D104–D114 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 5 October 2016*
doi: 10.1093/nar/gkw888

POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins

Boqin Hu[1,†], Yu-Cheng T. Yang[1,2,†], Yiming Huang[1], Yumin Zhu[1] and Zhi John Lu[1,*]

[1]MOE Key Laboratory of Bioinformatics, Center for Synthetic and Systems Biology, Center for Plant Biology and Tsinghua-Peking Joint Center for Life Sciences, School of Life Sciences, Tsinghua University, Beijing 100084, China and [2]Department of Statistics, University of California Los Angeles, Los Angeles, CA 90095-1554, USA

Received July 08, 2016; Revised September 23, 2016; Accepted September 27, 2016

ABSTRACT

We present POSTAR (http://POSTAR.ncrnalab.org), a resource of POST-trAnscriptional Regulation coordinated by RNA-binding proteins (RBPs). Precise characterization of post-transcriptional regulatory maps has accelerated dramatically in the past few years. Based on new studies and resources, POSTAR supplies the largest collection of experimentally probed (∼23 million) and computationally predicted (approximately 117 million) RBP binding sites in the human and mouse transcriptomes. POSTAR annotates every transcript and its RBP binding sites using extensive information regarding various molecular regulatory events (e.g., splicing, editing, and modification), RNA secondary structures, disease-associated variants, and gene expression and function. Moreover, POSTAR provides a friendly, multi-mode, integrated search interface, which helps users to connect multiple RBP binding sites with post-transcriptional regulatory events, phenotypes, and diseases. Based on our platform, we were able to obtain novel insights into post-transcriptional regulation, such as the putative association between CPSF6 binding, RNA structural domains, and Li-Fraumeni syndrome SNPs. In summary, POSTAR represents an early effort to systematically annotate post-transcriptional regulatory maps and explore the putative roles of RBPs in human diseases.

INTRODUCTION

The regulatory maps of genomes have been revealed by various high-throughput sequencing assays from individual groups and consortium efforts such as the ENCODE project (1) and Roadmap Epigenomics project (2). Most previous studies have focused on *cis*-regulation, the epigenome, the transcriptome, and the proteome, leaving post-transcriptional regulation not fully explored and connected with existing knowledge. Although RNA splicing has been well-studied, other post-regulatory signatures, such as RNA modification and RNA editing, have not been profiled until recently (3–5). Furthermore, most studies of genomic variants (e.g. GWAS SNPs (6) and cancer somatic mutations (7)) have mainly focused on the transcriptional level (8,9). Recently, there has been increasing interest in studying the association between post-transcriptional regulation and disease-associated variation (10). In addition, post-transcriptional regulation could regulate cell differentiation (11) and influence the scope of the druggable proteome (12). However, post-transcriptional interactions between cell state-associated genes and druggable genes and their impacts on cell differentiation, pathology and clinical treatment remain largely uncharacterized. To perform such studies, researchers require a platform that facilitates integration and association of multi-layer information to illuminate the mechanisms underlying post-transcriptional regulation.

RNA transcripts do not function as naked RNAs in eukaryotic cells from birth to death; instead, they are dynamically bound by various post-transcriptional regulatory factors, including RNA-binding proteins (RBPs) and microRNAs (miRNAs) (13,14). Recently developed high-throughput assays (e.g. CLIP-seq and RNAcompete technologies), together with computational tools, enabled researchers to obtain transcriptome-wide binding maps of RBPs and miRNAs at high resolution (15,16). Most RNA processing reactions (e.g. alternative splicing, alternative polyadenylation and nucleotide modification) and regulatory events (e.g. subcellular localization, RNA stability and translation efficiency) are mediated by miRNAs and RBPs. Constructing accurate and comprehensive RBP-RNA and miRNA-RNA interaction maps at high resolution is a necessary step toward interpreting their mechanistic roles in post-transcriptional regulation.

Although there are several databases for CLIP-seq-derived RBP binding sites, such as CLIPZ (17), starBase (18), DoRiNA (19) and CLIPdb (20), they merely provide

*To whom correspondence should be addressed. Tel: +86 10 62789217; Fax: +86 10 62789217; Email: zhilu@tsinghua.edu.cn
†These authors contributed equally to this work as first authors.

Table 1. Overview of data curated in POSTAR

	Category	Human	Mouse	Resource/calculation method[a]
RBP binding sites	RBP binding sites from experiments	1 752 329	1 003 984	**All CLIP-seq peaks called by Piranha (human: 65 RBPs; mouse: 30 RBPs)[b]**
		39 201	78 922	**HITS-CLIP peaks called by CIMS (human: 17 RBPs; mouse: 23 RBPs)[b]**
		7 731 846	96 346	**PAR-CLIP peaks called by PARalyzer (human: 44 RBPs; mouse: 4 RBPs)[b]**
		4 598 307	1 013 008	**iCLIP peaks called by CITS (human: 9 RBPs; mouse: 8 RBPs)[b]**
		6 703 559	NA	eCLIP-seq peaks called by ENCODE (human: 56 RBPs)[c]
		439 817	NA	PIP-seq peaks called by PMID24393486 (human: global RBPs)
	RBP binding site from predictions	25 623 567	18 540 386	**Peaks predicted by FIMO (human: 88 RBPs; mouse: 88 RBPs)[d]**
		19 447 967	24 621 203	**Peaks predicted by TESS (human: 88 RBPs; mouse: 88 RBPs)[d]**
		16 586 127	11 905 150	**Peaks predicted by DeepBind (human: 82 RBPs; mouse: 82 RBPs)[e]**
Data module I: Gene/RBP annotations	RBPs	132	104	Ensembl, PMID25365966
	Sequence motifs	726	180	**MEME, HOMER**
	Structural preferences	720	179	**RNApromo, RNAcontext**
	Gene Ontologies	15 677	13 849	GOBP, GOMF, GOCC[f]
	Biological pathways	186	105	KEGG
	Gene expression	34 cells/tissue types	18 cell/tissue types	**TopHat, Cufflinks[g]**
	Alternative splicing (skip exon)	34 cells/tissue types	18 cell/tissue types	**TopHat, MISO[g]**
Data module II: Molecular annotations	miRNA binding sites from experiments	3 906 955	1 588 861	**AGO CLIP-seq peaks called by Piranha, the targeting miRNAs identified by miRanda[h]**
	miRNA binding sites from predictions	70 516 087	38 336 372	**RNAhybrid, TargetScan, miRanda**
	RNA modification sites	177 049	91 930	RMBase, PMID26863196
	RNA editing sites	2 583 302	8846	RADAR, DARNED
	Splicing elements	1 995 574	1 152 186	Anno. in GENCODE human v19, mouse vM7
	Conserved structural regions	725	691	EvoFam
Data module III: Genomic variants	SNPs	149 398 310	77 785 586	dbSNP v146
	Tissue-specific eQTL	19 530 607	NA	GTEx
	GWAS SNPs	278 473	NA	GWASdb2, **RNAfold[i]**
	Clinically important SNPs	131 919	NA	ClinVar, **RNAfold[i]**
	Cancer TCGA whole-exome SNVs	828 119	NA	PMID24390350, **RNAfold[i]**
	Cancer TCGA whole-genome SNVs	4 745 891	NA	PMID23945592, **RNAfold[i]**
	Cancer COSMIC SNVs	2 371 219	NA	COSMIC v76, **RNAfold[i]**
Data module IV: Gene-Function associations	Tissue-specific genes	21 549	NA	TiGER, SpeCond
	Gene-Disease associations	419 906	NA	OMIM, DisGeNET
	Gene-Cancer associations	4485	NA	**Manually curated from 60 publications[j]**
	Gene-Drug associations	35 201	NA	DGIdb 2.0
Data module V: RNA secondary structures	Predicted local structures	82 242 543	57 095 233	**RNAfold with restraints from experimental structural probing data (human: DMS-seq, PARS; mouse: icSHAPE, Frag-seq, CIRS-seq)[k]**

[a]Results and data firstly generated by POSTAR are in bold font.
[b]We provide all CLIP-seq peaks called by Piranha with $P < 0.01$. For CIMS, CITS and PARalyzer, we provide peaks with default significance cutoffs.
[c]See Supplementary File 2 for the full list of eCLIP-seq data. The peaks were called by ENCODE.
[d]See Supplementary File 5 for the RBPs and motifs used for prediction.
[e]See Supplementary File 6 for the RBPs in DeepBind model.
[f]BP, Biological Process; MF, Molecular Function; CC, Cellular Component.
[g]See Supplementary File 4 for the full list of 230 RNA-seq data sets in human and mouse.
[h]We used all AGO CLIP-seq peaks called by Piranha ($P < 0.01$). The targeting miRNAs of the peaks were identified using miRanda with default parameters.
[i]We used RNAfold to calculate the minimal free energy changes of local RNA secondary structures that are induced by the mutations.
[j]See Supplementary File 3 for the full list of manually curated cancer genes.
[k]See Supplementary File 7 for the experimental structural probing datasets. We predicted one local structure centered on each RBP binding site (window size: 150nt).

repositories of transcriptome-wide RBP binding sites or focus on miRNA-mediated post-transcriptional regulation. Later, RBP-Var (10) was developed to incorporate SNV information for RBP binding sites, but it was limited to a few regulatory events in humans. Furthermore, experimental-data-constrained RNA secondary structures are not available for RBP binding sites in any current database. Finally, it is often desirable to know whether RBPs can interact with RNA molecules, especially novel lncRNAs; however, only a few online tools (21,22) are available for predicting RBP binding sites on given RNA sequences.

Previously, we developed CLIPdb (20), which simply provided RBP binding sites without further annotation and interpretation (http://CLIPdb.ncrnalab.org). Here, we constructed a new platform for CLIPdb version 2, POSTAR, which focuses on POST-trAnscriptional Regulation coordinated by RNA-binding proteins (RBPs), to facilitate searching, annotation, visualization, integration, connection, and interpretation of data regarding multiple post-transcriptional regulatory events in humans and mice. First and foremost, POSTAR provides a comprehensive repository of experimentally probed (i.e. derived from 498 CLIP-

seq and 151 eCLIP-seq data sets) and computationally predicted RBP binding sites in humans and mice. Based on these binding sites, POSTAR annotates a gene/lncRNA and its RBP binding sites using extensive information: (i) two kinds of RBP binding motifs/preferences, (ii) six types of molecular regulation events, (iii) six types of genomic variants, (iv) four types of gene-function associations and (v) predicted RNA secondary structure around every RBP binding site based on whole-transcriptome RNA structural profiling data (Figure 1, Table 1). Furthermore, we designed a multi-mode usage interface: (a) 'POSTAR' search, (b) 'RBP' search, (c) 'Structure' visualization, (d) 'Variation' search, (e) 'Functional gene' search and (f) 'Predict' server for RBP binding prediction for given RNA sequences (Figure 2A). POSTAR presents the search/prediction results in many ways (Figure 2B). Moreover, binding sites from multiple RBPs and their associations with various post-transcriptional regulatory events can be visualized and explored in an integrative manner (Figure 3), which will allow users to connect different pieces of data from various resources and layers.

DATA COLLECTION AND PROCESSING

Data source

POSTAR focuses on RBP binding sites in the human and mouse transcriptomes. We first obtained 338 processed data sets from CLIPdb (20). We also collected and processed 160 new CLIP-seq data sets using the same pipelines (Supplementary File 1). The data contain three CLIP-seq data types, including HITS-CLIP, PAR-CLIP and iCLIP. Moreover, we also incorporated 151 eCLIP-seq data sets (in HepG2 and K562 cells) that were released by the ENCODE consortium (23). The eCLIP-seq binding sites/peaks were directly downloaded from the ENCODE data portal (https://www.encodeproject.org, NOV 2015) (Supplementary File 2). In addition, we included genome-wide RBP binding sites profiled by PIP-seq technology (24).

To annotate and interpret RBP binding sites, we retrieved conserved structural regions from EvoFam (25), RNA modification sites from RMBase (26) and a recent publication (27), RNA editing sites from RADAR (28) and DARNED (29), single nucleotide polymorphisms (SNPs) from dbSNP version 146 (30), human trait/disease-associated SNPs from GWASdb2 (31) and ClinVar (32), and cancer somatic mutations from whole-exome sequencing data (33), whole-genome sequencing data (34), and the COSMIC database (35). We calculated the mean conservation scores for the RBP binding sites using genome-wide phastCons (36) and phyloP (37) intensities. We obtained human tissue-specific eQTLs from GTEx (38). We did not include eQTL annotation for mice because systematic eQTL mapping across multiple tissue types was unavailable. To better annotate RBP targets at the gene level, we collected cell-specific genes from TiGER (39) and SpeCond (40), human disease-associated genes from OMIM (41) and DisGeNET (42), cancer-associated genes from 60 publications (4485 cancer-associated genes across 36 cancer types, see Supplementary File 3), and druggable genes from DGIdb (43). In addition, we provided basic annotations of RBPs, including gene symbol, gene ID and domain information (44). We also collected 230 RNA-seq samples from 34 human tissues/cell types and 18 mouse tissues/cell types (Supplementary File 4). Detailed descriptions and statistics for these data resources can be found in Table 1 and Figure 1.

Data re-annotation

The raw data used in POSTAR were highly heterogeneous, because the data resources were collected from various publications and databases. Therefore, we processed and re-annotated the collected and computed data. First, the genomic coordinates of all data resources were converted to hg19 and mm10 using the LiftOver utility from the UCSC Genome Browser database (45). We also unified different IDs (e.g. RefSeq and UCSC gene ID) from various data sets into Ensembl gene IDs (46) using BioMart (47). We used GENCODE (human V19 and mouse V7) (48) for the annotation of regulatory elements (including validated and predicted RBP/miRNA binding sites, splicing *cis*-elements, and RNA modification and editing sites), trait/disease-associated variations, and various functional genes. The positions of splicing *cis*-elements were defined as: −3 to +8 nucleotides for 5′ splice sites and −12 to +2 nucleotides for 3′ splice sites (49). We annotated each regulatory element with its genomic strand, associated gene, genomic element, reference literature, etc. The annotation of genomic elements is based on the following priority: CDS, canonical ncRNA (including miRNA, snRNA, snoRNA, tRNA, rRNA and miscellaneous RNA), 3′ UTR, 5′ UTR, lncRNA exon, pseudogene, intron (mRNA and lncRNA), intergenic region, and others. Here, intergenic regions were defined as regions at a distance 2000 nt away from any genic regions (coding genes, ncRNAs and pseudogenes).

RBP binding site identification and prediction

We followed the computational pipeline used in our CLIPdb to identify binding sites from CLIP-seq data sets (20). First, we used the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit) to pre-process the CLIP-seq data sets in a uniform procedure. Next, we identified the RBP binding sites of all CLIP-seq samples using Piranha (P-value < 0.01), which is applicable to all variations of CLIP-seq technology (50). In addition, we also provided binding sites called by specialized tools for different CLIP-seq technologies: PARalyzer for PAR-CLIP (51), CIMS for HITS-CLIP (52) and CITS (a module in CIMS software) for iCLIP (53).

To expand our RBP binding site repository, we used several computational tools to predict putative RBP binding sites across the human and mouse transcriptomes. RBPs interact with their RNA targets via specific RNA recognition motifs, which have been extensively determined using various technologies (54,55). Therefore, we used position weight matrix (PWM) motif matching to predict genome-wide RBP binding sites. All PWMs for 88 human/mouse RBPs collected from the literature (Supplementary File 5) were used to call motif matches in the human and mouse transcriptomes. We scanned for each RBP PWM within 50-nt genome intervals using FIMO (56) (present if the P-value was $<$1e−4) and TESS (57) (present if the log-odds score

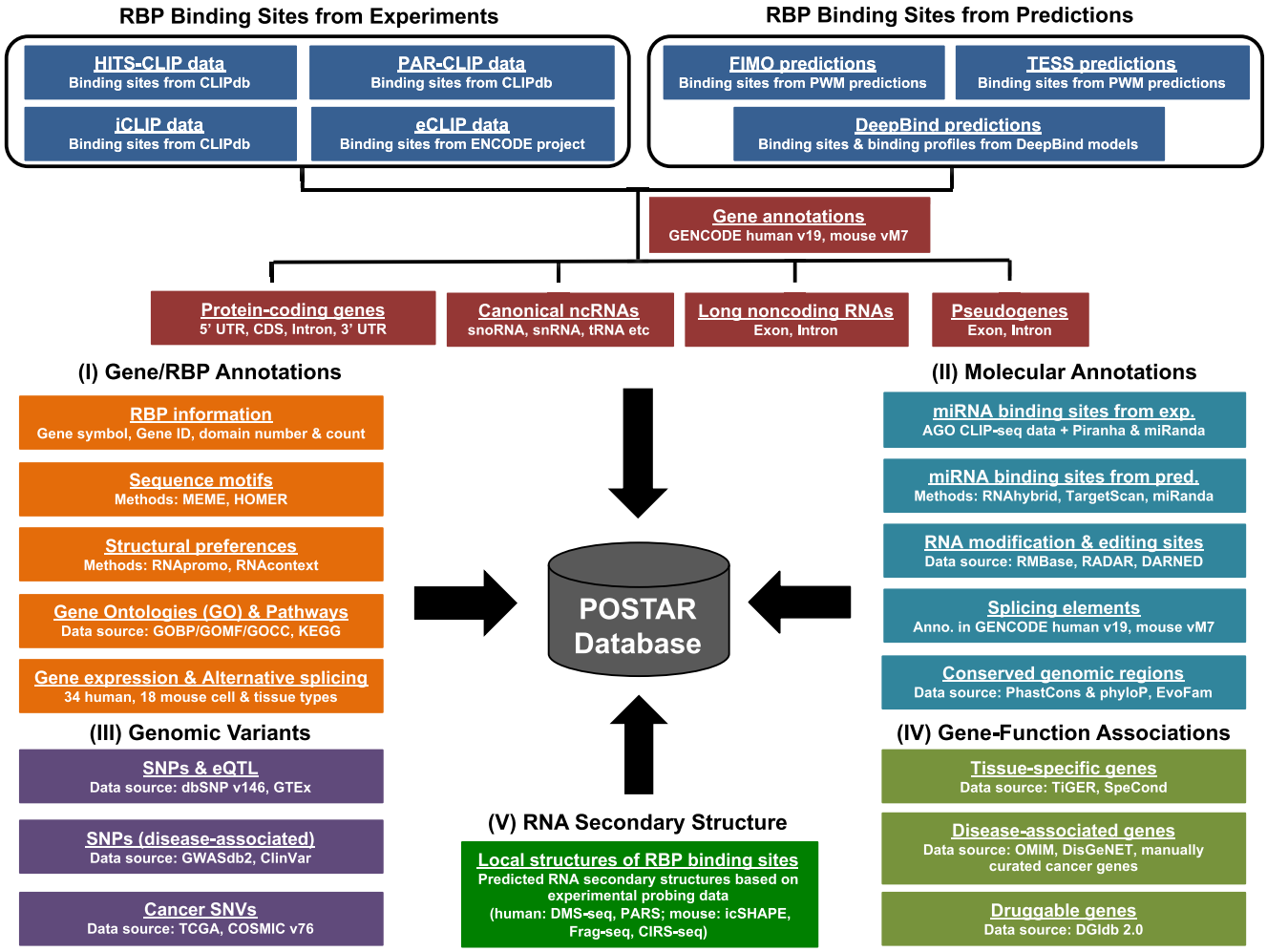

Figure 1. Multiple data modules in POSTAR can be used to annotate and interpret RBP binding sites at various levels. Experimentally probed and computationally predicted RBP binding sites were annotated with different genomic elements. The annotations and functions of RBPs and genes, as well as the predicted sequence motifs and structural preferences of RBPs, were provided (data module I). The RBP binding sites were annotated using extensive information at several levels, including molecular regulatory events (data module II), genomic variants (data module III), gene-function associations (data module IV), and RNA secondary structures (data module V).

was >7.0). We also used DeepBind (58), a deep learning-based tool, with default parameters to predict the binding strength of 50-nt genome intervals for 82 human/mouse RBPs (Supplementary File 6). The 50-nt genome intervals with top 5‰ binding strength were considered as the binding sites for each RBP model in DeepBind.

miRNA binding site identification and prediction within RBP binding sites

miRNA binding sites from experiments (i.e., derived from AGO CLIP-seq data sets) were annotated within the RBP binding sites. AGO CLIP-seq data sets experimentally identified miRNA–target interactions in a genome-wide manner. We used miRanda (59) to predict targeting miRNAs for AGO protein binding sites.

In addition to AGO binding regions, we used RNAhybrid (60), TargetScan (61) and miRanda (59) to screen possible miRNA binding sites within the RBP binding sites for sequences of all gene regions, including 3′ UTRs, 5′ UTRs, lncRNAs and pseudogenes. For RNAhybrid and miRanda, human and mouse miRNA sequences were obtained from miRBase (version 20) (62). We estimated proper parameters for each miRNA and gene sequence pair when using RNAhybrid. For TargetScan, human and mouse miRNA families were obtained from the TargetScan website (http://www.targetscan.org). We downloaded a 100-way genome alignment for humans and a 60-way genome alignment for mice from the UCSC Genome Browser database (45) when using TargetScan.

Sequence motifs and structural preferences of RBP binding sites

To identify the sequence and structural motifs of RBP binding sites, we used RBP binding sites identified using Piranha (P-value < 0.01). We also used default RBP binding sites downloaded from the ENCODE data portal for the eCLIP-seq data sets. Briefly, the binding sites in each CLIP-seq data set were separated into independent training (top 500 bind-

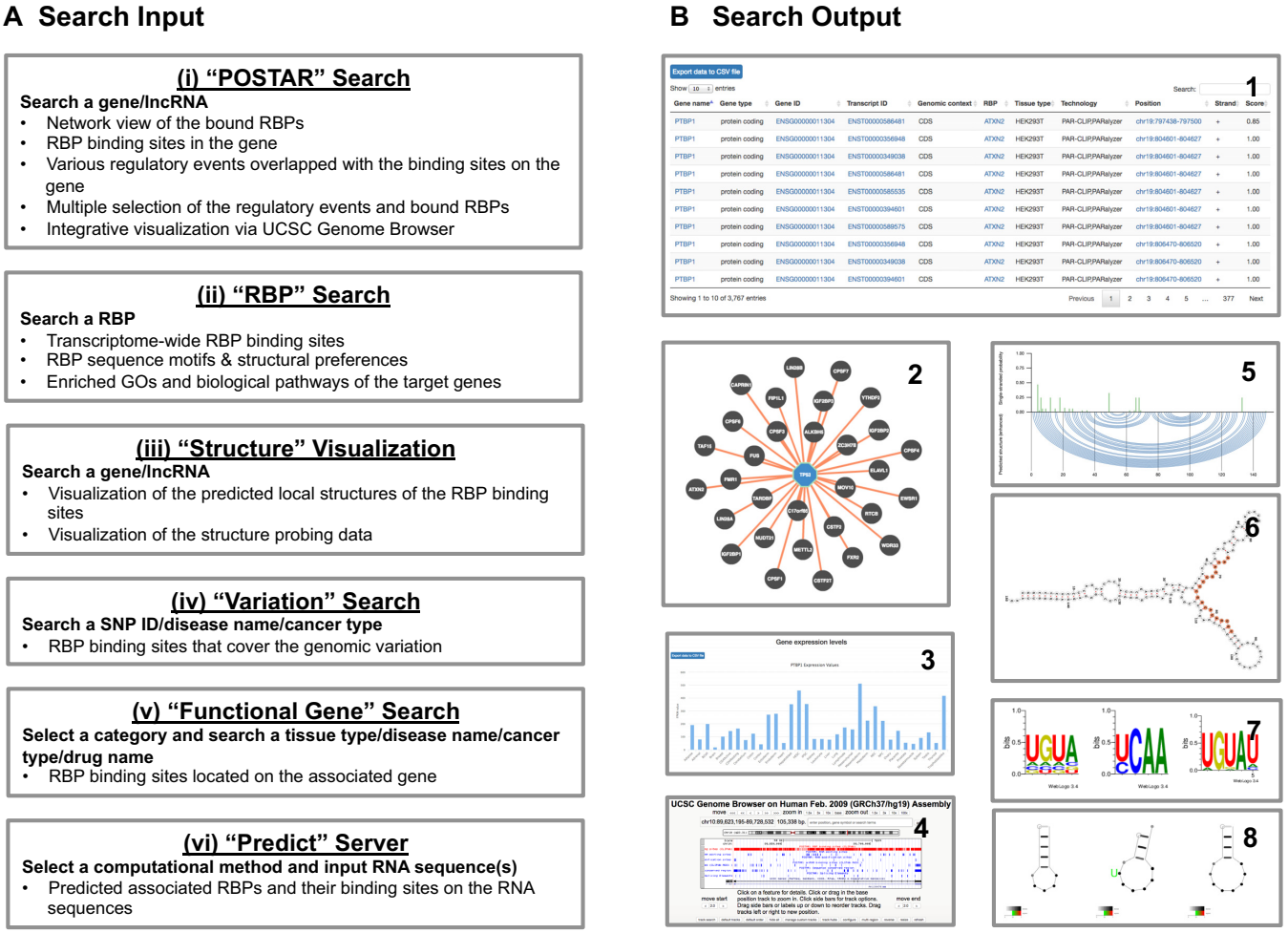

Figure 2. Input and output search interface of POSTAR: multiple search modes and multiple result viewers. (**A**) POSTAR provides six usage modes: (i) 'POSTAR' search, (ii) 'RBP' search, (iii) 'Structure' visualization, (iv) 'Variation' search, (v) 'Functional gene' search, and (vi) 'Predict' server. (**B**) POSTAR presents the search results in multiple ways. A table layout is the basic output format (1). In the 'POSTAR' search mode, the interactions between the target gene and multiple RBPs are visualized in a network (2). The expression levels of the target gene and splicing scores of skipped exons across multiple cell and tissue types are shown in a bar chart (3). Clicking on the genomic positions will direct the user to the UCSC Genome Browser, which will display any associated binding sites and regulatory events (4). In 'Structure' visualization mode, we provide RNA structural profiling data (5) and predicted RNA secondary structures based on these data (6). In 'RBP' search mode, we provide the sequence motifs (7) and structural preferences (8) of the RBP.

ing sites) and testing (binding sites ranked 501–1000) sets to ensure the quality of sequence motif discovery. If one data set contained <1000 binding sites, we defined the top half of the binding sites as the training set and the remaining sites as the testing set. We used MEME (63) to identify enriched sequence motifs in the training set for each CLIP-seq data set. We set MEME to report up to five motif models per data set, with motif width between 4 and 10 nucleotides. Next, we calculated enrichment for the initially detected motif models within the testing set using FIMO (56) and selected the three most enriched sequence motifs for each data set. In addition, we used another sequence motif finding tool, HOMER (64), to identify the three most enriched sequence motifs for each data set using the pipeline described above. Sequence motifs were visualized using WebLogo (65).

RBP binding sites were extended to at least 60 nt in length when structural preference was investigated. We used RNAcontext (66) to detect local structural motifs for each CLIP-seq data set using the pipeline described above. The structural annotation used in RNAcontext included paired (P), hairpin loop (L), bulge/internal/multi loop (M) and unstructured (U). In addition, we used another structural motif finding tool, RNApromo (67), to predict the three most enriched structural elements (P-value < 0.05) within the RBP binding sites for each CLIP-seq data set.

RNA secondary structure prediction for RBP binding sites

To explore the RNA secondary structure around every RBP binding site, we used RNAfold with optimized parameters (option: -D) (68) to predict local structure in a manner restrained by experimental structural-probing data (Supplementary File 7) (69,70). Experimental profiling data were processed and normalized based on the same RNAex protocol (70). For sequences without sufficient probing data, we used RNAfold to directly predict local structure based

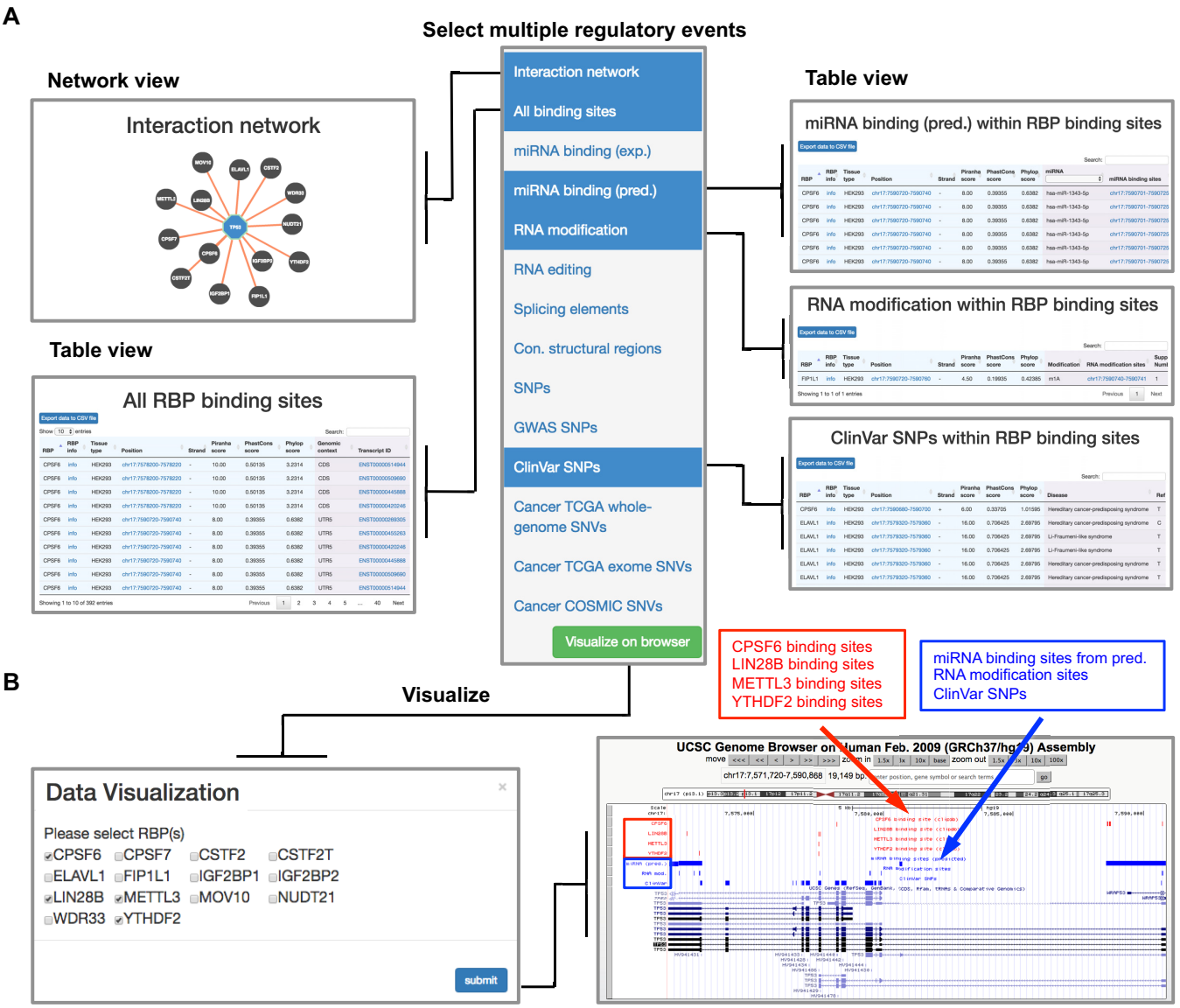

Figure 3. 'POSTAR' search enables integrative viewing of multiple RBP binding sites and their potential to post-transcriptionally regulate a target gene (TP53 as an example). (**A**) In the PAR-CLIP Piranha data module, users may select 'interaction network', 'all binding sites', and multiple regulatory elements, including 'miRNA binding (pred.)', 'RNA modification', and 'ClinVar SNPs', to obtain detailed information in one page. (**B**) By clicking on the 'Visualize in browser' button (green), a user can select four RBPs among all bound RBPs to simultaneously visualize their binding sites (red tracks) and regulatory events (blue tracks) in an integrative manner via the UCSC genome browser.

on free energy model. The folding size around the center of each RBP binding site was 150 nt.

Impact of SNVs on RNA secondary structure

SNPs (e.g. ClinVar), as well as binding sites, can be visualized on the predicted secondary structure. We calculated the folding free energy change of each SNP with RNAfold (71). Based on the human reference genome (hg19), we changed the reference allele to a corresponding alternative allele for each SNV and calculated changes in the minimal free energy (MFE) of RNA secondary structures. We used RNAfold (71) with default parameters to calculate changes in the MFE (ΔG) of RNA secondary structures ($\Delta\Delta G = \Delta G_{alt} - \Delta G_{ref}$) for each SNV.

Expression pattern and alternative splicing events detected by RNA-seq data

In addition to the cell specificity information derived from TiGER (39) and SpeCond (40), we also calculated expression specificity scores using raw RNA-seq data. All RNA-seq data sets in POSTAR were generated from Illumina GAII or HiSeq2000/HiSeq2500 systems. We filtered out low-quality reads in each RNA-seq data set using PRINSEQ (72). Next, we aligned the RNA-seq data to the human genome (hg19) or mouse genome (mm10) using Tophat v2.0.10 (73,74). Cufflinks (v2.1.1) (73,75) was used to calculate gene expression levels. Furthermore, we calculated percentage spliced in (PSI) scores for skipped exons in the human and mouse transcriptomes using MISO (76). The PSI

score denotes the fraction of mRNA that represents the inclusion isoform.

Enrichment analysis of gene ontology and biological pathways

For Gene Ontology (GO) analysis, we used topGO (77) to assess enrichment of biological process (BP), molecular function (MF) and cellular component (CC) terms for each CLIP-seq data set based on their target genes. For pathway analysis, we calculated the significance of pathway enrichment for every CLIP-seq data set using a hypergeometric test. We considered CDS, intron, 3′ UTR, and 5′ UTR separately in the analysis of GO terms and biological pathways.

Database architecture

All metadata in POSTAR are stored in a MySQL database. The web interface of POSTAR was implemented in Hyper Text Markup Language (HTML), Cascading Style Sheets (CSS) and Hypertext Preprocessor (PHP). Web design was based on the free templates of Bootstrap (http://getbootstrap.com). Visualization was implemented using the UCSC Genome Browser.

DATABASE FEATURES AND APPLICATIONS

Web interface

We provide a user-friendly web interface for users to query the database through multiple modes and predict RBP binding sites for a given RNA: (i) 'POSTAR' search, (ii) 'RBP' search, (iii) 'Structure' visualization, (iv) 'Variation' search, (v) 'Functional gene' search and (vi) 'Predict' server (Figure 2A). POSTAR presents the search/prediction results in many ways (table, bar plot, structure interface, network, etc.) (Figure 2B). We briefly introduce each query mode below.

POSTAR provides three main usage modes. (i) The 'POSTAR' search mode provides multiple RBP binding sites for a given protein coding gene or lncRNA, associated post-transcriptional regulatory events, *trans*-factors (i.e. miRNAs), and *cis*-elements (e.g. RNA modification/editing sites, splicing elements, and conserved structural regions), as well as functional SNPs (e.g. GWAS SNPs) and cancer somatic mutations. We applied the UCSC Genome Browser to visualize multiple RBP binding sites and their associated regulatory events or genomic variants. Users can search multiple RBPs and regulatory events for a gene/lncRNA to investigate potential crosstalk. Moreover, for a given gene/lncRNA, we also provide basic annotation, associated diseases, expression levels, and alternative splicing events across multiple cell and tissue types. (ii) The 'RBP' search mode provides an overview of the binding sites for a given RBP, as well as GO terms enriched in its set of target genes. In addition, users can also find the sequence and structural motifs of the binding sites of a given RBP. (iii) The 'Structure' visualization mode provides local RNA secondary structure visualization (78) for every RBP binding site. Previous studies suggest that RNA secondary structures can provide specific binding sites for RBPs and restrict protein binding by altering structural accessibility (79,80). Therefore, we provide the 'Structure' visualization mode so that users can visualize RBP binding sites and the effects of SNPs/SNVs (e.g., ClinVar) on local structure.

In addition to the main usage modes, POSTAR provides three additional usage modes. (iv) The 'Variation' search mode provides information about the effects of SNPs and disease-associated variants on RBP binding sites. (v) The 'Functional gene' search mode provides linkages between binding sites and cell state-associated genes, disease-associated genes, and druggable genes in a table layout. Although the downstream effects of these functional genes are understood relatively well, the manner in which they are regulated at the post-transcriptional level, and the impact of such regulation on their functionality, remains largely unclear. We believe that this mode could provide novel hypotheses regarding the physiological and pathological functions of RBP binding sites. (vi) Finally, in the 'Predict' mode, we provide web-based computational tools for predicting RBP binding sites on given RNA sequences provided by users. FIMO and TESS predict binding sites using the PWMs of 88 human/mouse RBPs collected from literature (Supplementary File 5). DeepBind predicts RBP binding affinity on an RNA sequence (length between 20 nt and 50 nt) using the DeepBind models of 82 human/mouse RBPs (Supplementary File 6).

Example findings using POSTAR search mode

We illustrate some example applications of POSTAR here to demonstrate exploration of multiple RBP binding sites and various post-transcriptional regulatory events in an integrative manner. Assume that users are interested in the post-transcriptional regulation of TP53, an important oncogene in humans. Users can go to the 'POSTAR' search mode and select all types of RBP binding sites. In the PAR-CLIP Piranha data module, users can simultaneously select multiple regulatory events to obtain the information shown in Figure 3A, including detailed information (e.g., genomic position, conservation score, and genomic context) about each binding site. TP53 contains 414 binding sites for 14 RBPs; for each of these RBP binding sites, users can discover associated RNA modification sites, predicted miRNA binding sites, and ClinVar SNPs. Users can associate RBP binding with other post-transcriptional regulatory events and disease-associated SNPs, obtaining novel insights into biological function and disease mechanisms. As an example, the post-transcriptional regulatory mechanism that underlies Li-Fraumeni syndrome, a condition characterized by early onset of several types of cancer and associated with germline mutations of TP53 (81), remains unexplored. Here, we identified nine Li-Fraumeni syndrome SNPs covered by five binding sites from six RBPs, CPSF6, ELAVL1, IGF2BP1, LIN28B, METTL3 and YTHDF2, which function as splicing factors, 3′ processors, and RNA modification (m^6A) 'writers' and 'readers'. Knowledge of these binding sites could help users generate new hypotheses about the molecular mechanisms of Li-Fraumeni syndrome at the post-transcriptional level. Furthermore, users can visualize the binding sites of multiple RBPs and their associated regulatory events via the UCSC genome browser (Figure 3B).

A

RBPs' binding sites | GWAS/ClinVar SNPs

RBP	Position	Strand	Disease	Ref (+)	Alt (+)	Mutation position
CPSF6	chr17:7578200-7578220	-	Li-Fraumeni syndrome	C	T	chr17:7578202-7578203
CPSF6	chr17:7578200-7578220	-	Li-Fraumeni syndrome 1	C	T	chr17:7578210-7578211
CPSF6	chr17:7578200-7578220	-	Li-Fraumeni syndrome 1	G	A	chr17:7578211-7578212
ELAVL1	chr17:7579320-7579360	-	Li-Fraumeni-like syndrome	T	A	chr17:7579328-7579329
ELAVL1	chr17:7579320-7579360	-	Li-Fraumeni-like syndrome	T	C	chr17:7579328-7579329
IGF2BP1	chr17:7577500-7577520	-	Li-Fraumeni syndrome	T	C	chr17:7577504-7577505
LIN28B	chr17:7577500-7577520	-	Li-Fraumeni syndrome	T	C	chr17:7577504-7577505
LIN28B	chr17:7577560-7577580	-	Li-Fraumeni syndrome	C	T	chr17:7577567-7577568
LIN28B	chr17:7577560-7577580	-	Li-Fraumeni syndrome	T	C	chr17:7577576-7577577
…	…	-	…	.	.	…

Search: Li-Fraumeni

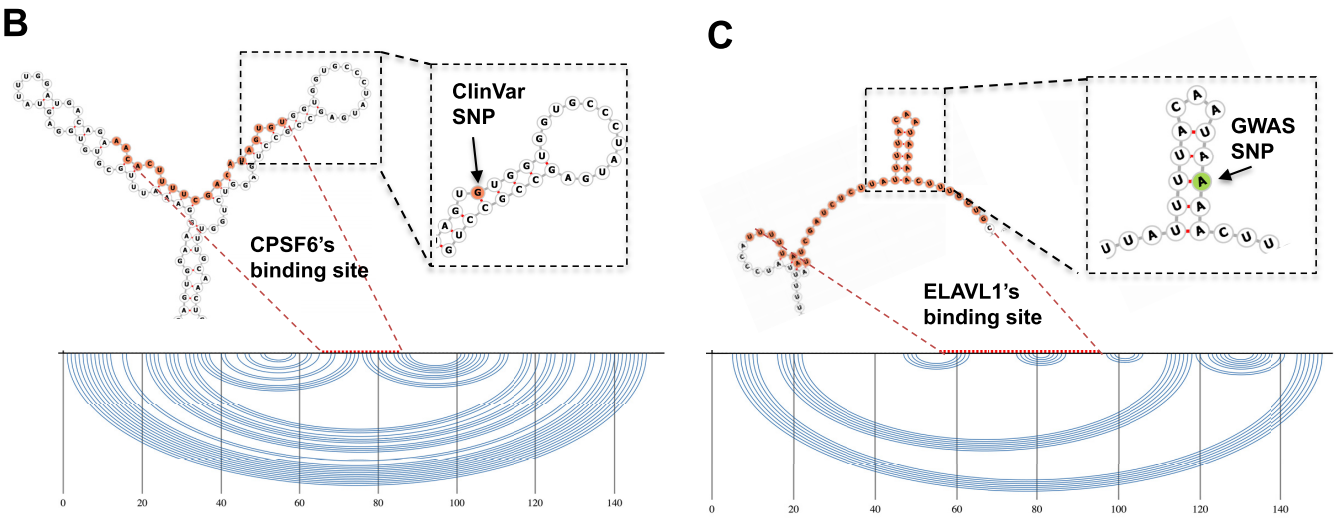

Figure 4. Local structures of RBP binding sites on TP53. (**A**) Users can search for RBP binding sites on TP53 that are associated with disease SNPs by searching for the disease name ('Li-Fraumeni syndrome' as an example here) in the table on the server. (**B**) Predicted local secondary structure centered on a CPSF6 binding site. The local secondary structure around a Li-Fraumeni syndrome SNP (from the ClinVar database), which is a G-to-A mutation on the TP53's transcript (minus-strand), is magnified; it disrupts the base pair (G-C pair) in the hairpin's stem. Note that the mutation is a C-to-T mutation (box highlight in (**A**)) as annotated by ClinVar on the plus-strand. (**C**) Another predicted local secondary structure, centered on an ELAVL1 binding site, which contains a GWAS SNP that is an A-to-U mutation.

Example findings using structure and other usage modes

Users can explore the local RNA secondary structures of RBP binding sites using the 'Structure' visualization mode. To extend the example described above, users can further investigate the local secondary structures of RBP binding sites that contain Li-Fraumeni syndrome SNPs (Figure 4A). One Li-Fraumeni syndrome SNP site that is bound by CPSF6 could disrupt the stem in the local structure (Figure 4B). The SNP and binding site are located in the coding sequence region of TP53. CPSF6 is involved in 3′ processing (82), mRNA export (83), and RNA splicing (84). Therefore, in addition to protein coding, our observation suggests that the RNA sequence in the coding region of TP53 influences protein expression via mechanisms that function at the post-transcriptional level. In another example, ELAVL1's binding site contains a skin cancer-associated GWAS SNP site, which is located in TP53's 3′ UTR. ELAVL1 (also known as HuR) functions in diverse mRNA metabolic processes, including splicing, degradation, and translation (85–87). The SNP in TP53's 3′ UTR could destroy the local stem structure of the ELAVL1 binding site (Figure 4C), which might alter the influence of ELAVL1 on TP53. These examples provide novel insight into the putative functions of RBP binding sites on RNA structural domains, which are poorly understood.

If a user desires information regarding the putative functions of RBP binding in heart diseases, the user could access the GWAS SNP interface in the 'Variation' search mode, choose the 'eCLIP ENCODE' data module, and select the phenotype term 'heart failure', at which point the database will return six heart failure-associated GWAS SNPs that overlap with 11 validated binding sites for nine RBPs. No-

tably, one heart failure-associated SNP located in the CDS of STXBP5 can be bound by four different RBPs. STXBP5 has been identified as a novel candidate gene for cardiovascular disease via GWAS (88).

DISCUSSION

We present a comprehensive resource, POSTAR, for easily exploring RBP-target interactions and their putative functions and consequences. POSTAR is the largest collection of RBP binding sites in the human and mouse transcriptomes. We combined a large amount of functional data sources to annotate RBP binding sites. POSTAR has a convenient interface, which provides multiple search modes and enables integrated navigation of RBP binding sites with various post-transcriptional regulatory events, phenotypes, diseases and other factors.

POSTAR enables integrated visualization and exploration of multiple RBPs and post-transcriptional regulatory events. Investigating the relationships between RBP binding sites, regulatory events, phenotypes, and diseases should facilitate the development of novel hypotheses. As mentioned in the STXBP5 example described above, the mechanism underlying the association between genetic variation in STXBP5 and cardiovascular disease is unclear. The search results from POSTAR suggest that binding of several RBPs (e.g. FMR1, FXR2, IGF2BP2 and IGF2BP3) to STXBP5 may be associated with the development of cardiovascular disease.

Establishing the functional roles of genetic variants remains a significant challenge in the post-genomic era. Existing studies have revealed that systematic annotation of *cis*-regulation and the epigenome can reveal many of the functional consequences of a variant (9,89). However, similar efforts regarding post-transcriptional regulation remain limited; this limitation is partly due to the lack of systematic profiling data on post-transcriptional regulation from current genomic studies, such as the ENCODE project (1) and Roadmap Epigenomics project (2). In comparison with methods used in previous studies (10) of functional variants involved in post-transcriptional interaction and regulation, our database has several notable advantages: (i) it includes RBP binding sites and genomic variants in mouse, (ii) it provides local RNA secondary structures around genomic variants, (iii) it enables integrated searching and visualization of RBP binding sites with other genomic variants and regulatory events and (iv) it provides more search and usage modes, such as the 'RBP' search, 'Functional gene' search and 'Predict' server.

Continued accumulation of multiple data resources related to post-transcriptional regulation will enable us to systematically identify post-transcriptional regulatory networks. For example, integrative analysis of RBP binding and miRNA binding data will allow identification of cooperative and competitive combinatorial patterning of these regulatory factors (90,91). POSTAR represents an early step toward achieving these goals. We believe that POSTAR will facilitate the generation of novel hypotheses regarding the biological functions of RBP binding sites through systematic annotation with post-transcriptional regulatory events, trait/disease-associated variants and functional genes. In the future, we will maintain and update POSTAR to ensure that it remains a useful resource for the research community.

AVAILABILITY

POSTAR is freely available at http://POSTAR.ncrnalab.org (redirected to http://lulab.life.tsinghua.edu.cn/POSTAR). The POSTAR data files can be downloaded and used in accordance with the GNU Public License and the license of their primary data sources.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the ENCODE Project Consortium for sharing the eCLIP-seq data publicly.

FUNDING

National Key Research and Development Plan of China [2016YFA0500803]; National High-Tech Research and Development Program of China [2014AA021103]; National Natural Science Foundation of China [31271402, 31522030]; Tsinghua University Initiative Scientific Research Program [2014z21045]; Computing Platform of National Protein Facilities (Tsinghua University). Funding for open access charge: National Key Research and Development Plan of China [2016YFA0500803].
Conflict of interest statement. None declared.

REFERENCES

1. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
2. Roadmap Epigenomics, C., Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
3. Li,S. and Mason,C.E. (2014) The pivotal regulatory landscape of RNA modifications. *Annu. Rev. Genomics Hum. Genet.*, **15**, 127–150.
4. Nishikura,K. (2010) Functions and regulation of RNA editing by ADAR deaminases. *Annu. Rev. Biochem.*, **79**, 321–349.
5. Vandivier,L.E., Campos,R., Kuksa,P.P., Silverman,I.M., Wang,L.S. and Gregory,B.D. (2015) Chemical modifications mark alternatively spliced and uncapped messenger RNAs in Arabidopsis. *Plant Cell*, **27**, 3024–3037.
6. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
7. Cancer Genome Atlas Research, N., Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
8. Ritchie,G.R. and Flicek,P. (2014) Computational approaches to interpreting genomic sequence variation. *Genome Med.*, **6**, 87.
9. Khurana,E., Fu,Y., Chakravarty,D., Demichelis,F., Rubin,M.A. and Gerstein,M. (2016) Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.*, **17**, 93–108.
10. Mao,F., Xiao,L., Li,X., Liang,J., Teng,H., Cai,W. and Sun,Z.S. (2016) RBP-Var: a database of functional variants involved in regulation mediated by RNA-binding proteins. *Nucleic Acids Res.*, **44**, D154–D163.

11. Ye,J. and Blelloch,R. (2014) Regulation of pluripotency by RNA binding proteins. *Cell Stem Cell*, **15**, 271–280.
12. Keller,T.H., Pichota,A. and Yin,Z. (2006) A practical view of 'druggability'. *Curr. Opin. Chem. Biol.*, **10**, 357–361.
13. Glisovic,T., Bachorik,J.L., Yong,J. and Dreyfuss,G. (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.*, **582**, 1977–1986.
14. Filipowicz,W., Bhattacharyya,S.N. and Sonenberg,N. (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.*, **9**, 102–114.
15. Ascano,M., Gerstberger,S. and Tuschl,T. (2013) Multi-disciplinary methods to define RNA-protein interactions and regulatory networks. *Curr. Opin. Genet. Dev.*, **23**, 20–28.
16. 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.
17. Khorshid,M., Rodak,C. and Zavolan,M. (2011) CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res.* **39**, D245–D252.
18. 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.
19. Blin,K., Dieterich,C., Wurmus,R., Rajewsky,N., Landthaler,M. and Akalin,A. (2015) DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.*, **43**, D160–D167.
20. Yang,Y.C., Di,C., Hu,B., Zhou,M., Liu,Y., Song,N., Li,Y., Umetsu,J. and Lu,Z.J. (2015) CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, **16**, 51.
21. Paz,I., Kosti,I., Ares,M. Jr, Cline,M. and Mandel-Gutfreund,Y. (2014) RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.*, **42**, W361–W367.
22. Agostini,F., Zanzoni,A., Klus,P., Marchese,D., Cirillo,D. and Tartaglia,G.G. (2013) catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics*, **29**, 2928–2930.
23. Van Nostrand,E.L., Pratt,G.A., Shishkin,A.A., Gelboin-Burkhart,C., Fang,M.Y., Sundararaman,B., Blue,S.M., Nguyen,T.B., Surka,C., Elkins,K. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.
24. Silverman,I.M., Li,F., Alexander,A., Goff,L., Trapnell,C., Rinn,J.L. and Gregory,B.D. (2014) RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome Biol.*, **15**, R3.
25. Parker,B.J., Moltke,I., Roth,A., Washietl,S., Wen,J., Kellis,M., Breaker,R. and Pedersen,J.S. (2011) New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res.*, **21**, 1929–1943.
26. Sun,W.J., Li,J.H., Liu,S., Wu,J., Zhou,H., Qu,L.H. and Yang,J.H. (2016) RMBase: a resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Res.*, **44**, D259–D265.
27. Dominissini,D., Nachtergaele,S., Moshitch-Moshkovitz,S., Peer,E., Kol,N., Ben-Haim,M.S., Dai,Q., Di Segni,A., Salmon-Divon,M., Clark,W.C. *et al.* (2016) The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA. *Nature*, **530**, 441–446.
28. Ramaswami,G. and Li,J.B. (2014) RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic Acids Res.*, **42**, D109–D113.
29. Kiran,A.M., O'Mahony,J.J., Sanjeev,K. and Baranov,P.V. (2013) Darned in 2013: inclusion of model organisms and linking with Wikipedia. *Nucleic Acids Res.*, **41**, D258–D261.
30. NCBI Resource Coordinators. (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **44**, D7–D19.
31. Li,M.J., Liu,Z., Wang,P., Wong,M.P., Nelson,M.R., Kocher,J.P., Yeager,M., Sham,P.C., Chanock,S.J., Xia,Z. *et al.* (2016) GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **44**, D869–D876.
32. Landrum,M.J., Lee,J.M., Riley,G.R., Jang,W., Rubinstein,W.S., Church,D.M. and Maglott,D.R. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.
33. Lawrence,M.S., Stojanov,P., Mermel,C.H., Robinson,J.T., Garraway,L.A., Golub,T.R., Meyerson,M., Gabriel,S.B., Lander,E.S. and Getz,G. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **505**, 495–501.
34. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Aparicio,S.A., Behjati,S., Biankin,A.V., Bignell,G.R., Bolli,N., Borg,A., Borresen-Dale,A.L. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.
35. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.
36. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
37. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
38. Consortium,G.T. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
39. Liu,X., Yu,X., Zack,D.J., Zhu,H. and Qian,J. (2008) TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.
40. Cavalli,F.M., Bourgon,R., Vaquerizas,J.M. and Luscombe,N.M. (2011) SpeCond: a method to detect condition-specific gene expression. *Genome Biol.*, **12**, R101.
41. Amberger,J.S., Bocchini,C.A., Schiettecatte,F., Scott,A.F. and Hamosh,A. (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.*, **43**, D789–D798.
42. Pinero,J., Queralt-Rosinach,N., Bravo,A., Deu-Pons,J., Bauer-Mehren,A., Baron,M., Sanz,F. and Furlong,L.I. (2015) DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, bav028.
43. Wagner,A.H., Coffman,A.C., Ainscough,B.J., Spies,N.C., Skidmore,Z.L., Campbell,K.M., Krysiak,K., Pan,D., McMichael,J.F., Eldred,J.M. *et al.* (2016) DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res.*, **44**, D1036–D1044.
44. Gerstberger,S., Hafner,M. and Tuschl,T. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.
45. Speir,M.L., Zweig,A.S., Rosenbloom,K.R., Raney,B.J., Paten,B., Nejad,P., Lee,B.T., Learned,K., Karolchik,D., Hinrichs,A.S. *et al.* (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, **44**, D717–D725.
46. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.
47. Kasprzyk,A. (2011) BioMart: driving a paradigm change in biological data management. *Database*, **2011**, bar049.
48. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
49. Gesteland,R.F., Cech,T. and Atkins,J.F. (1999) *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA*. 2nd edn. Cold Spring Harbor Laboratory Press, NY.
50. Uren,P.J., Bahrami-Samani,E., Burns,S.C., Qiao,M., Karginov,F.V., Hodges,E., Hannon,G.J., Sanford,J.R., Penalva,L.O. and Smith,A.D. (2012) Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, **28**, 3013–3020.
51. Corcoran,D.L., Georgiev,S., Mukherjee,N., Gottwein,E., Skalsky,R.L., Keene,J.D. and Ohler,U. (2011) PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biology*, **12**, R79.
52. Moore,M.J., Zhang,C., Gantman,E.C., Mele,A., Darnell,J.C. and Darnell,R.B. (2014) Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat. Protoc.*, **9**, 263–293.
53. Weyn-Vanhentenryck,S.M., Mele,A., Yan,Q., Sun,S., Farny,N., Zhang,Z., Xue,C., Herre,M., Silver,P.A., Zhang,M.Q. *et al.* (2014)

HITS-CLIP and integrative modeling define the Rbfox splicing-regulatory network linked to brain development and autism. *Cell Rep.*, **6**, 1139–1152.
54. Ray,D., Kazan,H., Cook,K.B., Weirauch,M.T., Najafabadi,H.S., Li,X., Gueroussov,S., Albu,M., Zheng,H., Yang,A. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172–177.
55. Cook,K.B., Kazan,H., Zuberi,K., Morris,Q. and Hughes,T.R. (2011) RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.*, **39**, D301–D308.
56. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
57. Schug,J. (2008) Using TESS to predict transcription factor binding sites in DNA sequence. *Curr. Protoc. Bioinformatics*, doi:10.1002/0471250953.bi0206s21.
58. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
59. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2003) MicroRNA targets in Drosophila. *Genome Biol.*, **5**, R1.
60. Kruger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.*, **34**, W451–W454.
61. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
62. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
63. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
64. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
65. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
66. Kazan,H., Ray,D., Chan,E.T., Hughes,T.R. and Morris,Q. (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol.*, **6**, e1000832.
67. Rabani,M., Kertesz,M. and Segal,E. (2008) Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 14885–14890.
68. Deigan,K.E., Li,T.W., Mathews,D.H. and Weeks,K.M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 97–102.
69. Berkowitz,N.D., Silverman,I.M., Childress,D.M., Kazan,H., Wang,L.S. and Gregory,B.D. (2016) A comprehensive database of high-throughput sequencing-based RNA secondary structure probing data (Structure Surfer). *BMC Bioinformatics*, **17**, 215.
70. Wu,Y., Qu,R., Huang,Y., Shi,B., Liu,M., Li,Y. and Lu,Z.J. (2016) RNAex: an RNA secondary structure prediction server enhanced by high-throughput structure-probing data. *Nucleic Acids Res.*, **44**, W294–W301.
71. Lorenz,R., Bernhart,S.H., Honer Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.: AMB*, **6**, 26.
72. Schmieder,R. and Edwards,R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, **27**, 863–864.
73. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
74. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
75. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
76. Katz,Y., Wang,E.T., Airoldi,E.M. and Burge,C.B. (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods*, **7**, 1009–1015.
77. Alexa,A., Rahnenfuhrer,J. and Lengauer,T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.
78. Kerpedjiev,P., Hammer,S. and Hofacker,I.L. (2015) Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, **31**, 3377–3379.
79. Wan,Y., Kertesz,M., Spitale,R.C., Segal,E. and Chang,H.Y. (2011) Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.*, **12**, 641–655.
80. Spitale,R.C., Flynn,R.A., Zhang,Q.C., Crisalli,P., Lee,B., Jung,J.W., Kuchelmeister,H.Y., Batista,P.J., Torre,E.A., Kool,E.T. *et al.* (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, **519**, 486–490.
81. Nichols,K.E., Malkin,D., Garber,J.E., Fraumeni,J.F. Jr and Li,F.P. (2001) Germ-line p53 mutations predispose to a wide spectrum of early-onset cancers. *Cancer Epidemiol. Biomarkers Prev.*, **10**, 83–87.
82. Martin,G., Gruber,A.R., Keller,W. and Zavolan,M. (2012) Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. *Cell Rep.*, **1**, 753–763.
83. Ruepp,M.D., Aringhieri,C., Vivarelli,S., Cardinale,S., Paro,S., Schumperli,D. and Barabino,S.M. (2009) Mammalian pre-mRNA 3′ end processing factor CF I m 68 functions in mRNA export. *Mol. Biol. Cell*, **20**, 5211–5223.
84. Dettwiler,S., Aringhieri,C., Cardinale,S., Keller,W. and Barabino,S.M. (2004) Distinct sequence motifs within the 68-kDa subunit of cleavage factor Im mediate RNA binding, protein-protein interactions, and subcellular localization. *J. Biol. Chem.*, **279**, 35788–35797.
85. Hinman,M.N. and Lou,H. (2008) Diverse molecular functions of Hu proteins. *Cell. Mol. Life Sci.: CMLS*, **65**, 3168–3181.
86. Lebedeva,S., Jens,M., Theil,K., Schwanhausser,B., Selbach,M., Landthaler,M. and Rajewsky,N. (2011) Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol. Cell*, **43**, 340–352.
87. Mukherjee,N., Corcoran,D.L., Nusbaum,J.D., Reid,D.W., Georgiev,S., Hafner,M., Ascano,M. Jr, Tuschl,T., Ohler,U. and Keene,J.D. (2011) Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell*, **43**, 327–339.
88. van Loon,J.E., Leebeek,F.W., Deckers,J.W., Dippel,D.W., Poldermans,D., Strachan,D.P., Tang,W., O'Donnell,C.J., Smith,N.L. and de Maat,M.P. (2010) Effect of genetic variations in syntaxin-binding protein-5 and syntaxin-2 on von Willebrand factor concentration and cardiovascular risk. *Circ. Cardiovasc. Genet.*, **3**, 507–512.
89. Ward,L.D. and Kellis,M. (2012) Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.*, **30**, 1095–1106.
90. Jens,M. and Rajewsky,N. (2015) Competition between target sites of regulators shapes post-transcriptional gene regulation. *Nat. Rev. Genet.*, **16**, 113–126.
91. HafezQorani,S., Lafzi,A., de Bruin,R.G., van Zonneveld,A.J., van der Veer,E.P., Son,Y.A. and Kazan,H. (2016) Modeling the combined effect of RNA-binding proteins and microRNAs in post-transcriptional regulation. *Nucleic Acids Res.*, **44**, e83.