Bioinformatics, 34, 2018, i664–i670
doi: 10.1093/bioinformatics/bty574
ECCB 2018

OXFORD

CisPi: a transcriptomic score for disclosing cis-acting disease-associated lincRNAs

Zhezhen Wang, John M. Cunningham and Xinan H. Yang*

Department of Pediatrics, University of Chicago, Chicago, IL 60637, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Long intergenic noncoding RNAs (lincRNAs) have risen to prominence in cancer biology as new biomarkers of disease. Those lincRNAs transcribed from active cis-regulatory elements (enhancers) have provided mechanistic insight into cis-acting regulation; however, in the absence of an enhancer hallmark, computational prediction of cis-acting transcription of lincRNAs remains challenging. Here, we introduce a novel transcriptomic method: a cis-regulatory lincRNA–gene associating metric, termed 'CisPi'. CisPi quantifies the mutual information between lincRNAs and local gene expression regarding their response to perturbation, such as disease risk-dependence. To predict risk-dependent lincRNAs in neuroblastoma, an aggressive pediatric cancer, we advance this scoring scheme to measure lincRNAs that represent the minority of reads in RNA-Seq libraries by a novel side-by-side analytical pipeline.

Results: Altered expression of lincRNAs that stratifies tumor risk is an informative readout of oncogenic enhancer activity. Our CisPi metric therefore provides a powerful computational model to identify enhancer-templated RNAs (eRNAs), eRNA-like lincRNAs, or active enhancers that regulate the expression of local genes. First, risk-dependent lincRNAs revealed active enhancers, over-represented neuroblastoma susceptibility loci, and uncovered novel clinical biomarkers. Second, the prioritized lincRNAs were significantly prognostic. Third, the predicted target genes further inherited the prognostic significance of these lincRNAs. In sum, RNA-Seq alone is sufficient to identify disease-associated lincRNAs using our methodologies, allowing broader applications to contexts in which enhancer hallmarks are not available or show limited sensitivity.

Availability and implementation: The source code is available on request. The prioritized lincRNAs and their target genes are in the Supplementary Material.

Contact: xyang2@uchicago.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Long intervening noncoding RNAs that do not overlap with protein-coding loci (lincRNAs) are found in evolutionarily conserved intergenic regions (GENCODE definition), and are expressed in a context-dependent manner, suggesting that they have functional relevance. While several lincRNAs have been shown to modulate gene expression and cell growth in cancers (Fernando *et al.*, 2017; Russell *et al.*, 2015), discerning and characterizing the cis-acting loci from thousands of transcribed lincRNAs is challenging. The majority of known functional lincRNAs emanate from active enhancers (Vance and Ponting, 2014); however, in the absence of an associated enhancer hallmark, computational prediction of cis-acting loci is a difficult task.

Owing to the growth in high-throughput sequencing data, adding the global contribution of lincRNA expression to the information flow from genotype to phenotype *in silico* is possible (Signal *et al.*, 2016). Currently, computational predictors of cis-acting activity of lincRNA are based on their association with either upstream DNAs or downstream genes. Upstream DNA-based identification methods assume that lincRNA expression level is a readout of enhancer activity. These methods focus on lincRNAs localizing to either disease variations (Signal *et al.*, 2016) or enhancer hallmarks. Although they are being widely applied, the essential limitation of these methods is their reliance on a high-affinity ChIP antibody to measure enhancer activity in a specific condition and cell type. On the other hand, downstream gene-based models follow the 'guilt-by-association' rule by which stimulus-dependent lincRNA abundance (Orom *et al.*, 2010) or positive correlations of expression levels between lincRNAs and their target genes (Iyer *et al.*, 2015) are prioritized. However, in highly epigenetically altered diseases, such as

cancer, pervasive transcription may occur as a byproduct (Signal *et al.*, 2016); thus, correlated expression levels with downstream genes may not imply cis-acting disease-association of the lincRNA loci.

In addition to the two types of lincRNA identification described above, layering information with seemingly independent statistical evidence increases the predictive power. For example, analyzing context-dependent expression of the long noncoding RNA (lncRNA) transcriptome, the mRNA transcriptome, and accessible chromatin, our 'mutually informative' method previously discovered cis-regulatory DNA elements from dysregulated lincRNAs (Yang *et al.*, 2017a). While successful, all of these approaches overlook lincRNAs that display mRNA-like features such as capped and polyadenylated post-transcriptional procession (Li *et al.*, 2016; Ulitsky and Bartel, 2013).

Additionally for genome-wide prediction, researchers have identified lincRNAs together with mRNAs from the same analytical pipeline (Iyer *et al.*, 2015). However, as the median lincRNA level is typically a tenth that of the median mRNA level (Ulitsky and Bartel, 2013), more sensitive prediction can arise from a pipeline that is specifically designed for identifying and measuring these lincRNAs.

Using the pipeline to predict transcription-regulating cis-acting and disease-associated lincRNAs, we designed a quantitative scoring system called 'CisPi'. Given our familiarity with neuroblastoma (NB) that has both published RNA-Seq profiling of a large cohort and enhancer landscapes (Yang *et al.*, 2017a,c), we tested this scoring system on the segregation of NB risk groups. Neuroblastoma is by far the most common cancer in infants (less than 1 year old), with about 800 new cases each year in the United States. While a child with non-high-risk (NHR) neuroblastoma can often be cured, the 5-year survival rate in children in the high-risk (HR) group is less than 50%. We therefore expect the lincRNAs expressed in a risk-dependent manner, and their target genes, to provide cis-regulatory insights into NB biology.

In this study, we introduce a novel cis-regulatory association scoring scheme to predict regulatory lincRNAs by assessing mutual biological relevance between lincRNAs and their target genes. Using an early version of this scoring system, we have previously discovered a cis-acting lincRNA relevant to cardiology (Yang *et al.*, 2017a). Here, we enhanced this methodology in three ways. First, we developed a novel side-by-side analytical pipeline for RNA-Seq data to measure lincRNAs with relatively low expression levels, since protein-coding transcripts are predominant in polyadenylated RNA libraries (Supplementary Fig. S1). Second, a dynamic cis-regulatory distance was estimated from the topologically associating domains (TADs) derived from Hi-C experiments rather than an arbitrarily fixed window around every lincRNA (Wang *et al.*, 2015). This rationale is derived from observations of spatial interactions between distant chromatin regions encoding lincRNA and their putative target genes (Cai *et al.*, 2016). Third, all risk-dependent genes, positively or negatively coherent with a risk-dependent lincRNA, were modeled to quantify the cis-regulatory activity of the lincRNA. Then, employing a robust individualized prognostic approach, we demonstrated the clinical application of prioritized lincRNAs and target genes using these three enhancements to our CisPi scoring system. These approaches are independent of experimental context thus can be applied broadly to the enormous amount of previously generated RNA-Seq, allowing new discovery in functional genomics.

2 Materials and methods

2.1 Multi-layer sequencing data collection

To identify risk-dependent polyadenylated noncoding transcripts, we obtained publicly-available RNA-Seq profiling of 68 selected primary tumors from TARGET (tumor > 60%, pair-end sequencing reads > 50M per sample) (Pugh *et al.*, 2013) (Supplementary Table S2). As a control, we also collected total RNA-Seq profiling of independent 498 primary NB patients which we termed as the 'Germany' dataset (tumor > 60%, sequencing depth around 100 M, Supplementary Table S1). A landscape of TADs in neuroblastoma cells was obtained from the 3D Genome Browser (Wang *et al.*, 2017).

2.2 CisPi-score to quantify coherent biological-relevance of lincRNAs and mRNAs

We hypothesized that there is a direct relationship between lincRNAs and their target genes leading to an expression difference between distinct clinical phenotypes: the HR and NHR groups. Therefore, we postulated that a risk-dependent regulatory linkage between a lincRNA and its target genes is measurable from the coherence of their expression levels. To address this two-feature question, we took advantage of a single-layer Pi-score that prioritizes gene-features by conjoining two-dimensional measurements (the fold change and statistical significance, ie, the *P*-value) into a one-dimensional estimation of biological relevance (Xiao *et al.*, 2014).

To model associated risk-dependence between lincRNAs and genes, we designed a cis-regulatory lincRNA–gene-association metric, termed 'CisPi'. We first calculated a $\pi.lincRNA$-score to prioritize each risk-dependent lincRNA i for its biological relevance to the disease risk (Eq. 1), where φ_i is the side effect (log fold-change of expression levels between HR and NHR patients) and P_i is the significance. We then calculated another $\pi.gene$-score to estimate the biological relevance of each of the two possibly risk-dependent gene x that sit adjacent to the lincRNA (Eq. 2). Following the 'guilt-by-association' rule, in case both adjacent genes were risk-dependent, the one with the absolute highest π value was chosen to represent the target of this lincRNA (Eq. 3). To further associate $\pi.lincRNA$ to $\pi.gene$, we build a CisPi-score to incorporate mutual risk-dependence of a lincRNA i and its putative target genes (Eq. 4), where the |.| is an absolute value operator and the sign() function extracts the directionality of risk-dependence:

$$\pi^i_{lincRNA} = [\varphi_i \cdot (-\log_{10} P_i)] \tag{1}$$

$$\pi^x_{gene} = [\varphi_x \cdot (-\log_{10} P_x)] \tag{2}$$

$$\pi^i_{target} = [\underset{x \in i\ targets}{\mathbf{argmax}}\{|\pi^x_{\text{gene}}|\}] \tag{3}$$

$$cis\pi^i = [\pi^i_{lincRNA} + \text{sign}(\pi^i_{target}) \cdot \pi^i_{target}]/2 \tag{4}$$

The max function further allowed us to compare the CisPi-metric assuming adjacent gene targeting of the lincRNA with another CisPi-metric that assumes distal targeting within a TAD. With the 'distal targeting' assumption, all genes residing in the same TAD (Won *et al.*, 2016) of a lincRNA were putative target genes to estimate the $\pi.target$ value (Eq. 3) for a CisPi-score (in Eq. 4). These two types of CisPi calculations will be equal when an adjacent gene has the highest $\pi.target$ value.

2.3 Relative expression analysis with lincRNA-set pairs

To bridge lincRNA expression levels with the clinic, we designed an individualized prognostic predictor: a Relative eXpression Analysis with LincRNA-Set Pairs (RXA-LSP) indicator. Given two sets of risk-dependent lincRNAs (upregulated and down-regulated), the RXA-LSP calculated the divergence of risk-dependence between the

two sets for each tumor. Thus, a higher indicator indicates a more unfavorable outcome. We have shown the robustness of this model on Gene-set pairs, which we previously termed RXA-GSP (Yang *et al.*, 2013; Yang *et al.*, 2017a,c). Importantly, this model of relative expression is technique-independent and population-independent.

2.4 RNA-Seq data analyses

Our overall analytic efforts were focused on the identification of noncoding RNAs that represent the minority of reads in polyadenylated RNA-Seq libraries. Using the publicly available RNA-Seq profiles (TARGET, n = 68, Supplementary Table S1), we assembled and defined 96.2k expressed transcripts (CPM > 0 in at least 10% of patients). For these transcripts, we then defined seven biotypes of lncRNAs, three types of coding transcripts and microRNAs (Supplementary Fig. S2a). Among those 60.9k expressed lncRNA transcripts, the vast majority (48.5k) were de novo transcripts, and the 2nd largest group (3.8k) were lincRNA transcripts of which at least 50% overlapped GENCODE-annotated lincRNAs (Supplementary Fig. S2b). Both de novo and lincRNA transcripts demonstrated low protein-coding potential (Supplementary Fig. S2c, the yellow and orange line, respectively), further suggesting that these represent polyadenylated lncRNAs.

Between-group differential expression was considered as significant at fold-change > 2 and false discovery rate (FDR) < 0.05 (Supplementary Methods).

2.5 Active enhancer candidates

We collected active enhancer candidates from three independent resources: 65k predicted human enhancers based on CAGE (Cap Analysis of Gene Expression)-Seq (FANTOM5) (de Hoon *et al.*, 2015), 785 validated active enhancers in neural tissues (VISTA) (Visel *et al.*, 2007), and 27k aggregate enhancers in NB identified using ChIP-Seq data of three canonical hallmarks of active enhancers (H3K27ac, H3K4me1 and P300, Supplementary Fig. S3a).

3 Results

3.1 Olig(dT)-primed RNA-Seq can capture polyadenylated lincRNAs with transcriptional regulatory potential

3.1.1 Expressed lincRNAs were significantly enriched for hallmarks of active enhancers in NB tumors

We first demonstrated that the Olig(dT)-primed polyadenylated RNA libraries, which were originally designed to measure mRNAs, can also be used to measure lincRNAs, a specific biotype of lncRNAs (Supplementary Fig. S2).

We then asked whether these polyadenylated lncRNAs are enriched for canonical functional noncoding loci involved in transcriptional regulation. Two out of seven biotypes of lncRNAs, lincRNAs and the 'PC_antisense' transcripts, were enriched for aggregate enhancers (Fig. 1a, Supplementary Fig. S3a–b). These observations extend the hypothesis that the production of lncRNAs could be indicative of enhancer function, from solely non-polyadenylated RNAs (Yang *et al.*, 2017a) to include polyadenylated RNAs as well.

3.1.2 Risk-dependent lincRNAs recaptured non-coding functional candidates identified by canonical hallmarks of active enhancers

As many cis-regulatory lincRNAs are transcribed from active enhancers and are highly tissue-specific and perturbation-dependent (Azofeifa *et al.*, 2018), we sought to determine whether lincRNAs that show risk-dependent expression could identify NB-associated enhancers. Applying our side-by-side analytical pipeline (Fig. 1b),

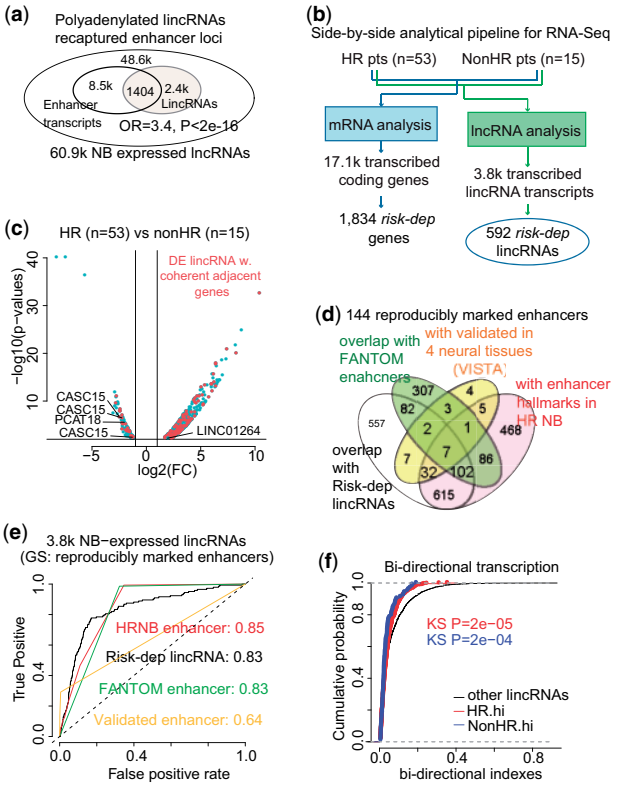

Fig. 1. Olig(dT)-primed RNA-Seq can capture functional noncoding loci. **(a)** Venn-diagram showing an enrichment between enhancer-marked transcripts and lincRNAs among all expressed lncRNAs. **(b)** Schematic illustrating a side-by-side RNA-Seq analytical pipeline designed for lincRNA signatures. **(c)** Volcano plot of altered expression of the risk-dependent lincRNAs [Fold change (FC) > 2, FDR < 0.05]; dot-color codes their coherence with adjacent genes and noted identifications were residing at NB susceptibility loci. **(d)** Venn-diagram displaying 144 enhancer candidates, being marked by at least 3 out of 4 methods, which were used as a 'proxy gold standard.' **(e)** ROC plots and AUC values comparing each enhancer-predicting method with the gold standard. **(f)** Empirical cumulative distribution plot of all 3.8k lincRNAs for their bi-directional indexes

we identified 592 significantly risk-dependent lincRNAs (Fig. 1c). These lincRNAs co-localized with 1404 enhancer candidates, and over half of these lincRNA-captured enhancer candidates (847) were independently evaluated (Fig. 1d). Given a global evaluation ratio of 6% (144 out of 2278 loci) among the four methods, risk-dependent lincRNAs indicate high cis-regulatory potential.

We observed an equally good predictive performance for active enhancer loci from two established methods, CAGE-seq and histone ChIP-seq, in comparison with NB risk-dependent lincRNA expression, with a performance score (AUC) ranging between 0.83 and 0.85 (Fig. 1e, Supplementary Method). The 785 validated active enhancers in neural tissues showed relatively lower performance (AUC = 0.64), which might arise from the restriction to elements with extreme sequence conservation, which was required for *in vivo* experiments but is not necessarily a required feature of functional cis-regulatory loci. We conclude that although most enhancer-templated noncoding RNAs are non-polyadenylated, polyadenylated RNA-Seq data is still capable of discovering enhancer-templated lincRNAs.

3.1.3 Risk-dependent lincRNAs tend to be bi-directional, long non-coding transcripts distal to coding genes

We then asked about the genomic features of risk-dependent lincRNAs. They tended to be long transcripts with a relative

enriched peak width of ~10kbp, and locate at a larger distance away from any coding-gene-TSS than other lincRNA transcripts (Supplementary Fig. S3c). These risk-dependent lincRNAs showed a significant preference for bi-directional transcription (Kolmogorov-Smirnov test $P < 2e-4$, Fig. 1f, Supplementary Method). Bi-directional expression at distal enhancers has been widely described (Kim *et al.*, 2010), in support of the hypothesis that risk-dependent lincRNAs represent regulatory enhancers affecting transcription.

3.2 Risk-dependent lincRNAs mark tumor-associated and tissue-specific enhancers

3.2.1 Risk-dependent lincRNAs reveal known and novel tumor susceptibility loci

We examined whether disease-specific or tissue-specific enhancers could be identified by risk-dependent alteration of lincRNA transcription. There exist 163 strong linkage-disequilibrium (LD) blocks at suggestive neuroblastoma-susceptibility loci ($P < 5e-5$, the GRASP repository) (Eicher *et al.*, 2015), using PLINK 1.9 (MAF>=≥0.05) (Chang *et al.*, 2015). 592 risk-dependent lincRNAs significantly captured seven of these NB susceptibility loci (Fig. 2a, FET $P = 0.004$, OR = 4.7; empirical $P < 0.001$, Supplementary Method). Note that lincRNA was the only biotype of lncRNAs that showed low coding potential, but over-represented NB susceptibility loci, confirming that disease risk-dependent alteration of lincRNAs are biologically relevant.

The risk-dependent lincRNAs revealed seven NB susceptibility loci, of which two were known tumor-suppressive lincRNAs at 6p22 and others were new functional lincRNA candidates (Supplementary Table S3). The NB suppressive roles of HR-downregulated lincRNAs (NBAT1 or CASC14, CASC15) have been extensively reported (Maris *et al.*, 2008; Mondal *et al.*, 2018; Russell *et al.*, 2015; Yao *et al.*, 2017), due to their significant NB-association derived from GWAS studies and aggregate hallmarks for enhancer activity (Fig. 2b1). Although identified with only a suggestive *P*-value (<5e−5) from GWAS (Fig. 2a3), we now identify a new oncogenic HR-upregulated lincRNA LINC01264 that is characterized by similar aggregate hallmarks for enhancer activity and is adjacent to the proto-oncogene RET (Fig. 2b2), which we previously defined as one of 2858 gene biomarkers up-regulated in HR NB (Yang *et al.*, 2017b). Another HR-downregulated lincRNA, produced from the same promoter as AQP4 but in the antisense direction, could also play a regulatory role in NB cell proliferation, as decreased AQP4 mRNA levels lead to an increase in both caspase activation and cell death within SHSY5Y neuroblastoma cells (Esposito *et al.*, 2008).

Collectively, these observations and the existing literature support two intriguing arguments: polyadenylated lincRNAs with risk-dependent expression may mark tumor-associated enhancers, and polyadenylated lincRNAs provide additional evidence to support suggestive GWAS findings in noncoding regions.

3.2.2 Risk-dependent lincRNAs are over-represented in validated neural enhancers

We next investigated the tissue-specificity of the lincRNA-captured enhancer candidates, using the 785 enhancers whose activity were positively evaluated in any of 22 tissues within the VISTA Enhancer Browser database (Visel *et al.*, 2007). Risk-dependent lincRNAs at the VISTA enhancers in four neural tissues showed preference for risk-dependence (empirical *P*-values from 0.1 to 0.001), with the midbrain being the highest (Fig. 2c, Supplementary Fig. S4). In contrast, the empirical *P*-values of the all other tissues were insignificant

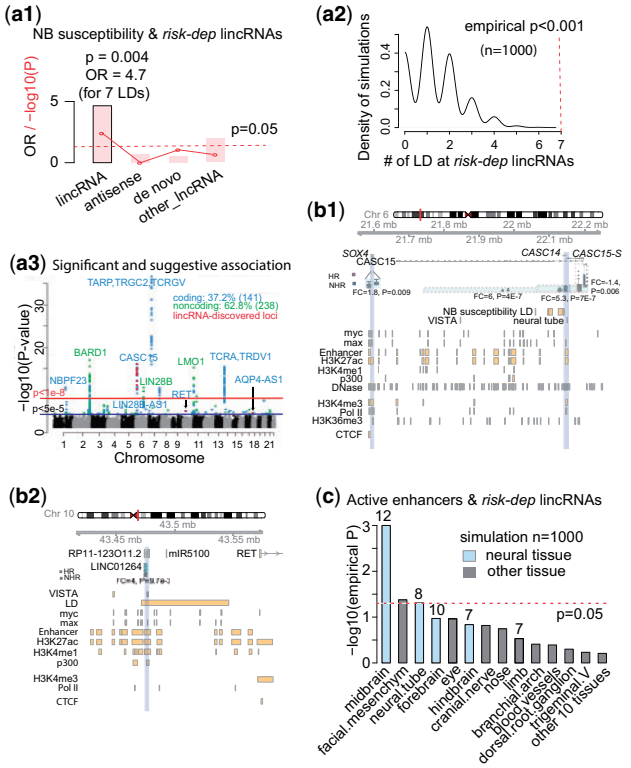

Fig. 2. Risk-dependent lincRNAs mark tumorigenesis and tissue-specific enhancers. **(a)** The theoretical (panel 1) and empirical evaluation (panel 2) for NB-susceptibility (panel 3). The suggestive association was cut at $P < 5e-5$ form each GWAS study. **(b)** A genomic view (hg19) of the CASC15/15S-SOX4 locus that was down-regulated in HR (panel 1) and the LINC01264 locus that was up-regulated in HR (panel 2). Blue horizon boxes indicate the transcripts. Orange boxes mark epigenetic hallmarks. **(c)** The corrected Fisher exact test for enrichment between lincRNA transcripts and tissue-specifically validated VISTA-enhancers. Blue color codes the results of four neural tissues

(>0.04). We concluded that risk-dependent lincRNA transcripts are a powerful metric to indicate tissue-specific transcriptional regulators.

3.3 CisPi score not only prioritizes the NB phenotype-associated lincRNAs but predicts target genes

To better explore the cis-regulatory potential of risk-dependent lincRNAs, we hypothesized that there existed a direct relationship between lincRNAs and their target genes regarding the expression level difference between distinct clinical phenotypes: the HR and NHR groups.

3.3.1 Quantitative and directional coherence existed between risk-dependent lincRNAs and local dysregulated mRNAs

We first asked if cisPi scoring systems is appropriate in the context of risk-dependent neuroblastoma transcriptome, ie, whether transcriptional association exists between lincRNAs and local genes that are risk-dependent. While lincRNAs globally co-expressed with local genes (both colored dot-lines shifted right off of a random control, the grey dot-line in Fig. 3a1), we further observed a quantitative and directional coherence of risk-dependent transcription between lincRNAs and neighboring genes: When co-localized within a TAD, risk-dependent and not risk-irrelevant lincRNA–gene pairs exhibited significantly positive correlation (Spearman $r > 0.6$, $n = 68$), especially for the directly adjacent pairs (the solid

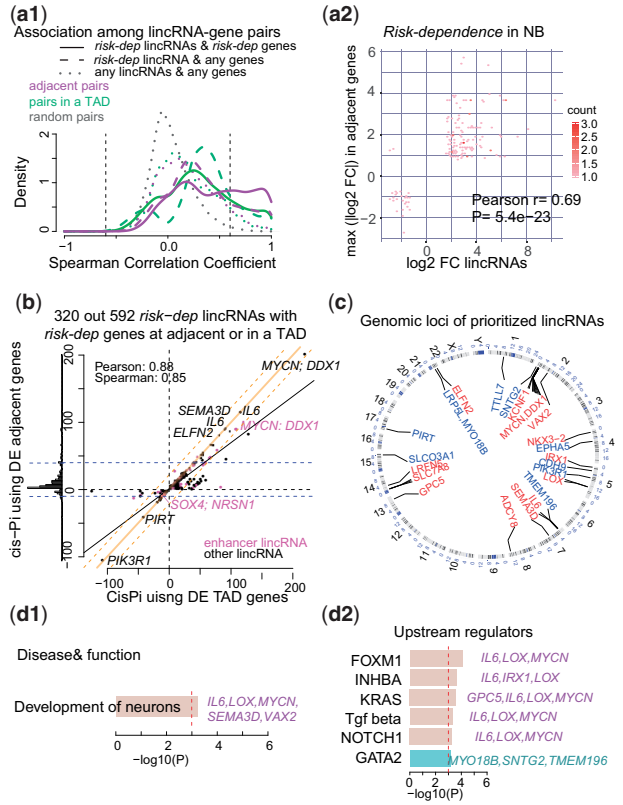

Fig. 3. CisPi metric not only prioritized the phenotype-associated lincRNAs but predicted target genes of lincRNAs. (**a1**) Density plot showing the Spearman's coefficients for five types of lincRNA–gene pairs across 68 samples. Vertical dashed lines are the significance cutoff at r = 0.6 where signatures dispatch from random controls (dotted lines). Risk-dependent adjacent pairs showing the highest proportion of positive correlation. (**a2**) Hexbin plot presenting the coherent risk-dependence of identified lincRNAs (x-axis) and genes (y-axis). (**c**) Sparse scatterplot for two types of cisPi-scores, one modeling the adjacent targeting (y-axis) and another modeling all potential targets within a TAD (x-axis). The orange diagonal line indicates y = x with two dashed lines for standard deviations. Annotated lincRNAs are labeled with the names of their nearest 'target' gene(s). Horizontal dashed lines indicate a 10% cutoff (cisPi <-10 or >40) for prioritization. (**b**) Circos plot of the 34 prioritized lincRNAs in the human genome, noted by 14 HR-upregulated genes (in red) and 10 HR-downregulated genes (in blue) sitting adjacent or in a TAD. (c) Ingenuity Pathway Analysis (IPA, www.Ingenuity.com) on these 24 predicted target genes ($P < 0.001$) showing an enriched disease and function (panel 1) and upstream regulatory molecules (Panel 2)

purple line in Fig. 3a1, Supplementary Method). Additionally, the fold-changes of risk-dependent lincRNAs significantly correlated with the fold-changes of their adjacent risk-dependent genes (Spearman rho = 0.69, $P < 2e-16$, Fig. 3a2). These coherences between differentially expressed lincRNAs and local dysregulated mRNAs fall in with a general feature of lncRNAs in other systems (Yang *et al.*, 2017a). What is new, however, is the statistical association of two seemingly irrelevant associations of lincRNA–gene pairs—between group difference and across-sample correlation—whether one agrees with the other.

To inspect the cis-regulatory effects of risk-dependent lincRNAs, we then employed our cis-regulatory lincRNA–gene-association score (CisPi-score). Theoretically, the more a score for a given lincRNA deviates from zero, the stronger this lincRNA impacts the cis-regulation of gene. For a given lincRNA, therefore, when we compare a CisPi-score that assumes adjacent gene targeting to the CisPi-score that assumes distal gene targeting within a TAD, similar CisPi values indicate that the most well-associated target gene of this lincRNA is local. For all risk-dependent lincRNAs, these two types of CisPi scores exhibited a global similarity ($P < 2e-16$, Pearson r = 0.88 and Spearman rho = 0.85) as well as focal similarity (90% of the 320 risk-dependent lincRNA–gene pairs displayed distal targeting scores similar to the estimated local targeting scores—sitting within two original dashed lines in Fig. 3b). Together, these data support the theory that these differentially transcribed lincRNAs often indicate cis-regulatory functions in the local space (Cai *et al.*, 2016).

3.3.2 Risk-dependent lincRNAs prioritized by CisPi may play a cis-regulatory role by targeting local biomarker genes

Prioritized by an early version of the CisPi scoring scheme for local targeting, we previously discovered a functional lincRNA relevant to cardiology (Yang *et al.*, 2017a). We now examined whether the advanced CisPi method prioritizes tumor-associated lincRNAs. From the distribution of calculated CisPi scores, we selected 34 lincRNAs (25 up-regulated and 9 down-regulated in HR patients) that exhibited the top 10% leading CisPi scores and local targeting (presented at the top-right and bottom-left panel in Fig. 3b, Supplementary Table S4). Except for the MYCN amplification locus, these 34 prioritized lincRNAs were independent of other recurrent genetic rearrangements, such as deletion of 1p or 11q and gain of 17q (Fig. 3c).

CisPi further associated these 34 prioritized lincRNAs with 14 up-regulated and 10 down-regulated putative target genes (Fig. 3c, Supplementary Table S5). Functional enrichment analysis of these 14 HR up-regulated genes revealed an enrichment in cell growth, proliferation and nervous system development (with the genes IL6, LOX, MYCN, SEMA3D and VAX2). These HR up-regulated genes also over-represented five upstream regulators (eg, FOXM1, KRAS, NOTCH1, etc) (Ingenuity IPA analysis, $P < 1e-3$, Fig. 3d) whose oncogenic roles had been reported. In contrast, the 10 HR down-regulated genes only indicated one upstream regulator, GATA2, that could independently operate on neuronal differentiation (El Wakil *et al.*, 2006). These results are expected, given a cis-regulatory model of lincRNA loci targeting to local genes, with both derived from risk-dependent alterations in expression.

3.3.3 CisPi-prioritized lincRNAs and target genes stratify risk-groups and are prognostic in NB, regardless of MYCN status

We next tested whether the abnormal expression of these 34 prioritized lincRNAs is robustly informative in predicting the HR status of new patients. We mapped these 34 lincRNAs into the non-coding transcripts measured by the RNA-Seq data of a larger, independent Germany population (n = 498), resulting in 32 recaptured lncRNAs (25 up-regulated and 7 down-regulated transcripts, Supplementary Table S4). With an unsupervised bi-clustering of the noncoding transcriptome profiling of these 32 lincRNA transcripts alone, 432 (87%) tumors were correctly classified for their risk-groups (Fig. 4a). This accurate risk-stratification occurred regardless of the MYCN status, which is currently the best-characterized genetic marker of risk in neuroblastoma, suggesting that the loci represent novel biomarkers.

High-risk neuroblastoma is a severe pediatric tumor characterized by poor prognosis. Therefore, we further examined the power of these prioritized lincRNAs to predict outcome endpoints, using the robust RXA-LSP indicator (Supplementary Method). Compared against the polyadenylated TARGET profiling, our 34 risk-dependent lincRNAs significantly predicted overall survival

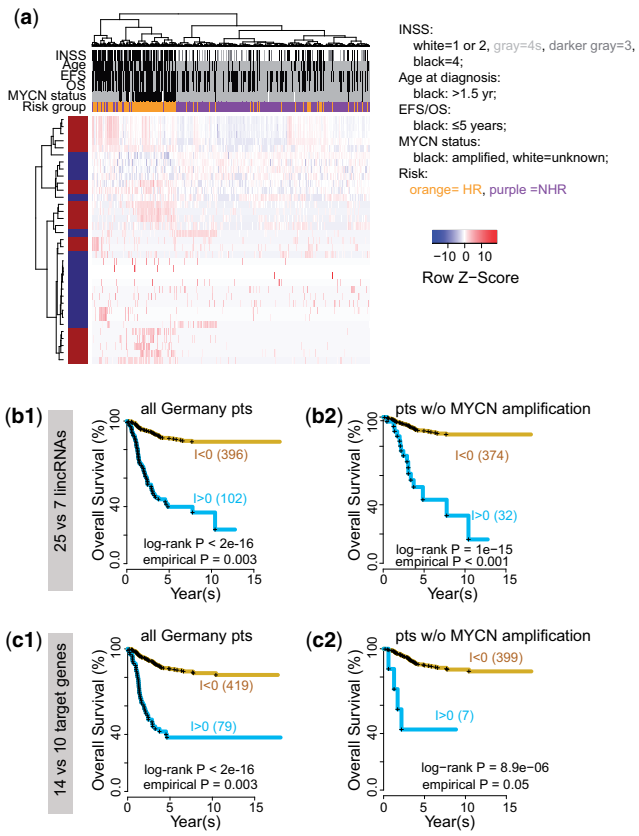

Fig. 4. Clinical implication of the prioritized lincRNAs, regardless of the MYCN status. (**a**) Heatmap with 32 prioritized lincRNAs in 498 independent patients, the Germany dataset, using a two-way hierarchical cluster of CPM (Complete linkage with Manhattan distance). The color in the left vertical bar codes risk-dependence, red for up-regulation and blue for down-regulation in HR. (**b**) Overall survival of patients stratified by the RXA-LSP indicator built on these lincRNAs for all patients (Panel 1) and patients without MYCN amplification (Panel 2). (**c**) Overall survival of patients stratified by the RXA-GSP indicator of 24 predicted target genes

(log-rank $P = 0.021$, empirical $P = 0.038$, Supplementary Fig. S5a). Evaluated in the independent Germany patients, the RXA-LSP indicator significantly separated favorable outcomes with a negative value from unfavorable outcomes with a positive value for all patients with neuroblastoma (Fig. 4b1) and patients without MYCN amplification, a cohort that is generally considered to be in the low risk group (Fig. 4b2) (severe empirical $P = 0.003$ and <0.001 respectively). We also observed a similar significance to predict event-free survival (EFS, Supplementary Fig. S5b). Notably, expression of the predicted target genes of these lincRNAs also accurately stratified overall survival prediction for neuroblastoma in all patients, and specifically in patients without evident MYCN-amplification (Fig. 4c, Supplementary Fig. S5 b, c2 and d2), suggesting a role for lincRNAs in guiding the risk-dependent expression of 'target' genes. To our best knowledge, this is the first attempt to infer gene biomarkers from lincRNA biomarkers that are associated with these genes, a successful implementation of the downstream gene-driven strategy for the prediction of functional lincRNAs.

4 Discussion

Altered expression of lncRNAs that stratifies tumor risk is an informative readout of oncogenic enhancer activity. These lncRNAs provide complementary information on cis-regulation, with or without a hallmark of enhancer activity. Therefore instead of focusing only on the enhancer hallmarks, our CisPi metric prompts a more quantitative consideration of differentially expressed lncRNAs with local target genes from RNA-Seq data alone, irrespective of enhancer detection. However, our analysis cannot distinguish the underlying cis-regulatory mechanisms which could be enhancer activity and/or lncRNA expression that targets the predicted local genes. Nevertheless, this methodology can be applied to the enormous amount of previously generated RNA-Seq data, opening the door wider to computational knowledge discovery.

LincRNAs, a subset of lncRNAs residing in evolutionarily conserved intergenic regions and mostly being polyadenylated, hold one-quarter of annotated human lncRNAs (GENCODE v19). We observed a significant co-localization of lincRNA transcripts at active enhancers, suggesting that lincRNAs could be enhancer-templated and active enhancers could be transcribed into polyadenylated RNAs. Given that only minorities of lncRNAs emanating from enhancers are polyadenylated, lincRNAs that have polyA tails and that are transcribed from enhancers without typical enhancer hallmarks are intriguing—indicating that either enhancer detection by ChIP-seq may have failed to identify these loci or that these loci represent a novel class of transcriptional regulatory lincRNAs. Our previous approach based on non-polyadenylated libraries will miss most of these lincRNAs, and so it will be incomplete to decipher functional cis-regulation (Yang *et al.*, 2017a). Instead, this practice of using deep-sequenced (>50 M reads per sample) polyadenylated RNA-Seq data provides complementary insights into cis-acting lincRNAs. In the future, applying this framework to data of total RNA-Seq profiling is practical and will lead to a more profound understanding of cis-regulatory gene network in disease.

Finally, with the increase in available sequencing data, incorporating population genetic evidence from GWAS, quantitative genomic evidence from the transcriptome, and epigenetic evidence from histone marks are encouraging for knowledge discovery. Our results demonstrate the promise of such computational incorporation for precision genomics. Notably, with evidence of an independent transcriptional association associated with disease, additional GWAS findings could lead to the discovery of new functional candidates, such as the LINC01264 at the RET locus that we predicted in this study. If validated, this strategy will provide post-GWAS research an essential opportunity to validate candidates computationally.

Acknowledgements

We specifically acknowledge Ivan Moskowitz, Susan Cohn, Megan Rowton and Jeff D Steimle for instructive discussions, Carlos Perez-Cervantes for downloading the TARGET and GRASP datasets, and the assistance of Lorenzo Pesce for data storage.

Funding

This work was supported by NIH R21LM012619 and through resources provided by the University of Chicago under grant 1S10OD018495-01. The results published here are partly based upon data generated by the Therapeutically Applicable Research to Generate Effective Treatments (TARGET), supported by NCI Grant UlO CA98543. The neuroblastoma data used for this analysis are available under accession [phs000467.v2] (which is now collected under [phs000218] in dbGaP. This specific grant is a collaboration of the Children's Oncology Group and The Children's Hospital of Philadelphia, Children's Hospital of Los Angeles and the National Cancer Institute.

Conflict of Interest: none declared.

References

Azofeifa,J.G. *et al*. (2018) Enhancer RNA profiling predicts transcription factor activity. *Genome Res*, **28**, 334–344.

Cai,L. *et al*. (2016) A comprehensive characterization of the function of LincRNAs in transcriptional regulation through long-range chromatin interactions. *Sci. Rep.,* **6**, 36572.

Chang,C.C. *et al*. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience,* **4**, 7.

de Hoon,M. *et al*. (2015) Paradigm shifts in genomics through the FANTOM projects. *Mamm. Genome,* **26**, 391–402.

Eicher,J.D. *et al*. (2015) GRASP v2.0: an update on the genome-wide repository of associations between SNPs and phenotypes. *Nucleic Acids Res.,* **43**, D799–D804.

El Wakil,A. *et al*. (2006) The GATA2 transcription factor negatively regulates the proliferation of neuronal progenitors. *Development,* **133**, 2155–2165.

Esposito,G. *et al*. (2008) Genomic and functional profiling of human Down syndrome neural progenitors implicates S100B and aquaporin 4 in cell injury. *Hum. Mol. Genet.,* **17**, 440–457.

Fernando,T.R. *et al*. (2017) The lncRNA CASC15 regulates SOX4 expression in RUNX1-rearranged acute leukemia. *Mol. Cancer,* **16**, 126.

Iyer,M.K. *et al*. (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.,* **47**, 199–208.

Kim,T.K. *et al*. (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature,* **465**, 182–187.

Li,W. *et al*. (2016) Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.,* **17**, 207–223.

Maris,J.M. *et al*. (2008) Chromosome 6p22 locus associated with clinically aggressive neuroblastoma. *N. Engl. J. Med.,* **358**, 2585–2593.

Mondal,T. *et al*. (2018) Sense-antisense lncRNA pair encoded by locus 6p22.3 determines neuroblastoma susceptibility via the USP36-CHD7-SOX9 regulatory axis. *Cancer Cell*, **33**, 417–434. e417.

Orom,U.A. *et al*. (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell*, **143**, 46–58.

Pugh,T.J. *et al*. (2013) The genetic landscape of high-risk neuroblastoma. *Nat. Genet.,* **45**, 279–284.

Russell,M.R. *et al*. (2015) CASC15-S is a tumor suppressor lncRNA at the 6p22 neuroblastoma susceptibility locus. *Cancer Res.,* **75**, 3155–3166.

Signal,B. *et al*. (2016) Computational approaches for functional prediction and characterisation of long noncoding RNAs. *Trends Genet.,* **32**, 620–637.

Ulitsky,I. and Bartel,D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell,* **154**, 26–46.

Vance,K.W. and Ponting,C.P. (2014) Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends Genet.,* **30**, 348–355.

Visel,A. *et al*. (2007) VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.,* **35**, D88–D92.

Wang,Y. *et al*. (2017) The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. http://biorxiv.org/content/early/2017/02/27/112268.

Wang,Z. *et al*. (2015) Seq2pathway: an R/Bioconductor package for pathway analysis of next-generation sequencing data. *Bioinformatics,* **31**, 3043–3045.

Won,H. *et al*. (2016) Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature,* **538**, 523–527.

Xiao,Y. *et al*. (2014) A novel significance score for gene selection and ranking. *Bioinformatics,* **30**, 801–807.

Yang,X. *et al*. (2013) Bridging cancer biology with the clinic: relative expression of a GRHL2-mediated gene-set pair predicts breast cancer metastasis. *PLoS One,* **8**, e56195.

Yang,X.H. *et al*. (2017a) Transcription-factor-dependent enhancer transcription defines a gene regulatory network for cardiac rhythm. *Elife*, 6:e31683.

Yang,X.H. *et al*. (2017b) A c-Myc-regulated stem cell-like signature in high-risk neuroblastoma: a systematic discovery (Target neuroblastoma ESC-like signature). *Sci. Rep.,* **7**, 41.

Yang,X.H. *et al*. (2017c) Incorporating genomic, transcriptomic and clinical data: a prognostic and stem cell-like MYC and PRC imbalance in high-risk neuroblastoma. *BMC Syst. Biol.,* **11**, 92.

Yao,X.M. *et al*. (2017) High expression of LncRNA CASC15 is a risk factor for gastric cancer prognosis and promote the proliferation of gastric cancer. *Eur. Rev. Med. Pharmacol. Sci.,* **21**, 5661–5667.