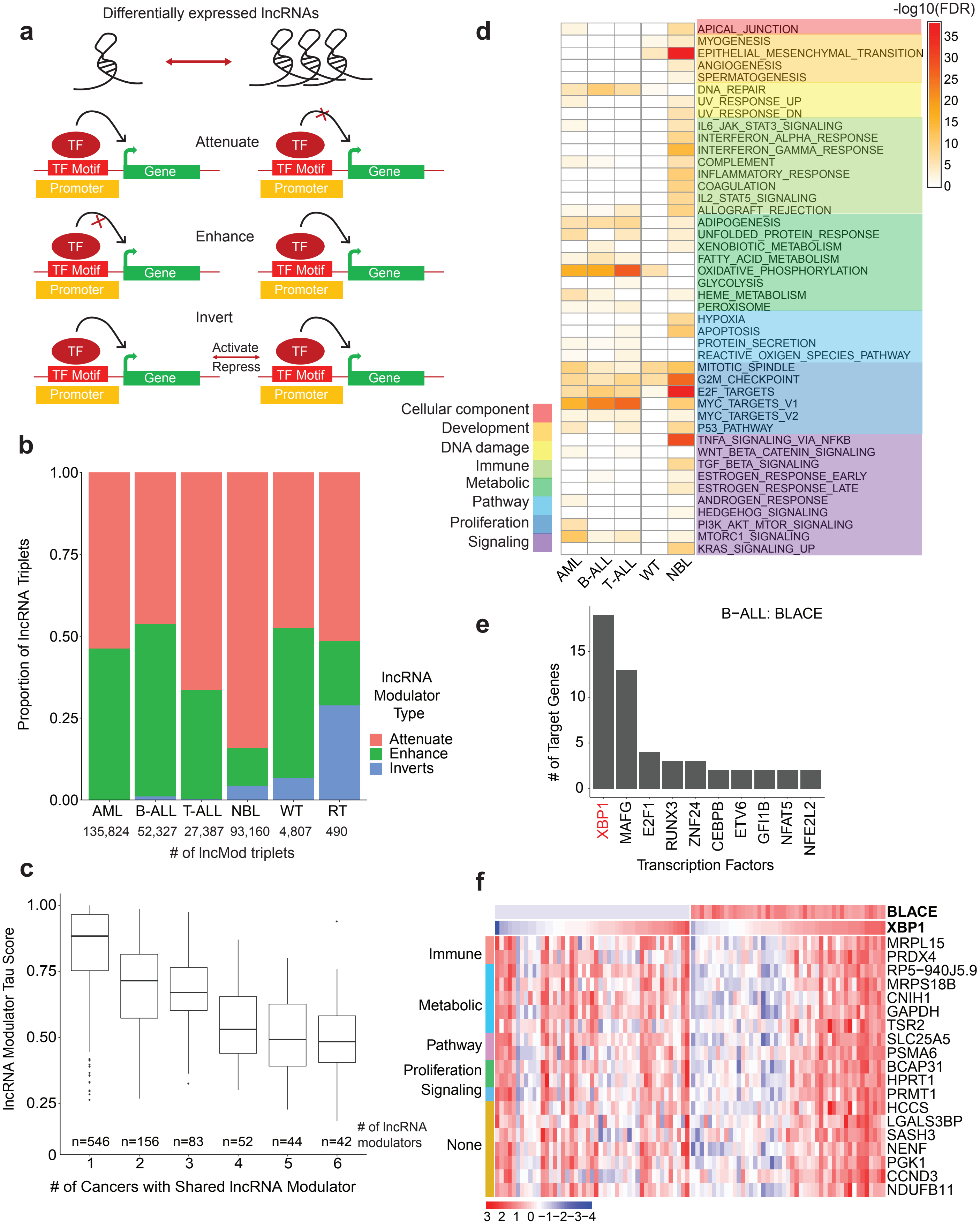

a
Differentially expressed lncRNAs
TF
TF Motif
Promoter
Gene
Attenuate
Enhance
Invert
Activate
Repress
b
1.00
0.75
0.50
0.25
0.00
Proportion of lncRNA Triplets
AML
B-ALL
T-ALL
NBL
WT
RT
135,824
52,327
27,387
93,160
4,807
490
# of lncMod triplets
lncRNA Modulator Type
Attenuate
Enhance
Inverts
c
1.00
0.75
0.50
0.25
0
lncRNA Modulator Tau Score
n=546
n=156
n=83
n=52
n=44
n=42
# of lncRNA modulators
1
2
3
4
5
6
# of Cancers with Shared lncRNA Modulator
d
-log10(FDR)
35
30
25
20
15
10
5
0
APICAL_JUNCTION
MYOGENESIS
EPITHELIAL_MESENCHYMAL_TRANSITION
ANGIOGENESIS
SPERMATOGENESIS
DNA_REPAIR
UV_RESPONSE_UP
UV_RESPONSE_DN
IL6_JAK_STAT3_SIGNALING
INTERFERON_ALPHA_RESPONSE
INTERFERON_GAMMA_RESPONSE
COMPLEMENT
INFLAMMATORY_RESPONSE
COAGULATION
IL2_STAT5_SIGNALING
ALLOGRAFT_REJECTION
ADIPOGENESIS
UNFOLDED_PROTEIN_RESPONSE
XENOBIOTIC_METABOLISM
FATTY_ACID_METABOLISM
OXIDATIVE_PHOSPHORYLATION
GLYCOLYSIS
HEME_METABOLISM
PEROXISOME
HYPOXIA
APOPTOSIS
PROTEIN_SECRETION
REACTIVE_OXIGEN_SPECIES_PATHWAY
MITOTIC_SPINDLE
G2M_CHECKPOINT
E2F_TARGETS
MYC_TARGETS_V1
MYC_TARGETS_V2
P53_PATHWAY
TNFA_SIGNALING_VIA_NFKB
WNT_BETA_CATENIN_SIGNALING
TGF_BETA_SIGNALING
ESTROGEN_RESPONSE_EARLY
ESTROGEN_RESPONSE_LATE
ANDROGEN_RESPONSE
HEDGEHOG_SIGNALING
PI3K_AKT_MTOR_SIGNALING
MTORC1_SIGNALING
KRAS_SIGNALING_UP
Cellular component
Development
DNA damage
Immune
Metabolic
Pathway
Proliferation
Signaling
AML
B-ALL
T-ALL
WT
NBL
e
B−ALL: BLACE
15
10
5
0
# of Target Genes
XBP1
MAFG
E2F1
RUNX3
ZNF24
CEBPB
ETV6
GFI1B
NFAT5
NFE2L2
Transcription Factors
f
BLACE
XBP1
Immune
Metabolic
Pathway
Proliferation
Signaling
None
MRPL15
PRDX4
RP5−940J5.9
MRPS18B
CNIH1
GAPDH
TSR2
SLC25A5
PSMA6
BCAP31
HPRT1
PRMT1
HCCS
LGALS3BP
SASH3
NENF
PGK1
CCND3
NDUFB11
3 2 1 0 −1−2−3−4

Fig 5: lncRNA modulators impact transcriptional networks involving proliferation.
a. Schematic that shows the three ways (attenuate, enhance, or invert) in which differentially expressed lncRNA modulators can impact transcription factor and target gene relationships. lncRNA modulators are associated with a TF-target gene pair based on a significant difference between TF-target gene expression correlation in samples with low lncRNA expression (lowest quartile) vs samples with high lncRNA expression (highest quartile). **b.** The proportion of lncRNA modulator types associated with significantly dysregulated lncRNA modulator- TF-target gene (lncMod) triplets. The number of significantly dysregulated lncMod triplets is listed per cancer. **c.** Number of lncRNA modulators genes that are common in lncMod triplets across cancers. Pan-cancer lncRNA modulator genes tend to have a lower tau score compared to lncRNA modulators only associated with one cancer. **d.** Gene set enrichment using the MsigDB Hallmark gene set, of target genes associated with lncRNA modulators in each cancer (Fisher exact test, FDR < 0.1). **e.** Transcription factors associated with the B-ALL expression specific lncRNA, *BLACE,* ranked based on number of regulated target genes. **f**. Expression heatmap of *BLACE* and the target genes of the *XBP1* transcription factor, grouped by associated hallmark gene set, in samples that fall within the bottom and top quartiles of *BLACE* expression across all B-ALL samples.

Figure 6

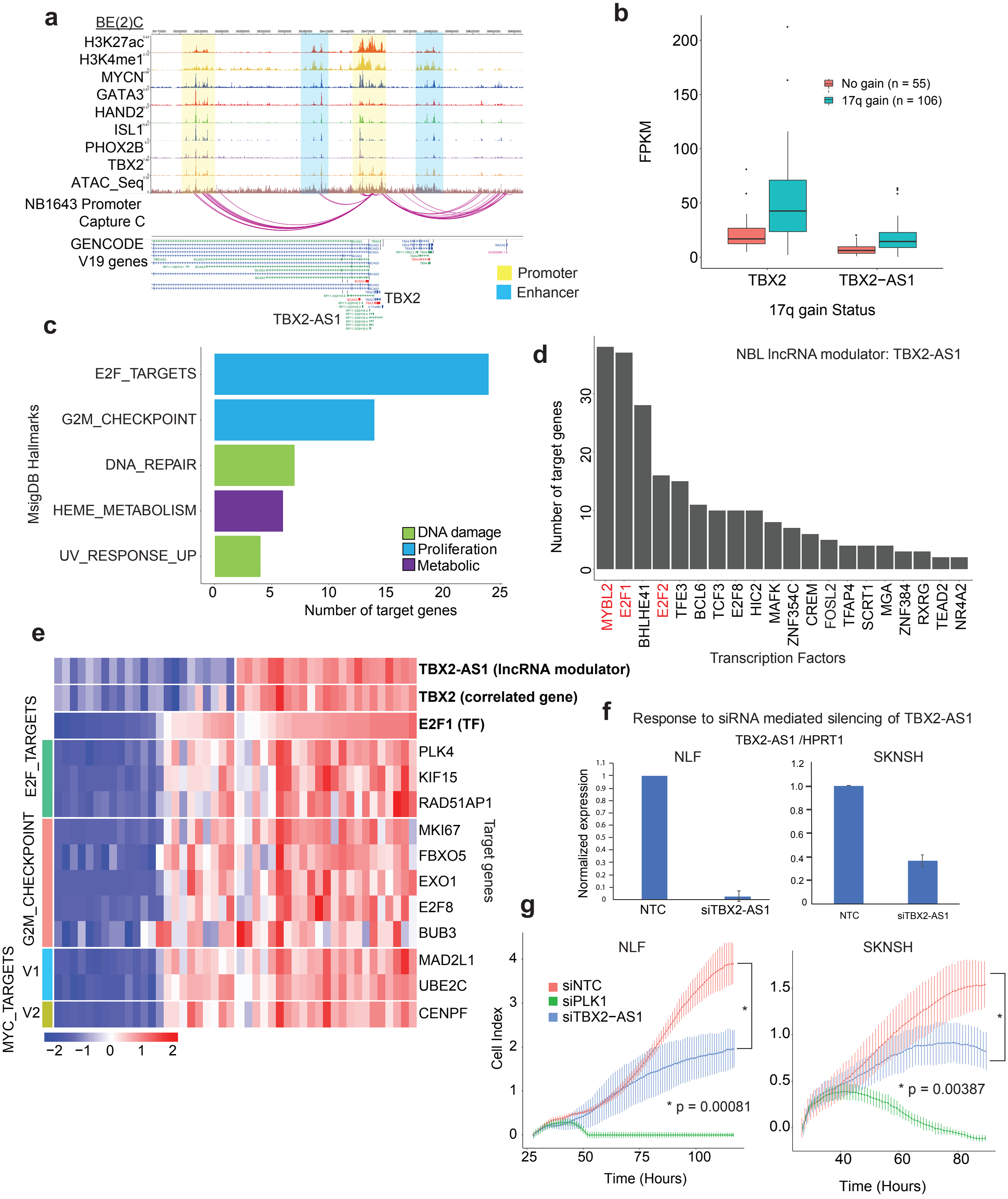

a
BE(2)C
H3K27ac
H3K4me1
MYCN
GATA3
HAND2
ISL1
PHOX2B
TBX2
ATAC_Seq
NB1643 Promoter Capture C
GENCODE V19 genes
TBX2
TBX2-AS1
Promoter
Enhancer
b
FPKM
200
150
100
50
0
No gain (n = 55)
17q gain (n = 106)
TBX2
TBX2−AS1
17q gain Status
c
MsigDB Hallmarks
E2F_TARGETS
G2M_CHECKPOINT
DNA_REPAIR
HEME_METABOLISM
UV_RESPONSE_UP
DNA damage
Proliferation
Metabolic
0
5
10
15
20
25
Number of target genes
d
NBL lncRNA modulator: TBX2-AS1
Number of target genes
0
10
20
30
MYBL2
E2F1
BHLHE41
E2F2
TFE3
BCL6
TCF3
E2F8
HIC2
MAFK
ZNF354C
CREM
FOSL2
TFAP4
SCRT1
MGA
ZNF384
RXRG
TEAD2
NR4A2
Transcription Factors
e
TBX2-AS1 (lncRNA modulator)
TBX2 (correlated gene)
E2F1 (TF)
E2F_TARGETS
G2M_CHECKPOINT
MYC_TARGETS
V1
V2
PLK4
KIF15
RAD51AP1
MKI67
FBXO5
EXO1
E2F8
BUB3
MAD2L1
UBE2C
CENPF
Target genes
−2
−1
0
1
2
f
Response to siRNA mediated silencing of TBX2-AS1
TBX2-AS1 /HPRT1
NLF
SKNSH
Normalized expression
NTC
siTBX2-AS1
g
NLF
SKNSH
Cell Index
siNTC
siPLK1
siTBX2−AS1
* p = 0.00081
* p = 0.00387
Time (Hours)
25
50
75
100
40
60
80

Fig 6: The *TBX2-AS1* lncRNA plays a role in neuroblastoma proliferation
a. Chromatin interactions and ChIP-seq tracks for NBL-CRC transcription factors in the NBL cell lines, Be(2)C and NB1643, at the *TBX2/TBX2-AS1* locus. **b.** Expression of *TBX2* and *TBX2-AS1* in NBL tumor samples with and without 17q gain. **c.** The top MsigDB Hallmarks enriched across targets genes (p-value < 0.01) regulated by *TBX2-AS1* as predicted using the lncMod analysis. **d.** The transcription factors with most target genes regulated by *TBX2-AS1* as predicted from lncMod analysis. **e.** Expression of gene targets of the E2F1 transcription factor that are enriched for proliferation hallmarks, in samples with low and high *TBX2* and *TBX2-AS1* expression. *TBX2* expression is highly correlated with that of *TBX2-AS1* (Pearson's r=0.77). **f.** siRNA knockdown efficiency of TBX2-AS1 in the NBL cell line: NLF is 98% and in the SKNSH cell 63% knockdown was achieved. **g.** Representative image of cell growth (as measured by RT-Ces assay) of the NBL cell lines: NLF and SKNSH. Cell index is normalized to time point when siRNA reagent is added at 24 hours post cell plating.