frontiers in
BIOENGINEERING AND BIOTECHNOLOGY

ORIGINAL RESEARCH ARTICLE
published: 14 January 2015
doi: 10.3389/fbioe.2014.00088

Discovery of protein–lncRNA interactions by integrating large-scale CLIP-Seq and RNA-Seq datasets

Jun-Hao Li†, Shun Liu†, Ling-Ling Zheng†, Jie Wu, Wen-Ju Sun, Ze-Lin Wang, Hui Zhou, Liang-Hu Qu* and Jian-Hua Yang*

RNA Information Center, Key Laboratory of Gene Engineering of the Ministry of Education, State Key Laboratory for Biocontrol, Sun Yat-sen University, Guangzhou, China

Edited by:
Alessandro Laganà, The Ohio State University, USA

Reviewed by:
Igor B. Rogozin, National Institutes of Health, USA
Thiruvarangan Ramaraj, National Center for Genome Resources, USA

****Correspondence:***
Liang-Hu Qu and Jian-Hua Yang, Biotechnology Research Center, Sun Yat-sen University, Guangzhou 510275, China
e-mail: lssqlh@mail.sysu.edu.cn; yangjh7@mail.sysu.edu.cn

† Jun-Hao Li, Shun Liu and Ling-Ling Zheng have contributed equally to this work.

Long non-coding RNAs (lncRNAs) are emerging as important regulatory molecules in developmental, physiological, and pathological processes. However, the precise mechanism and functions of most of lncRNAs remain largely unknown. Recent advances in high-throughput sequencing of immunoprecipitated RNAs after cross-linking (CLIP-Seq) provide powerful ways to identify biologically relevant protein–lncRNA interactions. In this study, by analyzing millions of RNA-binding protein (RBP) binding sites from 117 CLIP-Seq datasets generated by 50 independent studies, we identified 22,735 RBP–lncRNA regulatory relationships. We found that one single lncRNA will generally be bound and regulated by one or multiple RBPs, the combination of which may coordinately regulate gene expression. We also revealed the expression correlation of these interaction networks by mining expression profiles of over 6000 normal and tumor samples from 14 cancer types. Our combined analysis of CLIP-Seq data and genome-wide association studies data discovered hundreds of disease-related single nucleotide polymorphisms resided in the RBP binding sites of lncRNAs. Finally, we developed interactive web implementations to provide visualization, analysis, and downloading of the aforementioned large-scale datasets. Our study represented an important step in identification and analysis of RBP–lncRNA interactions and showed that these interactions may play crucial roles in cancer and genetic diseases.

Keywords: long non-coding RNA, RNA-binding protein, GWAS, CLIP-Seq, RNA-Seq

INTRODUCTION

Mammalian genomes encode thousands of long non-coding RNAs (lncRNAs) (Wang and Chang, 2011; Guttman and Rinn, 2012). lncRNAs play important roles in a variety of biological processes that have been implicated in regulating tumorigenesis through interaction with RNA-binding proteins (RBPs) (Konig et al., 2011; Wang and Chang, 2011; Guttman and Rinn, 2012; Ulitsky and Bartel, 2013). However, for the majority of lncRNAs, the mechanism underlying their interaction with RBPs remains unknown (Konig et al., 2011; Wang and Chang, 2011; Guttman and Rinn, 2012; Ulitsky and Bartel, 2013).

The control and function of lncRNA are governed by the specificity of RBPs (Wang and Chang, 2011; Guttman and Rinn, 2012). Increasing evidence suggests that many RBP–lncRNA interactions play important roles in correct transcriptional regulation (Konig et al., 2011; Wang and Chang, 2011; Guttman and Rinn, 2012; Ulitsky and Bartel, 2013). One emerging theme that many lncRNAs regulate gene expression by directing chromatin modificators to specific target regions (Ulitsky and Bartel, 2013). Significant fractions (20% in human) of lincRNAs are interacted with PRC2 and other chromatin-modifying complexes (Khalil et al., 2009; Guttman et al., 2011). The functional outcomes of some binding events have been revealed. For example, HOTAIR, which is transcribed from human HOX locus, guides repressor PRC2 to specific mammalian loci to silence gene expression and to promote cancer metastasis (Rinn et al., 2007; Wang et al., 2011). Besides, many lncRNAs have been shown to interact with other types of RBPs, including DNA methyltransferases (Schmitz et al., 2010; Di Ruscio et al., 2013), transcription factors (Wang et al., 2014), and splicing factors (Tripathi et al., 2010; Gong and Maquat, 2011; Yin et al., 2012). However, deciphering the interactions between hundreds of RBPs and thousands of lncRNAs remains a daunting challenge.

Genome-wide association studies (GWAS) have identified thousands of common genetic variants related to specific traits or disease phenotypes, and many of these variants (about 88%) lie in non-coding regions, which could potentially influence processing and expression of ncRNAs (Sethupathy and Collins, 2008; Hindorff et al., 2009; Ryan et al., 2010; Cabili et al., 2011; Kumar et al., 2013; Ning et al., 2014). For example, single nucleotide polymorphism (SNP) within miR-125a gene alters the processing of pri-miRNA by DGCR8 and causes recurrent pregnancy loss in a Han-Chinese population (Duan et al., 2007; Hu et al., 2011). Another study found that a papillary thyroid carcinoma-associated SNP, rs944289 affects the expression of lncRNA PTCSC3 by changing the binding activity of C/EBPα transcription factor (Cabili et al., 2011; Jendrzejewski et al., 2012). Although the genetic variants in interaction sites of RBP–lncRNA may interfere lncRNA functions and affected the susceptibility to human diseases, the relationships between genetic variants and interaction sites were yet unexplored.

Recent advances in high-throughput sequencing of RNA isolated by cross-linking immunoprecipitation (HITS-CLIP, CLIP-Seq, PAR-CLIP, CLASH, iCLIP) have provided powerful ways to identify RBP-associated RNAs and map such interactions in the genome (Chi et al., 2009; Hafner et al., 2010; Konig et al., 2011; Helwak et al., 2013; Fu, 2014; Fu and Ares, 2014). The application of CLIP-Seq methods has reliably identified Argonaute (Ago) binding sites and miRNA-target interactome (Chi et al., 2009; Hafner et al., 2010; Helwak et al., 2013). In fact, many more studies to date have been focused on understanding the function of RBPs in RNA metabolism (Konig et al., 2011; Fu, 2014), such as pre-mRNA splicing (Fu and Ares, 2014). While an increasing number of RBPs have been explored using CLIP technologies, binding peaks mapped to non-protein-coding genes have been routinely discarded and not further analyzed. However, this data will be a rich trove well worthy of mining RBP–lncRNA relationships.

In this study, we performed a large-scale integration of public RBP binding sites generated by high-throughput CLIP-Seq technology and identified thousands of RBP–lncRNA interactions. Furthermore, by combining GWAS and RNA-Seq data, we explored clinically relevant RPB–lncRNA interactions that may facilitate the translation of genetic studies of complex diseases into therapeutics.

MATERIALS AND METHODS

INTEGRATION OF RBP BINDING SITES FROM PUBLISHED CLIP DATA

HITS-CLIP, PAR-CLIP, and iCLIP binding clusters/peaks data were retrieved from the gene expression omnibus and sequence read archive (SRA) (Barrett et al., 2013), the supplementary data of original references or directly from authors upon request. All binding sites coordinates were converted to hg19 and mm10 assemblies using the UCSC LiftOver Tool (Meyer et al., 2013).

RBP TARGET SITES SCANNING IN ANNOTATED lncRNA TRANSCRIPTS

Human gene annotations were acquired from GENCODE Version 17 (Harrow et al., 2012). Mouse gene annotations were extracted from Ensembl Gene Release 72 (Hubbard et al., 2009) and LiftOver to mm10 assembly. lncRNAs were further filtered to remove the transcripts overlapping with protein-coding genes. The aforementioned RBP CLIP clusters were used to intersect with the coordinates of all annotated transcripts to find their RBP binding sites, which were fed to Circos (Krzywinski et al., 2009) for visualization.

TCGA TUMOR EXPRESSION DATA AND EXPRESSION CORRELATION OF RBPs AND lncRNAs

The Cancer Genome Atlas (TCGA) RNA-Seq expression datasets (level 3, IlluminaHiSeq_RNASeqV2) for 14 cancer types and gene annotation file (TCGA.hg19.June2011.gaf) were downloaded from TCGA Data Portal (Cancer Genome Atlas Research Network, 2008). Expression of 397 known lncRNAs can be measured in TCGA level 3 RNA-Seq data. Expression correlation (Pearson correlation coefficient) between lncRNAs and RBPs was estimated using co-expression program (the program is available from the authors upon request), which was written in C language and ALGLIB library, and p-value was adjusted with the false discovery rate (FDR) correction (Benjamini and Hochberg, 1995).

IDENTIFICATION OF DISEASE-RELATED SNPs IN RBP BINDING SITES ASSOCIATED WITH lncRNAs

Disease/phenotype associated SNPs were curated from published GWAS data provided by the NHGRI GWAS Catalog (Welter et al., 2014), Johnson and O'Donnell (2009), dbGAP (Mailman et al., 2007), and GAD (Becker et al., 2004). Additional SNPs in linkage disequilibrium (LD) with reported disease-related loci were selected with the criteria requiring an r^2 value over 0.5 in at least one of the four populations (CEU, CHB, JPT, and YRI) genotype data of the HapMap project (release 28) (International HapMap 3 Consortium et al., 2010). For each SNP, rs ID were lifted to dbSNP build 141 based on the "RsMergeArch.bcp" and "SNPHistory.bcp" table from dbSNP, and genomic coordinates were lifted to the hg19 assembly using the UCSC LiftOver tool. All these disease-related SNPs or LD SNPs were mapped to exons and splicing sites (2 nt in the intron that is close to an exon) of the annotated lncRNA transcripts and further examined whether they were located in any RBP binding clusters.

DATA VISUALIZATIONS

RNA-binding protein–lncRNA interactions were deposited in our starBase V2.0 (Li et al., 2014) under the "Protein–RNA" section[1]. For each interaction, we provided links to our enhanced deepView genome browser[2], which was written using a GD graphics library for PHP, to visualize RBP binding sites, lncRNAs, and other annotation tracks in an integrated display style similar to that of UCSC genome browser.

RESULTS

THE GENOME-WIDE BINDING MAP OF RNA-BINDING PROTEINS AND THE ANNOTATION OF RBP–lncRNA INTERACTIONS

We curated 117 published CLIP-Seq datasets to profile the genome-wide binding maps of 65 RBP. Unique binding sites of distinct RBPs varied from thousands to millions, and the genomic context distributions of binding sites for different RBPs distinguished from each other (**Figure 1**; Table S1 in Supplementary Material). For example, PUM2, a translational repressor during embryonic development and cell differentiation (Huang et al., 2011), predominately bound to 3′UTR regions of protein genes, while another translation inhibitor FMRP (Napoli et al., 2008) tended to interact with CDS. The discrepancy in binding context preferences for RBPs could root from different amounts of available datasets, usages of various variants of CLIP-Seq, varying sequencing depth, and/or genuine distinctions in the underlying recognition mechanism of RBPs.

Despite that the majority of RBP binding sites were mapped to protein-coding genes, on average 1.1% of RBP binding sites lay within exons of human lncRNAs. In total, 21,073 and 1,662 RBP–lncRNA interactions were identified in human and mouse, respectively (**Table 1**). It is noteworthy that most well-studied lncRNAs interacted with chromatin modificators, acting as tethers or scaffolds (Khalil et al., 2009; Kung et al., 2013). Thus, we considered the binding features of Ezh2, a subunit of PRC2 complexes, by analyzing the CLIP-Seq data from mouse embryonic stem cells

[1]http://starbase.sysu.edu.cn/rbpLncRNA.php
[2]http://starbase.sysu.edu.cn/browser.php

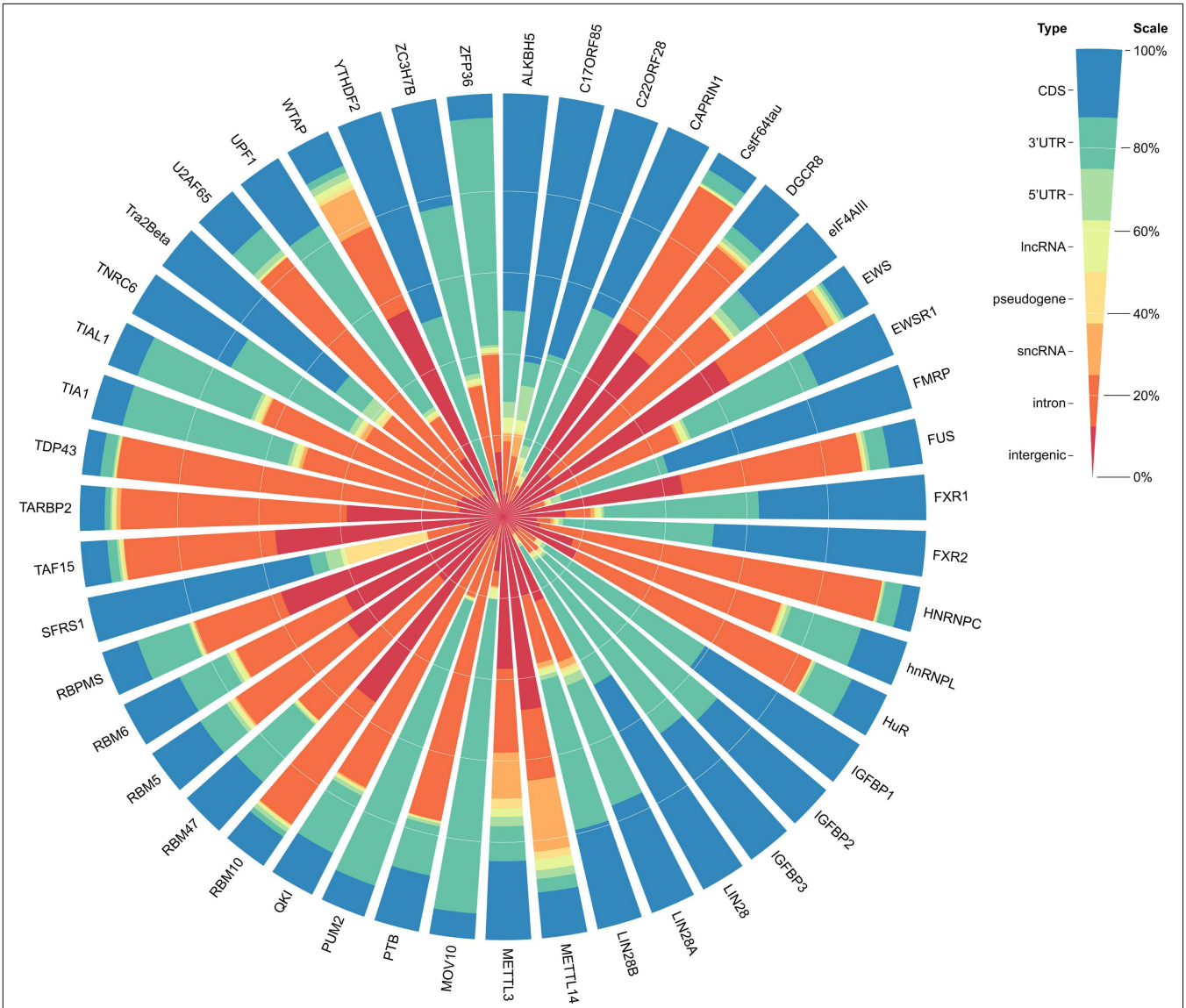

FIGURE 1 | The genomic context distributions of binding sites for 47 human RBPs. Binding sites are mapped to genomic features in the following priority order: CDS, 3′UTR, 5′UTR, lncRNA, pseudogene, sncRNA, intron, intergenic.

Table 1 | The summary of CLIP-Seq datasets used in this study and the resulting RBP–lncRNA interactions.

Species	Experiments	RBPs	Cell lines/ tissues	RBP binding sites mapped to lncRNAs	RBP–lncRNA interactions
Human	90	47	13	84,356	21,073
Mouse	27	18	20	5,330	1,662

(Kaneko et al., 2013). Our results demonstrated that Ezh2 interacted with 35 lncRNAs including many imprinted RNAs, such as Tsix, Meg3, Rian, and Pvt1 (**Figure 2**), which was consistent with the epigenetic features of PRC2 (Zhao et al., 2010).

EXPLORING COMBINATORIAL EFFECTS AMONG RBPs

For the 12,255 human lncRNAs, 56.8% were found bound to at least 1 RBP. Surprisingly, 16 lncRNAs, including GAS5 and NEAT1, harbored binding sites of over 30 RBPs (**Figure 3**; Table S2 in Supplementary Material), indicating their diverse roles in biological processes when accompanied with different RBPs. Since one lncRNA could interact with multiple RBPs, it could be expected that some RBP binding sites were overlapped with each other. Therefore, we explored combinatorial effects among RBPs by employing integrated CLIP-Seq datasets. For example, we utilized PAR-CLIP data generated in HEK293 and intersect binding sites of three RNA destabilizer HuR, Ago2, and MOV10. The results showed that tens of lncRNAs, including cancer-related lncRNAs TUG1, DLEU2, and GAS5, were bound by at least two of the

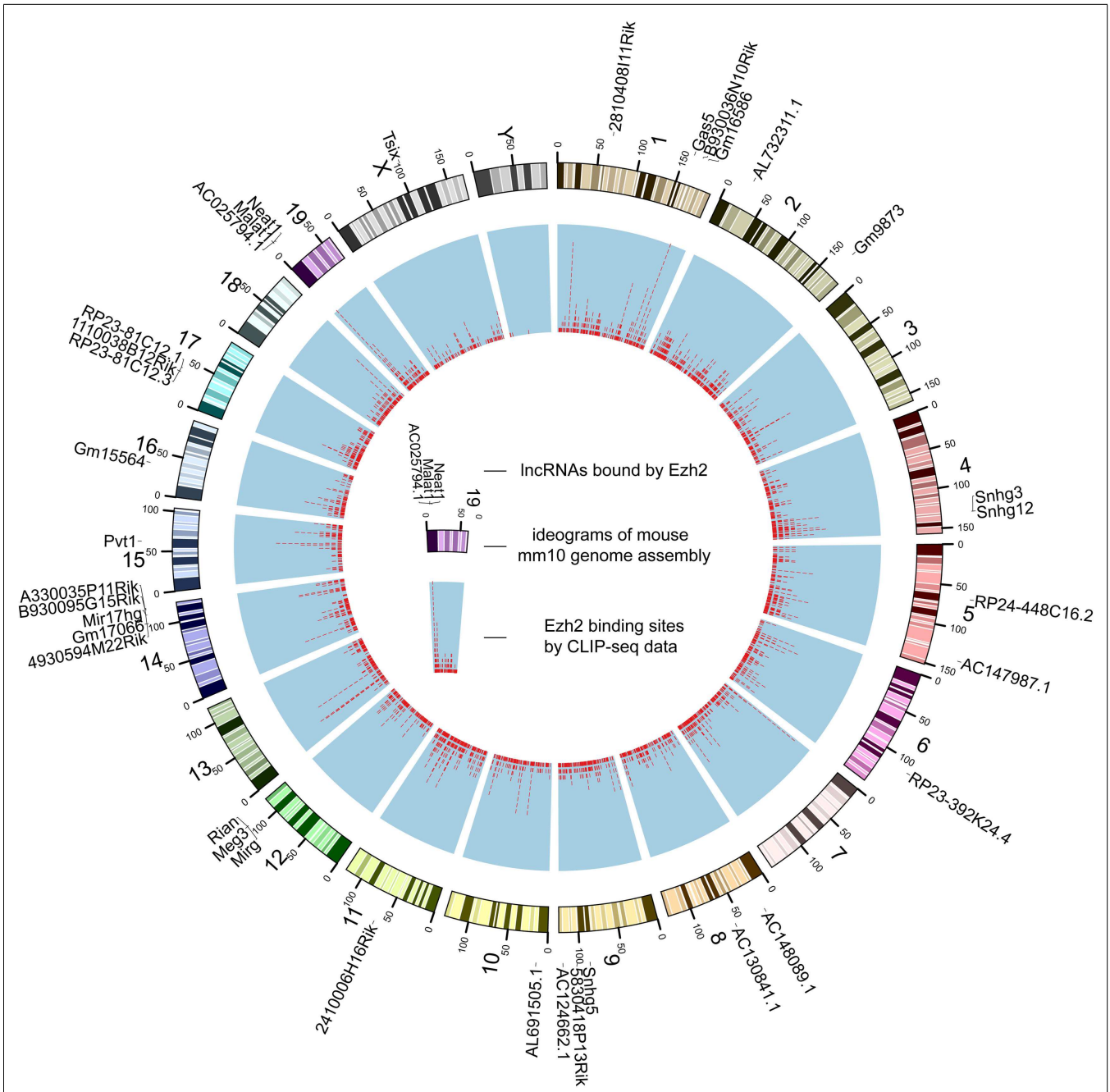

FIGURE 2 | The genome-wide binding map of Ezh2 in mouse. The outer track is mouse chromosomes labeled with lncRNAs bound by Ezh2. The red tiles of the inner track represent the genomic coordinates of corresponding binding sites.

three RBPs at identical binding sites (**Figure 4**). This phenomenon suggested that the stabilities of these lncRNAs were likely under joint control of these three RBPs, which could be explained by their confirmed interplays in HEK293 (Chendrimada et al., 2007) and Hela cells (Kim et al., 2009).

EXPRESSION ASSOCIATION OF RBP–lncRNA INTERACTIONS

To realize the roles of RBP–lncRNA interactions in cancer, we preformed co-expression analysis between RBPs and lncRNAs by virtue of 90 human CLIP-Seq datasets and expression data from more than 6,000 tumor samples in 14 types of cancer. Up to 583 pairs concerning 47 RBPs and 49 lncRNAs showed strong correlation at expression levels in at least 1 cancer type (**Figure 5A**). Marvelously, PUM2 and TUG1 involved with cell cycle regulation (Khalil et al., 2009; Huang et al., 2011) showed significant positive expression correlation ($p < 0.05$) in all 14 cancer types (**Figure 5B**). Two potential PUM2 binding sites on TUG1 have the consensus recognition

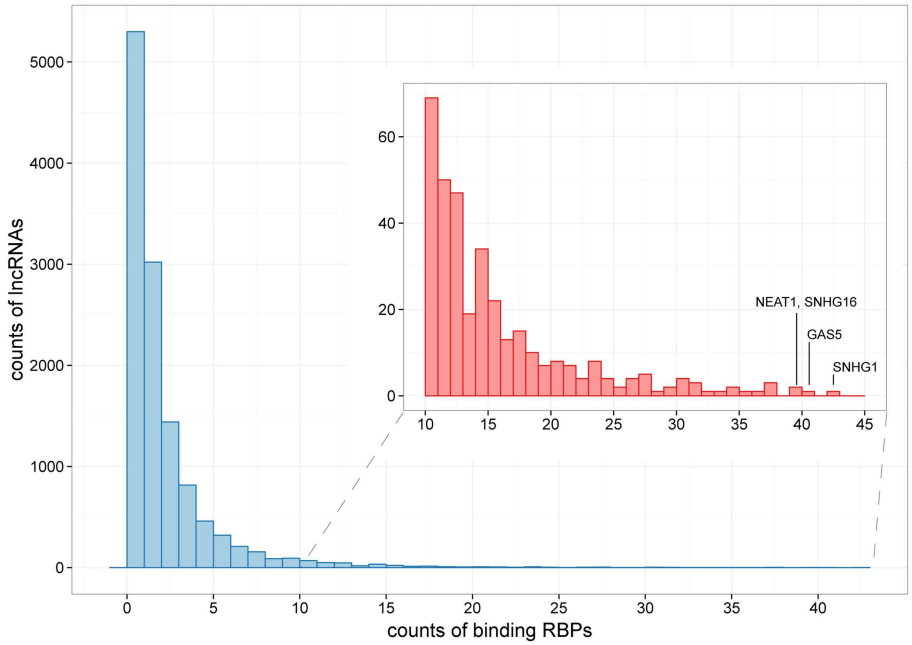

FIGURE 3 | The distribution of lncRNAs bound by different numbers of RBPs. Histograms showing counts of lncRNAs bound by over 10 RBPs are zoomed in at the subpanel. SNHG1, GAS5, NEAT1, and SHNG16 are marked, which are bound by 42, 40, 39, and 39 RBPs, respectively.

motif UGURUAUA, which was highly conserved in mammals (**Figure 5C**).

PREDICTING GWAS-ASSOCIATED RBP BINDING SITES IN lncRNAs

Although GWAS over the years have revealed a significant number of genetic variants related to diseases or phenotypes, a considerable portion of these identified loci are not within protein-coding genes and therefore not functionally explained to date (Hindorff et al., 2009). Here, we tried to fill this gap by connecting RBP binding sites in lncRNAs and potential disease-related SNPs.

Altogether, 87,677 unique disease-related SNPs were collected from four public GWAS data source (Table S3 in Supplementary Material, detailed in Section "Materials and Methods"). Considering that additional SNPs in LD with reported disease-related loci may also map to RBP binding sites in lncRNAs, we perform LD analysis to extracted SNPs that had high LD relationship with disease-related SNPs using a threshold of $r^2 > 0.5$ in at least one population from the HapMap CEU, CHB, JPT, and YRI genotype data, which yielded a total of 895,968 disease-related or LD SNPs.

We found that 2431 of these SNPs were mapped to the exons of 2089 transcripts of 1489 lncRNA genes, among which 162 SNPs were also located in at least 1 binding sites of 29 RBPs (Table S4 in Supplementary Material). For example, three disease-related SNPs, namely, rs16902485, rs10283090, and rs2720659, resided in the exons of lncRNA PVT1. According to the GWAS annotations of Johnson and O'Donnell (2009), the latter two of the three SNPs were associated with "type II diabetes mellitus," which was in good accordance with the recent reports showing that PVT1 may contribute to diabetic nephropathy (Hanson et al., 2007; Alvarez and DiStefano, 2011; Alwohhaib et al., 2014). These SNPs were also overlapped with binding sites of U2AF65, HuR, and eIF4AIII, respectively (**Figure 6**), suggesting variants in these sites might result in impaired binding of these RBPs to PVT1, which thereby might lead to the development of corresponding diseases.

Next, we checked whether disease-related SNPs might be located in the splicing sites of lncRNAs and affect the alternative splicing of lncRNAs. We defined a splicing site as the 2 nt within an intron close to the exon–intron junction. As a result, we found that only 24 SNPs lay within lncRNA splicing sites (Table S4 in Supplementary Material), among which only 1 SNP, rs17207481, was overlapped with binding sites of FUS and HuR. These results suggested that SNPs exerted limited effects on disease occurrence through the mechanism of disturbing alternative splicing of lncRNAs.

COMPARATIVE ANALYSIS OF RBP TARGETS USING THE deepView GENOME BROWSER

To facilitate comparative analysis of the CLIP-Seq datasets and exploration of RBP–lncRNA interactions, we developed the improved deepView Genome Browser[2] in starBase V2.0 (Li et al., 2014). In the query page of the browser, users can input one interested genomic region in the "search term" and select corresponding genome assembly to gain an integrated view of various genomic features. Information on binding sites of RBPs, predicted miRNA-target sites overlapped with CLIP-Seq data, as well as gene annotations from RefSeq and Ensembl were provided in toggleable tracks. The image of the browser will be updated immediately by clicking the "refresh tracks" button when users change track options. **Figure 7** illustrated the visualization of FUS–MEG3 interactions with deepView. Users can click the "zoom in" or "zoom out"

FIGURE 4 | The genome-wide binding map of HuR, Ago2, and MOV10 in human. The outermost track represents ideograms of chr1, chr13, and chr22 in human genome. lncRNAs bound by these RBPs are labeled on the periphery, and those bound by at least two of the three RBPs at identical binding sites are colored red. The blue, green, and purple tracks indicate the binding positions of HuR, Ago2, and MOV10, respectively.

button at the top to shrink or extend on the center of the annotation tracks window by 1.5-, 3-, or 10-folds. Clicking the "View region at UCSC" button will redirect users to the UCSC page and exhibit the current region on the UCSC genome browser. To explore RBP binding sites on a particular gene, users can type its gene symbol in the position textbox and then click the "GO" button to update the display image to determine, which RBPs might participate in regulating the gene. Our visualization method allows a direct comparison of binding patterns of different RBPs, binding preferences of one particular RBP in different cell lines and tissues, and genomic contexts of RBP binding sites.

DISCUSSION

Although a few dozen lncRNAs have been characterized to some extent and reported to function in important cellular processes, the functions of most annotated lncRNAs are

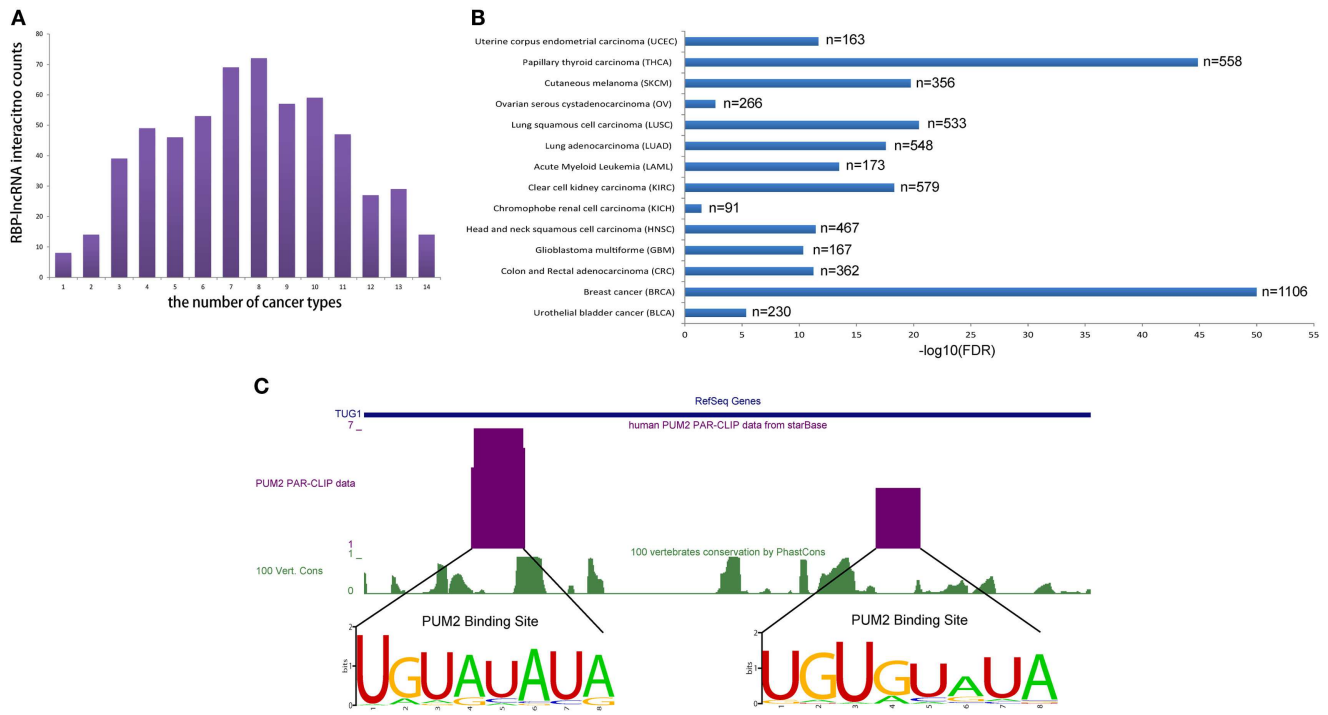

FIGURE 5 | RBP–lncRNA interactions are supported by co-expression analysis in 14 types of cancers. **(A)** Histograms show RBP–lncRNA interactions with expression association (Pearson correlation, $p < 0.05$) in at least one cancer type. **(B)** The expression levels of PUM2 and TUG1 are positively correlated ($p < 0.05$) in all 14 cancers. **(C)** The PUM2 binding sites on TUG1 are inferred from PAR-CLIP data, and the consensus recognition motif UGURUAUA are conserved in mammals.

unknown (Guttman and Rinn, 2012; Ulitsky and Bartel, 2013). Several bioinformatics resources and tools have made efforts to functionally annotate lncRNAs (Da Sacco et al., 2012), such as fRNAdb (Kin et al., 2007) and ncFANs (Liao et al., 2011). These tools mainly inferred lncRNA function by their differential expression in distinct biological states or their co-expression patterns with protein-coding genes, but little attention was paid to the relationship of lncRNAs and their bounded proteins. In this study, by analyzing a large set of RBP binding sites derived from all available CLIP-Seq experimental techniques (PAR-CLIP, HITS-CLIP, iCLIP, CLASH), we have shown extensive and complex RBP–lncRNA interaction networks (**Figure 1**).

Recent studies have revealed that many lncRNAs function through specific interactions with RBPs, but whether these interactions are direct and specific remains controversial. RBP–lncRNA interactions identified by low stringent immunoprecipitation of non-cross-linked RNA–protein complexes, such as RIP-Chip and RIP-Seq, may contain indirect binding relationships (Konig et al., 2011). In comparison to previously reported significant fractions (10% in mouse) of PRC2-associated lncRNAs (Zhao et al., 2008), we found that a relatively small fraction (~1%) of lncRNAs were bound by Ezh2 in mouse (**Figure 2**). Therefore, we provide enhanced resolution to determine lncRNA functional networks based on RBP–lncRNA interactions supported by high-throughput CLIP-Seq data. More than 80,000 binding clusters identified from 65 different RBPs represent a valuable resource for resolving some obstacles that have arisen in efforts to understand lncRNA action. Nevertheless, although CLIP-Seq is designed to detect direct binding events of proteins and RNAs, the resulting data might still contain false positives and false negatives, which may root from every cumbersome step of this technique. To minimize the impact of such false discoveries, we filtered the origin results by the reported FDR and provided evidences such as number of CLIP reads and number of supporting experiments, which may help users to gain RBP–lncRNA interactions of high-confidence.

By cross analysis of binding maps for multiple RBPs, this study offers a new resource to understanding joint control of target lncRNA expression. While only 65 RBPs were analyzed, we found that many of the RBPs bound to the same lncRNA (**Figure 3**). This is consistent with the compelling idea that lncRNAs can serve as scaffolds that assemble many relevant RBPs to regulate gene expression (Wang and Chang, 2011; Ulitsky and Bartel, 2013). At the same time, we also identified hundreds of identical binding sites that bound by multiple different RBPs in lncRNAs (**Figure 4**), probably reflecting competition among RBPs that binding on a given lncRNA.

Our combined analysis of CLIP-Seq data and GWAS data revealed hundreds of disease-related SNPs resided in the RBP binding sites of lncRNAs (Table S4 in Supplementary Material). Unlike the sporadic attempts on simply finding genetic variants associated with disease susceptibility within lncRNA genes

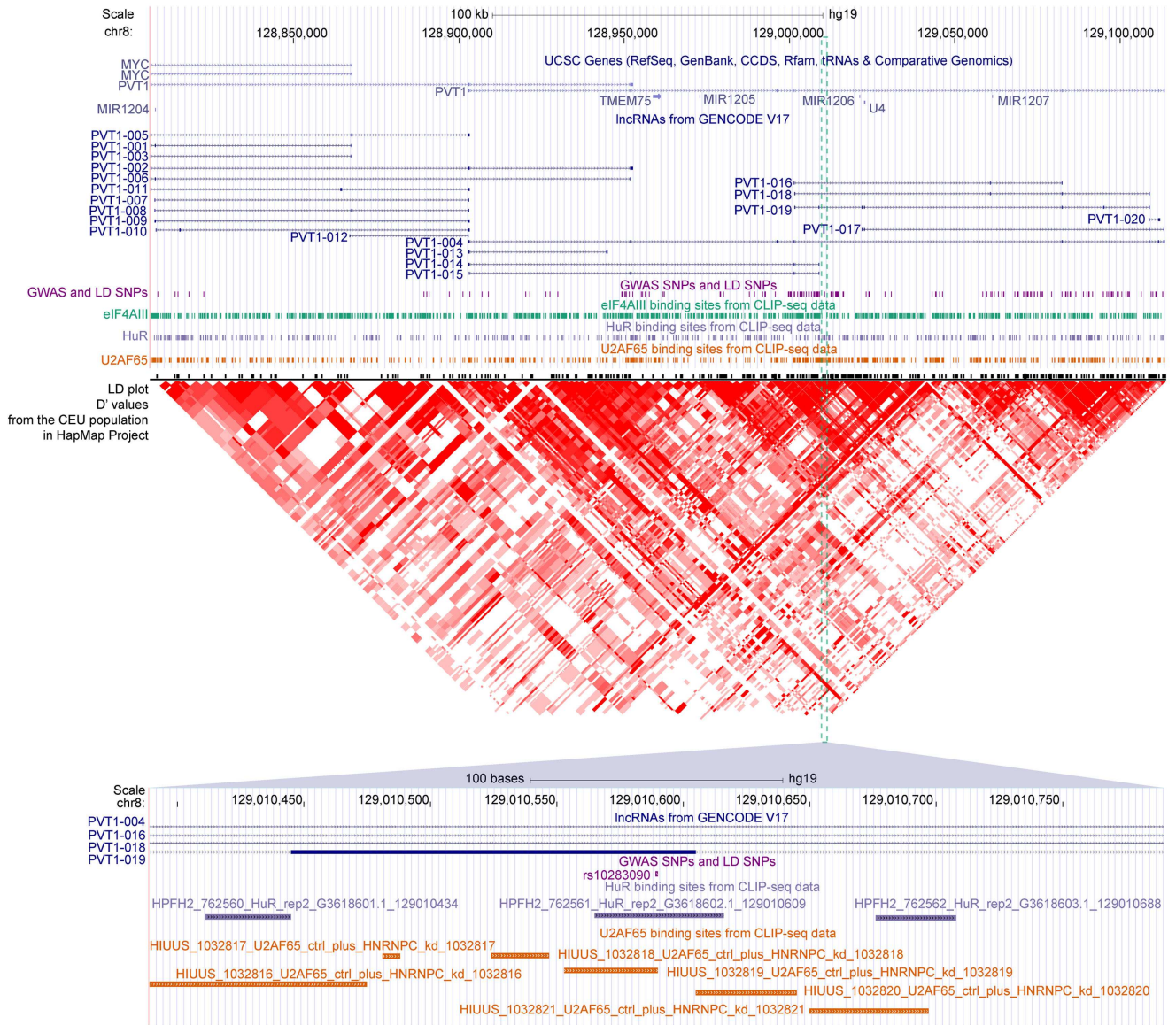

FIGURE 6 | The GWAS-associated SNPs and binding sites of three RBPs in the locus of PVT1. Gene annotations from UCSC, lncRNAs from GENCODE, GWAS, and LD SNPs, binding sites of eIF4AIII/HuR/U2AF65 and LD plot from HapMap are shown accordingly. The SNP rs10283090 overlapped with binding sites of HuR and U2AF65 are zoomed in at the bottom panel.

(Bochenek et al., 2013; Mirza et al., 2014), our approaches focused on SNPs that might impact on the binding events between RBPs and lncRNAs. Since most lncRNAs fulfill their roles through by forming complex with their protein partners, our results provide insights on the functions of lncRNAs from the perspective of RBP binding malfunction in diseases, which in turn may contribute to disease etiology.

Overall, our studies and the accompanying datasets demonstrated that one single lncRNA will generally be bound and regulated by one or multiple RBPs, the combination of which may coordinately determine the final regulatory outcome. We have also shown that an exhaustive and high-resolution RBP–lncRNA interaction map will help to discover genetic variations that contribute to complex genetic diseases by affecting post-transcriptional gene regulation.

AUTHOR CONTRIBUTIONS

Jian-Hua Yang, Liang-Hu Qu, and Jun-Hao Li conceived the project. Jun-Hao Li, Shun Liu, Ling-Ling Zheng, and Jian-Hua Yang performed the computational and statistical analysis. Jun-Hao Li, Shun Liu, Ling-Ling Zheng, Jian-Hua Yang, Liang-Hu Qu, Jie Wu, Wen-Ju Sun, Ze-Lin Wang, and Hui Zhou wrote the manuscript. Liang-Hu Qu and Jian-Hua Yang supervised the project. All authors read and approved the final manuscript.

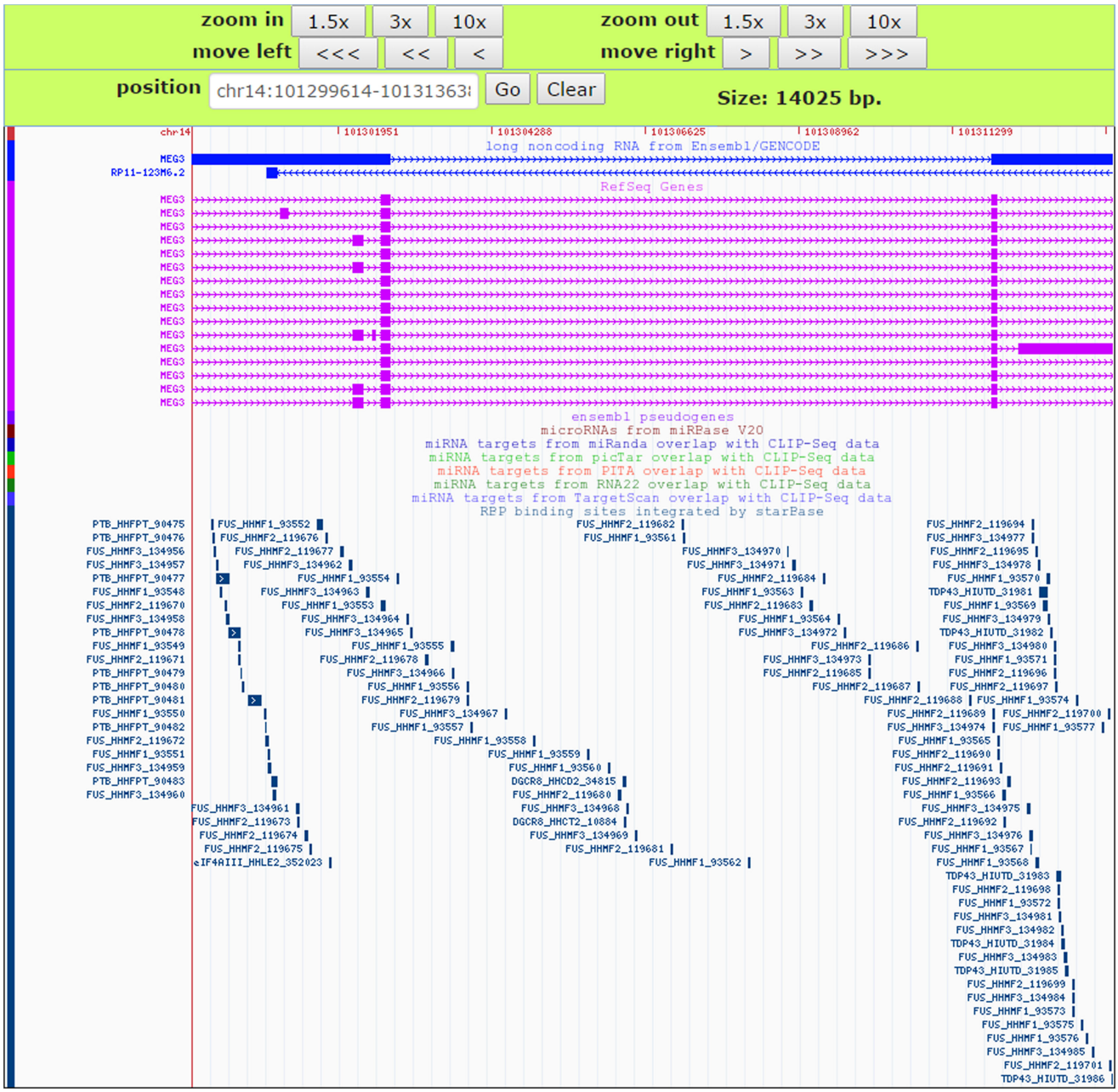

FIGURE 7 | An instance for displaying RBPs target sites in the deepView Browser of starBase V2.0. The predictive FUS binding sites on MEG3 are visible in the RBP binding sites track. In this track, the binding sites of other RBPs such as TDP-43 and PTB on MEG3 are also showed, which facilitates comparative analysis of binding events of multiple RBPs.

ACKNOWLEDGMENTS

This research is supported by the Ministry of Science and Technology of China, National Basic Research Program (No. 2011CB811300); the National Natural Science Foundation of China (No. 91440110, 31230042, 31370791, 31471223, 31401975); the funds from Guangdong Province (No. S2012010010510, S2013010012457); the project of Science and Technology New Star in ZhuJiang Guangzhou city (No. 2012J2200025); Fundamental Research Funds for the Central Universities (No. 2011330003161070, 14lgjc18); China Postdoctoral Science Foundation (No. 200902348). This research is supported in part by the Guangdong Province Key Laboratory of Computational Science and the Guangdong Province Computational Science Innovative Research Team.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at http://www.frontiersin.org/Journal/10.3389/fbioe.2014.00088/abstract

Table S1 | The summary of binding sites distribution across genomic features for 47 human RBPs.

Table S2 | Counts of binding RBPs in the 12,255 human lncRNAs.

Table S3 | Disease/trait related SNPs collected from four public GWAS databases.

Table S4 | Disease/trait related SNPs overlapped with RBP binding sites in lncRNAs.

REFERENCES

Alvarez, M. L., and DiStefano, J. K. (2011). Functional characterization of the plasmacytoma variant translocation 1 gene (PVT1) in diabetic nephropathy. *PLoS ONE* 6:e18671. doi:10.1371/journal.pone.0018671

Alwohhaib, M., Alwaheeb, S., Alyatama, N., Dashti, A. A., Abdelghani, A., and Hussain, N. (2014). Single nucleotide polymorphisms at erythropoietin, superoxide dismutase 1, splicing factor, arginine/serin-rich 15 and plasmacytoma variant translocation genes association with diabetic nephropathy. *Saudi J. Kidney Dis. Transpl.* 25, 577–581.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets – update. *Nucleic Acids Res.* 41, D991–D995. doi:10.1093/nar/gks1193

Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The genetic association database. *Nat. Genet.* 36, 431–432. doi:10.1038/ng0504-431

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* 57, 289–300.

Bochenek, G., Hasler, R., El Mokhtari, N. E., Konig, I. R., Loos, B. G., Jepsen, S., et al. (2013). The large non-coding RNA ANRIL, which is associated with atherosclerosis, periodontitis and several forms of cancer, regulates ADIPOR1, VAMP3 and C11ORF10. *Hum. Mol. Genet.* 22, 4516–4527. doi:10.1093/hmg/ddt299

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927. doi:10.1101/gad.17446611

Cancer Genome Atlas Research Network. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068. doi:10.1038/nature07385

Chendrimada, T. P., Finn, K. J., Ji, X. J., Baillat, D., Gregory, R. I., Liebhaber, S. A., et al. (2007). MicroRNA silencing through RISC recruitment of eIF6. *Nature* 447, 823–U821. doi:10.1038/nature05841

Chi, S. W., Zang, J. B., Mele, A., and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460, 479–486. doi:10.1038/nature08170

Da Sacco, L., Baldassarre, A., and Masotti, A. (2012). Bioinformatics tools and novel challenges in long non-coding RNAs (lncRNAs) functional analysis. *Int. J. Mol. Sci.* 13, 97–114. doi:10.3390/ijms13010097

Di Ruscio, A., Ebralidze, A. K., Benoukraf, T., Amabile, G., Goff, L. A., Terragni, J., et al. (2013). DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature* 503, 371–376. doi:10.1038/nature12598

Duan, R., Pak, C., and Jin, P. (2007). Single nucleotide polymorphism associated with mature miR-125a alters the processing of pri-miRNA. *Hum. Mol. Genet.* 16, 1124–1131. doi:10.1093/hmg/ddm062

Fu, X.-D. (2014). Non-coding RNA: a new frontier in regulatory biology. *Natl. Sci. Rev.* 1, 190–204. doi:10.1016/j.pt.2013.04.003

Fu, X. D., and Ares, M. Jr. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.* 15, 689–701. doi:10.1038/nrg3778

Gong, C., and Maquat, L. E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3'UTRs via Alu elements. *Nature* 470, 284–288. doi:10.1038/nature09701

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 477, 295–300. doi:10.1038/nature10398

Guttman, M., and Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature* 482, 339–346. doi:10.1038/nature10887

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* 141, 129–141. doi:10.1016/j.cell.2010.03.009

Hanson, R. L., Craig, D. W., Millis, M. P., Yeatts, K. A., Kobes, S., Pearson, J. V., et al. (2007). Identification of PVT1 as a candidate gene for end-stage renal disease in type 2 diabetes using a pooling-based genome-wide single nucleotide polymorphism association study. *Diabetes* 56, 975–983. doi:10.2337/db06-1072

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774. doi:10.1101/gr.135350.111

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 153, 654–665. doi:10.1016/j.cell.2013.03.043

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., et al. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367. doi:10.1073/pnas.0903103106

Hu, Y., Liu, C. M., Qi, L., He, T. Z., Shi-Guo, L., Hao, C. J., et al. (2011). Two common SNPs in pri-miR-125a alter the mature miRNA expression and associate with recurrent pregnancy loss in a Han-Chinese population. *RNA Biol.* 8, 861–872. doi:10.4161/rna.8.5.16034

Huang, Y. H., Wu, C. C., Chou, C. K., and Huang, C. Y. (2011). A translational regulator, PUM2, promotes both protein stability and kinase activity of Aurora-A. *PLoS ONE* 6:e19718. doi:10.1371/journal.pone.0019718

Hubbard, T. J., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., et al. (2009). Ensembl 2009. *Nucleic Acids Res.* 37, D690–D697. doi:10.1093/nar/gkn828

International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58. doi:10.1038/nature09298

Jendrzejewski, J., He, H., Radomska, H. S., Li, W., Tomsic, J., Liyanarachchi, S., et al. (2012). The polymorphism rs944289 predisposes to papillary thyroid carcinoma through a large intergenic noncoding RNA gene of tumor suppressor type. *Proc. Natl. Acad. Sci. U.S.A.* 109, 8646–8651. doi:10.1073/pnas.1205654109

Johnson, A. D., and O'Donnell, C. J. (2009). An open access database of genome-wide association results. *BMC Med. Genet.* 10:6. doi:10.1186/1471-2350-10-6

Kaneko, S., Son, J., Shen, S. S., Reinberg, D., and Bonasio, R. (2013). PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1258–1264. doi:10.1038/nsmb.2700

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., et al. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 106, 11667–11672. doi:10.1073/pnas.0904715106

Kim, H. H., Kuwano, Y., Srikantan, S., Lee, E. K., Martindale, J. L., and Gorospe, M. (2009). HuR recruits let-7/RISC to repress c-Myc expression. *Genes Dev.* 23, 1743–1748. doi:10.1101/gad.1812509

Kin, T., Yamada, K., Terai, G., Okida, H., Yoshinari, Y., Ono, Y., et al. (2007). fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.* 35, D145–D148. doi:10.1093/nar/gkl837

Konig, J., Zarnack, K., Luscombe, N. M., and Ule, J. (2011). Protein-RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet.* 13, 77–83. doi:10.1038/nrg3141

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi:10.1101/gr.092759.109

Kumar, V., Westra, H. J., Karjalainen, J., Zhernakova, D. V., Esko, T., Hrdlickova, B., et al. (2013). Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet.* 9:e1003201. doi:10.1371/journal.pgen.1003201

Kung, J. T. Y., Colognori, D., and Lee, J. T. (2013). Long noncoding RNAs: past, present, and future. *Genetics* 193, 651–669. doi:10.1534/genetics.112.146704

Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42, D92–D97. doi:10.1093/nar/gkt1248

Liao, Q., Xiao, H., Bu, D., Xie, C., Miao, R., Luo, H., et al. (2011). ncFANs: a web server for functional annotation of long non-coding RNAs. *Nucleic Acids Res.* 39, W118–W124. doi:10.1093/nar/gkr432

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39, 1181–1186. doi:10.1038/ng1007-1181

Meyer, L. R., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Kuhn, R. M., Wong, M., et al. (2013). The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.* 41, D64–D69. doi:10.1093/nar/gks1048

Mirza, A. H., Kaur, S., Brorsson, C. A., and Pociot, F. (2014). Effects of GWAS-associated genetic variants on lncRNAs within IBD and T1D candidate loci. *PLoS ONE* 9:e105723. doi:10.1371/journal.pone.0105723

Napoli, I., Mercaldo, V., Boyl, P. P., Eleuteri, B., Zalfa, F., De Rubeis, S., et al. (2008). The fragile X syndrome protein represses activity-dependent translation through CYFIP1, a new 4E-BP. *Cell* 134, 1042–1054. doi:10.1016/j.cell.2008.07.031

Ning, S., Zhao, Z., Ye, J., Wang, P., Zhi, H., Li, R., et al. (2014). LincSNP: a database of linking disease-associated SNPs to human large intergenic non-coding RNAs. *BMC Bioinformatics* 15:152. doi:10.1186/1471-2105-15-152

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129, 1311–1323. doi:10.1016/j.cell.2007.05.022

Ryan, B. M., Robles, A. I., and Harris, C. C. (2010). Genetic variation in microRNA networks: the implications for cancer research. *Nat. Rev. Cancer* 10, 389–402. doi:10.1038/nrc2867

Schmitz, K. M., Mayer, C., Postepska, A., and Grummt, I. (2010). Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev.* 24, 2264–2269. doi:10.1101/gad.590910

Sethupathy, P., and Collins, F. S. (2008). MicroRNA target site polymorphisms and human disease. *Trends Genet.* 24, 489–497. doi:10.1016/j.tig.2008.07.004

Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., et al. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* 39, 925–938. doi:10.1016/j.molcel.2010.08.011

Ulitsky, I., and Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell* 154, 26–46. doi:10.1016/j.cell.2013.06.020

Wang, K. C., and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Mol. Cell* 43, 904–914. doi:10.1016/j.molcel.2011.08.018

Wang, K. C., Yang, Y. W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472, 120–124. doi:10.1038/nature09819

Wang, P., Xue, Y., Han, Y., Lin, L., Wu, C., Xu, S., et al. (2014). The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation. *Science* 344, 310–313. doi:10.1126/science.1251456

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006. doi:10.1093/nar/gkt1229

Yin, Q. F., Yang, L., Zhang, Y., Xiang, J. F., Wu, Y. W., Carmichael, G. G., et al. (2012). Long noncoding RNAs with snoRNA ends. *Mol. Cell* 48, 219–230. doi:10.1016/j.molcel.2012.07.033

Zhao, J., Ohsumi, T. K., Kung, J. T., Ogawa, Y., Grau, D. J., Sarma, K., et al. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell* 40, 939–953. doi:10.1016/j.molcel.2010.12.011

Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J., and Lee, J. T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* 322, 750–756. doi:10.1126/science.1163045

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 19 October 2014; accepted: 22 December 2014; published online: 14 January 2015.

Citation: Li J-H, Liu S, Zheng L-L, Wu J, Sun W-J, Wang Z-L, Zhou H, Qu L-H and Yang J-H (2015) Discovery of protein–lncRNA interactions by integrating large-scale CLIP-Seq and RNA-Seq datasets. Front. Bioeng. Biotechnol. ***2****:88. doi: 10.3389/fbioe.2014.00088*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Bioengineering and Biotechnology.