D322–D329 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 24 November 2018*
doi: 10.1093/nar/gky1112

ENPD - A Database of Eukaryotic Nucleic Acid Binding Proteins: Linking Gene Regulations to Proteins

Ricky Wai Tak Leung[1], Xiaosen Jiang[2,3,4], Ka Hou Chu[1,2] and Jing Qin[1,2,5,*]

[1]School of Life Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China, [2]Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen 518057, China, [3]School of Future Technology, The University of Chinese Academy of Sciences, Beijing 100049, China, [4]College of Life Science & Technology, Huazhong University of Science and Technology, Wuhan 430074, China and [5]School of Pharmaceutical Sciences (Shenzhen), Sun Yat-sen University, Guangzhou 510275, China

Received August 15, 2018; Revised October 09, 2018; Editorial Decision October 20, 2018; Accepted October 23, 2018

ABSTRACT

Eukaryotic nucleic acid binding protein database (ENPD, http://qinlab.sls.cuhk.edu.hk/ENPD/) is a library of nucleic acid binding proteins (NBPs) and their functional information. NBPs such as DNA binding proteins (DBPs), RNA binding proteins (RBPs), and DNA and RNA binding proteins (DRBPs) are involved in every stage of gene regulation through their interactions with DNA and RNA. Due to the importance of NBPs, the database was constructed based on manual curation and a newly developed pipeline utilizing both sequenced transcriptomes and genomes. In total the database has recorded 2.8 million of NBPs and their binding motifs from 662 NBP families and 2423 species, constituting the largest NBP database. ENPD covers evolutionarily important lineages which have never been included in the previous NBP databases, while lineage-specific NBP family expansions were also found. ENPD also focuses on the involvements of DBPs, RBPs and DRBPs in non-coding RNA (ncRNA) mediated gene regulation. The predicted and experimentally validated targets of NBPs have both been recorded and manually curated in ENPD, linking the interactions between ncRNAs, DNA regulatory elements and NBPs in gene regulation. This database provides key resources for the scientific community, laying a solid foundation for future gene regulatory studies from both functional and evolutionary perspectives.

INTRODUCTION

Nucleic acid binding proteins and gene regulations

Gene regulations are vital processes in all organisms, enabling an organism to respond to its surroundings and promote its overall complexity. This can be achieved with the help of nucleic acid binding proteins (NBPs) such as RNA binding proteins (RBPs) and DNA binding proteins (DBPs). DBPs possess the ability to bind DNAs, including transcription factors (TFs), chromatin modifiers, histones, etc. As the major group of DBPs, TFs are present in all organisms and are capable of regulating gene expression (1). In order for TFs to perform their function, they would generally bind to a regulatory DNA element, such as a promoter or an enhancer, which is essential for gene transcriptional regulation. A TF acquires such autonomous DNA binding ability through possessing at least one DNA binding domain (DBD) (2). As DNA binding domains are responsible for directing the proteins to a DNA target sites, which thus are grouped into different families based on their similarities and homologies of the DNA binding domains (3). In general, such DNA bound transcription factor can, in turn, interact with RNA polymerase or other protein factors (transcription coregulators), resulting in up- or down-regulation of the target gene (4). With TFs' utmost importance in gene regulation and numerous vital biological processes including cell development, cell growth, and differentiation, many resources are fed into the field to continue revealing everything about them.

Just as DBPs would generally possess the ability to bind to a DNA molecule, RBPs are characterized as proteins with RNA-binding ability. They acquire this ability from having RNA binding domains (RBDs). RBPs have diverse

*To whom correspondence should be addressed. Tel: +852 3943 6391; Fax: +852 2603 5391; Email: qinj29@mail.sysu.edu.cn
Present address: School of Life Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, China.

functions and they are involved in almost all central cellular processes as important regulators of gene expression (5). They regulate gene expression by taking part in non-coding RNA (ncRNA) mediated gene regulation as well as every stage of mRNA processing including splicing, polyadenylation, capping, modification, export, localization and translation of mRNAs (6–8). Although DBPs and RBPs have been studied separately as two distinct classes of proteins for decades, there are growing shreds of evidence in recent years showing that the two groups are closely related, with the designation of a new protein subclass - the DNA and RNA binding proteins (DRBPs) which can bind to both DNA and RNA and are evolutionarily close to both DBPs and RBPs (9). Because of their duel nucleic acid binding ability, they could perform functions that DBP and RBP do but many more such as gene silencing (10), microRNA biogenesis (11) and telomere maintenance (12), making them an equally if not more important protein class than DBPs and RBPs. As DRBPs, DBPs and RBPs are evolutionarily and functionally closely related, we hereby consider them as a single protein group, NBPs.

Current NBP databases

Due to the importance of DBPs and RBPs, their repertoires in many organisms have been recorded in various published databases, most of which rely mainly on sequenced genomes. With the emergence and advancement of next generation sequencing, sequencing a genome has become faster and less costly, so that hundreds of thousands of DBPs or RBPs can also be screened and identified across the tree of life in a cost-effective approach. For example, AnimalTFDB is a comprehensive database of animal transcription factors, recording TFs, transcription co-factors and chromatin modifiers from 97 animal species (13); DBD is a database that contains predicted DBP information from 972 species with completely sequenced genomes across Archaea, Bacteria, and Eukaryota (14); CIS-BP—the Catalog of Inferred Sequence Binding Preferences (15) is an online library of predicted DBPs and their inferred DNA binding motifs from 339 eukaryotic species; while CISBP-RNA (16) and ATtRACT (17) are online libraries of RBPs and their motifs from 289 and 38 eukaryotic species respectively. These databases have served the community very well, as they often provide basic information for experimental design and act as a foundation for many gene regulatory researches. Yet there are two common inadequacies. First, these databases are confined to cover only one type of NBPs; there are neither databases for all three types of NBPs nor any databases documenting DRBPs. Secondly, the number of species covered in these databases is restricted by the number of sequenced genomes and low-throughput experiments. Due to the difficulties in sequencing some eukaryotic genomes, a relatively small number of eukaryotes had their TF repertoires been recorded, ranging from 38 in ATtRACT to 339 eukaryotic species in CIS-BP (Supplementary Table S1).

Fortunately, due to the ease of sequencing transcriptomes, the Transcriptome Shotgun Assembly (TSA) database contains more than 1.7k eukaryotic species. Most of these species' NBP repertoires were not recorded before, such as for Alismatales (basal monocots), early-diverging eudicots (primitive dicots), Palaeoptera (an ancestral group of winged insects), Bivalvia (molluscs with shells divided into two halves) and Cephalopoda (the lineage containing squid and octopus), providing us the opportunity to fill the knowledge gaps of NBP genes in such lineages and increases the taxon coverage in lineages with sequenced genomes (Supplementary Figure S1). To tackle the inadequacies in previous databases, we have constructed the Eukaryotic Nucleic acid binding Protein Database (ENPD, http://qinlab.sls.cuhk.edu.hk/ENPD/) based on manual curation and a newly developed pipeline utilizing both sequenced transcriptomes and genomes, aiming to document all three types of NBPs and provide gene regulatory information linking ncRNAs, DNA regulatory elements and NBPs interactions.

MATERIALS AND METHODS

Database overview

ENPD is a database of eukaryotic NBPs derived from genomes, transcriptomes and published literature used for manual curation (Figure 1). We explored a comprehensive collection of eukaryotic genomes and transcriptomes to identify huge number of NBPs. The database allows users to search NBPs with keywords matching NBP names or NBP identifiers, to select species and NBP family of interest. Free batch download is available for NBP protein sequences and domain sequences of selected species or NBP families. Besides NBP sequences, ENPD also contains functional and regulatory information of NBPs including NBPs' binding motifs with their validated and predicted targets.

Database implementation

ENPD implements a Linux-Apache-MySQL-PHP (LAMP) system. Data were saved in MySQL database, including NBP sequences, NBP information, domains, species information, and motifs. The web is constructed based on the powerful PHP framework CodeIgniter, which provides an Application Programming Interface (API) to connect the web to MySQL database. We also used JavaScript libraries including jQuery (2.2.0), jQuery-labelauty and some additional plugins to perform dynamic web services.

Data resources

Assembled eukaryotic transcriptomes were downloaded from National Center for Biotechnology Information (NCBI) TSA database (https://www.ncbi.nlm.nih.gov/genbank/tsa/). The detailed information of all transcriptomes from TSA are summarized in Supplementary Table S2. As of August 2018, our transcriptome collection composes of 2560 eukaryotic transcriptome samples from 1770 species. Proteins predicted from eukaryotic genomes were downloaded from NCBI genome database (ftp://ftp.ncbi.nlm.nih.gov/genomes/). The detailed information of all NCBI genomes are summarized in Supplementary Table S3. As of August 2018, our genome protein collection consists of 790 eukaryotic species.

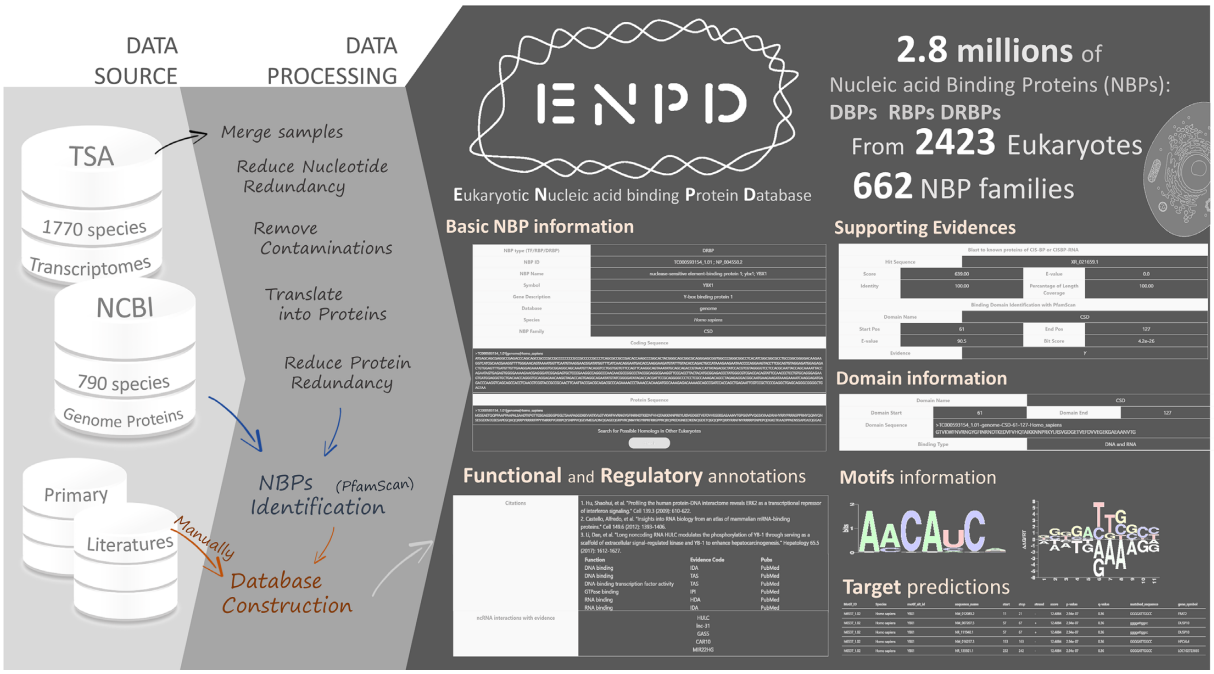

Figure 1. (**A**) Overview of ENPD including data source, data processing and features. Transcriptomes from TSA, genomes from NCBI and primary literature were used as data sources for the construction of eukaryotic nucleic acid binding protein database (ENPD). Recording the 3 types of NBPs in eukaryotes: DBPs, RBPs and DRBPs. The basic information of NBPs as well as their functional and regulatory information are also curated by both prediction and manual. See text for details.

Information from primary literature concerning proteins with nucleic acid binding validations from low throughput experiments were also compiled manually and included in the database. Combining all data sources, there is 2423 species in total recorded in the database.

Data processing

TSA assembled transcriptomes were directly downloaded from the database, contigs from different transcriptomes originated from the same species were merged. The contig redundancies were reduced by cd-hit-est (18) with a sequence identity threshold of 0.98. Potentially contaminated sequences from bacteria, virus or archaea were filtered by Kraken (19). Coding sequences (CDSs) and amino acid sequences were deduced from assembled contigs with Transdecoder (20). The redundancies of the proteins predicted from transcriptomes and the downloaded proteins originated from NCBI genomes were reduced by cd-hit with a sequence identity threshold of 0.95.

Functional and regulatory annotations of NBPs

DBDs and RBDs were identified by scanning all proteins with hidden Makov models (HMMs) of known DBDs and RBDs from Pfam database by PfamScan (21) that implements HMMER3 with default setting (22). In brief, NBPs which contain only DBDs were classified as DBPs; NBPs with only RBDs detected were classified as RBPs; while NBPs with both binding domains or domains with both DNA and RNA binding abilities detected were listed as DRBPs. Proteins with direct evidence from primary literature that can bind with both DNA and RNA were also manually curated as DRBPs. Furthermore, to infer the DNA or RNA binding preference of eukaryotic NBPs, their nucleic acid binding domains (DBDs or RBDs) were compared with those with known DNA or RNA binding motifs, and the domain similarity between eukaryotic NBPs and known NBPs was calculated. Based on the observation on co-evolution of nucleic acid binding domain sequences and their nucleic acid binding motifs, it has been reported that the nucleic acid binding motifs of an NBP could be inferred from those of homologous nucleic acid binding domains when the identity between the two nucleic acid binding domains is greater than a threshold (15). The thresholds of all nucleic acid binding domain families could be found in both CIS-BP and CISBP-RNA (Supplementary Table S4). Known DNA and RNA binding motifs were downloaded from CIS-BP (15), CISBP-RNA (16), JASPAR (23), UniPROBE (24) and hPDI (25). Nucleic acid binding motifs of eukaryotic NBPs could be inferred when the nucleic acid binding domains were highly similar to those with nucleic acid binding motifs that were detected experimentally. The predicted binding motifs of NBPs were used to predict NBP targets of six model organisms namely *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Gallus gallus* and *Danio retio* by scanning gene promoter DNA sequences (defined as 1000 bp upstream of transcription start site) from UCSC genome browser (https://genome.ucsc.edu/cgi-bin/hgTables) and ncRNA targets from LNCipedia (https://lncipedia.org/) (26) or NONCODE (www.noncode.org/) (27) by MEME Suite (28). Validated gene targets were manually curated from published literature. The protein func-

tionality of NBPs recorded in literature was also curated (Figure 1, Supplementary Figure S9).

Confidence quantification of NBPs

The confidence level of a predicted NBP was estimated using several criteria: (i) the percentage of the top hit that matches known NBPs, E-value, Blast score and sequence identity when Blast to the NBPs in DBP and RBP databases CIS-BP and CISBP-RNA, respectively; (ii) the bit-score and E-value of PfamScan that identifies the DBDs and RBDs and (iii) whether or not there are direct experimental evidence documenting the DNA and RNA binding ability of the NBP.

RESULTS AND DISCUSSION

The largest NBP database

ENPD utilizes all low-throughput experiments, transcriptomes and genomes as its data sources, in comparison to previous databases which only utilize data from low-throughput experiments or genomes, leading to a substantial explosion in NBP repertoire (Supplementary Figure S2). ENDP contains NBP information from a massive 2423 eukaryotic species covering multiple lineages in four kingdoms and >20 phyla (Supplementary Figure S3), its species coverage is seven times more than that of CIS-BP/CISBP-RNA, and 2.8 million NBPs, outnumbering CIS-BP and CISBP-RNA combined by 12-fold. Similar to CIS-BP and CISBP-RNA, NBPs are classified into different families according to their nucleic acid binding domain combinations identified in their sequences. ENDP records a total number of 662 NBP families with 344 families unannotated before (Supplementary Figure S2). In short, the vast amount of data gives rise to ENPD, the largest NBP database available so far that includes all three classes of NBPs, i.e., DBPs, RBPs and DRBPs.

Transcriptomes as a great source of data

ENPD includes several important lineages of organisms that were not covered by the two largest NBPs databases, DBD and CIS-BP, as there are only transcriptomes but no complete genome data from species of these lineages (Supplementary Figure S1). By constructing the NBP repertoires of these lineages, insights and clues concerning the evolution of NBPs in monocots, dicots, insects and molluscs can be acquired. Therefore, when we utilized transcriptomes as an additional source of data, ENPD can provide NBP repertoires for 1633 more species, on top of the 790 species with sequenced genomes. Moreover, out of the 2.8 million recorded NBPs, more than half of them (1.5 million) were predicted from transcriptomes. Furthermore, the depth and quality of the NBP repertoires based on transcriptome data are comparable to those originated from genomes. Genome and transcriptome-based repertoires are both often found within individual lineages, with the NBP family member abundance pattern highly similar and indistinguishable (Supplementary Figure S4). Although the amount of NBPs that can be retrieved from transcriptomes is affected by sequencing depth and constrained by the tissues being extracted, transcriptome mined NBPs have a higher confidence that the proteins were truly expressed while non-expressed pseudogenes may be present in genomes. The use of transcriptomes as an additional data source tremendously increases the species and lineage coverage with comparable quality of NBP prediction as that based on genome data.

Lineage-specific NBP family expansion pattern in eukaryotes

Kingdom-specific expansion of NBP family. The NBP family RNA recognition motif (RRM) has vast diverse functions including but not limited to RNA alternative splicing, RNA stability and translation. It is the only NBP family found to be extensively abundant in all eukaryotes. Certain NBP families are found to show kingdom-specific expansion across eukaryotes (Supplementary Figure S4). The NBP families C2H2-ZF and Forkhead are found to be abundant in Fungi and Metazoa. There are ten NBP families, namely mTERF, AP2, NAC/NAM, WRKY, B3, FAR1, TCP, RWP-RK, SBP and Dof, found to be expanded specifically in plants. These NBP families are highly related to plant specific functions, explaining their high abundance in plants. The NBP family nuclear receptor (zf-C4) and ETS are found to be expanded extensively and specifically in metazoans. For more details please refer to Supplementary file 1.

Nuclear receptors (zf-C4) in Metazoa. Nuclear receptors (zf-C4) are responsible for detecting steroid and thyroid hormones and have been found in some of the earliest branching animal lineages such as sponge, sea anemone, *Trichoplax* and comb jelly (29,30). In ENDP, nuclear receptors (zf-C4) are found to be rather abundant in some lineages such as Eutardigrada, Rotifera and the basal bilaterian Acoela, and very abundant in *Caenorhabditis*, a genus of nematodes living in bacteria-rich environment. On the other hand, this NBP family is found to be much less abundant in two groups of soft bodied creatures: Cnidaria and Cephalopoda (Supplementary Figure S5).

Lineage-specific pattern in Viridiplantae. Within kingdom Viridiplantae, clade Chlorophyta (aquatic green algae) and Streptophyta (mostly land plants) possess very distinct NBP family abundance patterns (Supplementary Figure S6), perhaps due to their contrasting habitats. Families S1 and SAP domain are highly abundant in Chlorophyta, while families NAC/NAM, B3 and WRKY are highly abundant in Streptophyta.

Lineage-specific pattern in Protoctista. The now invalid kingdom Protoctista is a convenient grouping of eukaryotes that are not members of Metazoa, Fungi or Viridiplantae. This group is truly highly diverse, accounting for the distinctive patterns as shown in the NBP family protein abundance among different clades. Clades Trypanosomatidae, Florideophyceae, Dinophyceae, Apicomplexa and Phaeophyceae each has a highly unified abundance pattern among members within the same clade but exhibits extremely different patterns among different clades (Supplementary Figure S7).

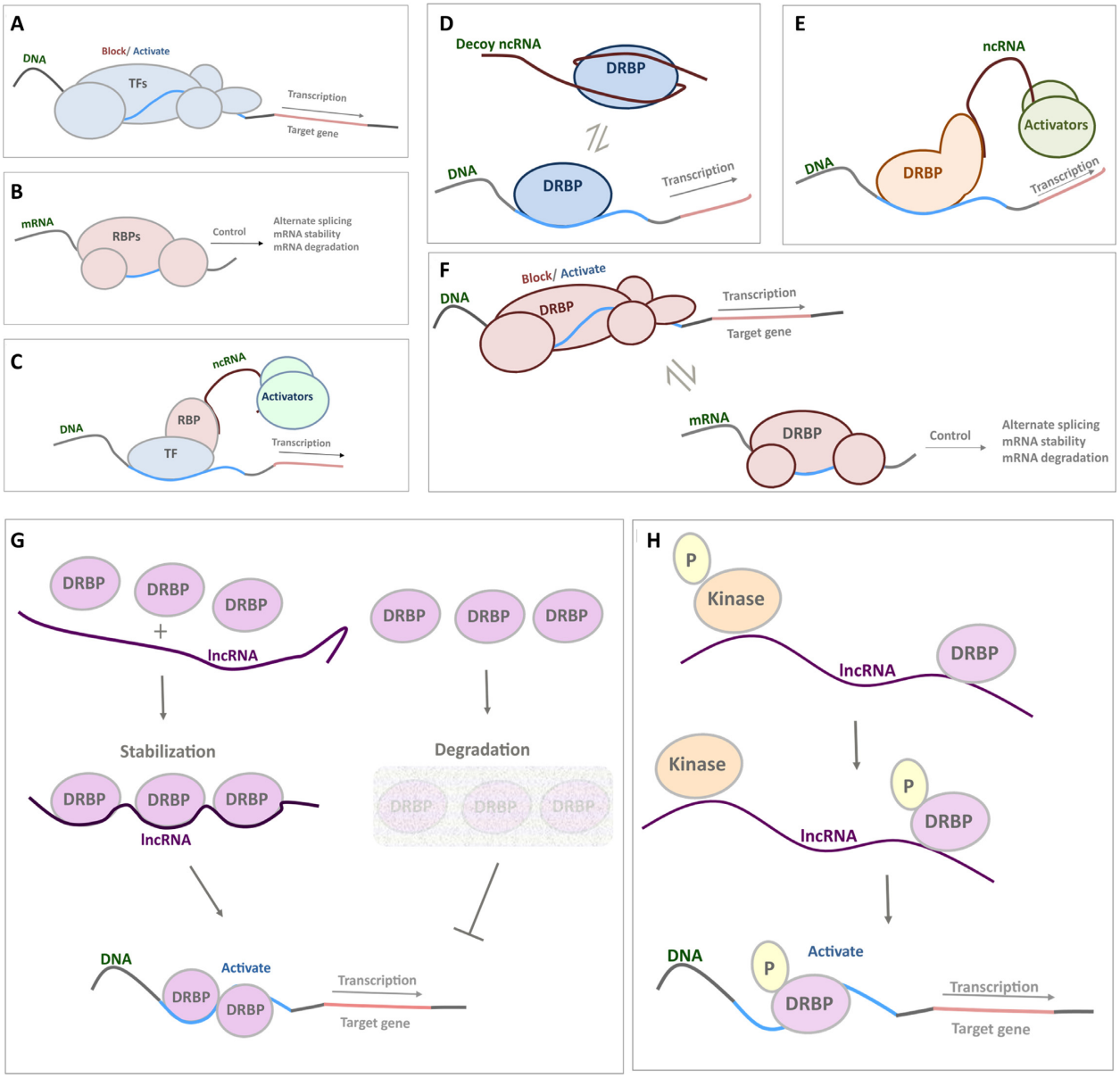

Figure 2. Functions of TFs, RBPs and DRBPs in ENPD. NBPs are involved in every stage of gene regulation and are important in most if not all biological processes through interactions with DNAs and RNAs, controlling cell identity and pathological status. (**A**) TFs usually bind to DNA regulatory elements of a gene and regulate its transcription activity, while (**B**) RBPs have diverse functions related to mRNA regulation, including alternative splicing, mRNA stability control, and degradation. (**C**) The two types of proteins can cooperate to recruit ncRNA for subsequent transcriptional activator recruitment. DRBPs have a wide range of distinct functions: (**D**) functioning as a typical TF with add-on abilities such as interacting with ncRNA decoys; (**E**) recruiting ncRNA with co-activator function; (**F**) regulating mRNAs and DNA regulatory elements just as typical RBPs and TFs, with the two functions competing for the same protein; (**G**) interacting with lncRNA for stability enhancement, thus increasing DRBP abundance and further activating downstream gene transcriptions and (**H**) altering the phosphorylation status by the mediation of lncRNA, enhancing the gene activating activity.

Features and examples in ENPD

ENPD is an online library of NBPs and their functional information. NBPs, i.e., DBPs, RBPs as well as DRBPs can be searched, browsed and downloaded with the user-friendly interface (Supplementary Figures S8–S10). Users are able to search their interested NBPs through several filters including data origin (from genomes or transcriptomes), species origin, binding domain identity and nucleic acid binding preference. Users can also recover their interested NBPs through the keyword search box function by directly inputting the gene name (Supplementary Figure S8). The resulting NBP entry pages contain abundant useful information, such as NBP name, NBP ID, species origin, NBP type, domain information, binding preferences of each domain, CDS and protein sequences. The NBP entry page also provides a supporting proof section where user can find best hit Blast result scores of this protein blasted towards NBPs in CIS-BP and CISBP-RNA as well as the PfamScan identification scores. Users can then be able to estimate the reliability of the protein under view base on

the information provided. Information from all of the NBPs recorded in ENPD are also available to download for free; users are able to select their target NBPs by three filters namely species, binding domain family and data source (from genome or transcriptome) and download the NBPs' CDSs, protein sequences and domain sequences (Supplementary Figure S10).

ENPD also records functional information for NBPs (Supplementary Figure S9). TFs usually bind to DNA regulatory elements of a gene, and up-regulate or down-regulate its transcription activity (4) (Figure 2A). While RBPs have diverse functions related to mRNA regulation, including alternative splicing, mRNA stability control, and degradation (6–8) (Figure 2B). For example, the mouse RBP KSRP was found to interact with long non-coding RNA (lncRNA) H19, and this interaction further favours the binding of the protein to myogenin mRNA, and induces degradation of the mRNA, thus mediating the maintenance of undifferentiated cell state (31). In some occasions, TF and RBP can cooperate to recruit lncRNA for further activators recruitment (32) (Figure 2C). Publications supporting these functions are thus also annotated in the web page of each DBP or RBP.

ENPD is the first database curating the third kind of NBP, DRBPs. Some DRBPs are just typical DBPs with add-on abilities such as interacting with ncRNA decoys. Taking glucocorticoid receptor (GR) as an example, it is a TF that can bind with glucocorticoid and up-regulates the expression of anti-inflammatory genes through activating the glucocorticoid response element on DNA. It is also capable of binding with an RNA decoy of a glucocorticoid response element called growth arrest-specific 5 (Gas5), a ncRNA present in the cell. The interaction between GR and Gas5 prevents GR from activating the glucocorticoid response element, in turn down-regulating anti-inflammatory proteins (33) (Figure 2D). Functions of other DRBPs in our database include recruitment of ncRNA with co-activator function (34,35) (Figure 2E); carrying out dual functions by regulating mRNAs and DNA regulatory elements just as typical RBPs and DBPs (36–39) (Figure 2F); and interacting with lncRNA for stability enhancement, thus increasing their own abundance and further activating downstream genes (40) (Figure 2G). DRBPs can also alter their phosphorylation status by the mediation of lncRNA, enhancing their gene activating activity (41) (Figure 2H). Manual curation of these experimental evidences for DRBPs in ENPD provides users useful information for understanding their functional roles in gene regulations.

Furthermore, ENPD focuses on the involvement of DBPs, RBPs, and DRBPs in ncRNA mediated gene regulation. For example, the lncRNA HULC was found to interact with the DRBP ybx1. HULC serves as a scaffold between ybx1 and kinases, and in turn promotes the phosphorylation of ybx1. Other than HULC, lncRNAs lnc-31, GAS-5 and CAR10 were also experimentally validated to interact with ybx1 in different pathways, serving different functions (40–44). Apart from validated targets, numerous lncRNA targets were also predicted *in silico* and documented in ENPD (Supplementary Figure S9E). NBPs, their previously annotated molecular function and the predicted and experimentally validated targets have all been recorded and manually curated in ENPD (Figure 1 and S9), linking the interactions between ncRNAs, DNA regulatory elements and NBPs in gene regulation.

CONCLUSION

ENPD is an online library of NBPs and their functional information. It is constructed based on manual curation and a newly developed pipeline utilizing both sequenced transcriptomes and genomes. From transcriptomes and genomes available from 1770 and 790 eukaryotic species respectively, 662 NBP families with more than 2.8 million of NBPs and their binding motifs were predicted for a total number of 2423 species, constituting the largest NBP database that includes DBPs, RBPs, and DRBPs. As the real pictures of gene regulation networks are complicated, and requires the involvement of DNA regulatory elements, different types of RNAs and protein factors, DRBPs often serve as the key and the bridge of them all, linking different parts of the system, this makes the curation of DRBP information in ENPD as a crucial function to the field. This database also covers evolutionarily important lineages such as Alismatales, early-diverging eudicots, Palaeoptera, Bivalvia and Cephalopoda, none of which are included in the previously available databases. To conclude, ENPD can fill the knowledge gaps of NBPs in certain lineages; the information provided in the database can serve as a starting point for future experimental designs and serve as the key resources for the scientific community, providing insights, fundamental data and information for future gene regulatory studies from both functional and evolutionary perspectives.

DATA AVAILABILITY

Eukaryotic Nucleic acid binding Protein Database (ENPD) is free and publicly available at http://qinlab.sls.cuhk.edu.hk/ENPD/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Ka Yan Ma and Miss Fiona Ka Man Cheung for their constructive suggestions and technical supports for this study.

FUNDING

National Natural Science Foundation of China [41606143]; Direct Grants from The Chinese University of Hong Kong [4053187 and 4053260]. Funding for open access charge: National Natural Science Foundation of China.
Conflict of interest statement. None declared.

REFERENCES

1. Latchman,D.S. (1997) Transcription factors: an overview. *Int. J. Biochem. Cell Biol.*, **29**, 1305–1312.

2. Garvie,C.W. and Wolberger,C. (2001) Recognition of specific DNA sequences. *Mol. Cell*, **8**, 937–946.
3. Marmorstein,R. and Fitzgerald,M.X. (2003) Modulation of DNA-binding domains for sequence-specific DNA recognition. *Gene*, **304**, 1–12.
4. Gill,G. (2001) Regulation of the initiation of eukaryotic transcription. *Essays Biochem.*, **37**, 33–43.
5. Re,A., Joshi,T., Kulberkyte,E., Morris,Q. and Workman,C.T. (2014) RNA-protein interactions: an overview. *Methods Mol Biol.*, **1097**, 491–521.
6. Glisovic,T., Bachorik,J.L., Yong,J. and Dreyfuss,G. (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.*, **582**, 1977–1986.
7. Keene,J.D. (2007) RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.*, **8**, 533–543.
8. 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.
9. Hudson,W.H. and Ortlund,E.A. (2014) The structure, function and evolution of proteins that bind DNA and RNA. *Nat. Rev. Mol. Cell Biol.*, **15**, 749–760.
10. Di Ruscio,A., Ebralidze,A.K., Benoukraf,T., Amabile,G., Goff,L.A., Terragni,J., Figueroa,M.E., De Figueiredo Pontes,L.L., Alberich-Jorda,M., Zhang,P. *et al.* (2013) DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature*, **503**, 371–376.
11. Kawahara,Y. and Mieda-Sato,A. (2012) TDP-43 promotes microRNA biogenesis as a component of the Drosha and Dicer complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 3347–3352.
12. Deng,Z., Norseen,J., Wiedmer,A., Riethman,H. and Lieberman,P.M. (2009) TERRA RNA binding to TRF2 facilitates heterochromatin formation and ORC recruitment at telomeres. *Mol. Cell*, **35**, 403–413.
13. Hu,H., Miao,Y.R., Jia,L.H., Yu,Q.Y., Zhang,Q. and Guo,A.Y. (2018) AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.*, doi:10.1093/nar/gky822.
14. Wilson,D., Charoensawan,V., Kummerfeld,S.K. and Teichmann,S.A. (2008) DBD–taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.*, **36**, D88–D92.
15. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
16. 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.
17. Giudice,G., Sanchez-Cabo,F., Torroja,C. and Lara-Pezzi,E. (2016) ATtRACT—a database of RNA-binding proteins and associated motifs. *Database (Oxford)*, **2016**, baw035.
18. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
19. Wood,D.E. and Salzberg,S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, **15**, R46.
20. Haas,B.J., Papanicolaou,A., Yassour,M., Grabherr,M., Blood,P.D., Bowden,J., Couger,M.B., Eccles,D., Li,B., Lieber,M. *et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protoc*, **8**, 1494–1512.
21. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
22. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
23. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Cheneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D1284.
24. Hume,M.A., Barrera,L.A., Gisselbrecht,S.S. and Bulyk,M.L. (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.*, **43**, D117–D122.
25. Xie,Z., Hu,S., Blackshaw,S., Zhu,H. and Qian,J. (2010) hPDI: a database of experimental human protein-DNA interactions. *Bioinformatics*, **26**, 287–289.
26. Volders,P.J., Verheggen,K., Menschaert,G., Vandepoele,K., Martens,L., Vandesompele,J. and Mestdagh,P. (2015) An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.*, **43**, 4363–4364.
27. Zhao,Y., Li,H., Fang,S., Kang,Y., Wu,W., Hao,Y., Li,Z., Bu,D., Sun,N., Zhang,M.Q. *et al.* (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.*, **44**, D203–D208.
28. Bailey,T.L., Johnson,J., Grant,C.E. and Noble,W.S. (2015) The MEME Suite. *Nucleic Acids Res.*, **43**, W39–W49.
29. Reitzel,A.M., Pang,K., Ryan,J.F., Mullikin,J.C., Martindale,M.Q., Baxevanis,A.D. and Tarrant,A.M. (2011) Nuclear receptors from the ctenophore *Mnemiopsis leidyi* lack a zinc-finger DNA-binding domain: lineage-specific loss or ancestral condition in the emergence of the nuclear receptor superfamily? *EvoDevo*, **2**, 3.
30. Bridgham,J.T., Eick,G.N., Larroux,C., Deshpande,K., Harms,M.J., Gauthier,M.E., Ortlund,E.A., Degnan,B.M. and Thornton,J.W. (2010) Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS Biol.*, **8**, E100497.
31. Giovarelli,M., Bucci,G., Ramos,A., Bordo,D., Wilusz,C.J., Chen,C.Y., Puppo,M., Briata,P. and Gherzi,R. (2014) H19 long noncoding RNA controls the mRNA decay promoting function of KSRP. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E5023–E5028.
32. Marchese,F.P., Raimondi,I. and Huarte,M. (2017) The multidimensional mechanisms of long noncoding RNA function. *Genome Biol.*, **18**, 206.
33. Kino,T., Hurt,D.E., Ichijo,T., Nader,N. and Chrousos,G.P. (2010) Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci. Signal*, **3**, ra8.
34. Xu,B., Yang,W.H., Gerin,I., Hu,C.D., Hammer,G.D. and Koenig,R.J. (2009) Dax-1 and steroid receptor RNA activator (SRA) function as transcriptional coactivators for steroidogenic factor 1 in steroidogenesis. *Methods Cell Biol.*, **29**, 1719–1734.
35. Poon,M.M. and Chen,L. (2008) Retinoic acid-gated sequence-specific translational control by RARalpha. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 20303–20308.
36. Auphan,N., DiDonato,J.A., Rosette,C., Helmberg,A. and Karin,M. (1995) Immunosuppression by glucocorticoids: inhibition of NF-kappa B activity through induction of I kappa B synthesis. *Science*, **270**, 286–290.
37. Surjit,M., Ganti,K.P., Mukherji,A., Ye,T., Hua,G., Metzger,D., Li,M. and Chambon,P. (2011) Widespread negative response elements mediate direct repression by agonist-liganded glucocorticoid receptor. *Cell*, **145**, 224–241.
38. Jonat,C., Rahmsdorf,H.J., Park,K.K., Cato,A.C., Gebel,S., Ponta,H. and Herrlich,P. (1990) Antitumor promotion and antiinflammation: down-modulation of AP-1 (Fos/Jun) activity by glucocorticoid hormone. *Cell*, **62**, 1189–1204.
39. Dhawan,L., Liu,B., Blaxall,B.C. and Taubman,M.B. (2007) A novel role for the glucocorticoid receptor in the regulation of monocyte chemoattractant protein-1 mRNA stability. *J. Biol. Chem.*, **282**, 10146–10152.
40. Wei,M.M., Zhou,Y.C., Wen,Z.S., Zhou,B., Huang,Y.C., Wang,G.Z., Zhao,X.C., Pan,H.L., Qu,L.W., Zhang,J. *et al.* (2016) Long non-coding RNA stabilizes the Y-box-binding protein 1 and regulates the epidermal growth factor receptor to promote lung carcinogenesis. *Oncotarget*, **7**, 59556–59571.
41. Li,D., Liu,X., Zhou,J., Hu,J., Zhang,D., Liu,J., Qiao,Y. and Zhan,Q. (2017) Long noncoding RNA HULC modulates the phosphorylation of YB-1 through serving as a scaffold of extracellular signal-regulated kinase and YB-1 to enhance hepatocarcinogenesis. *Hepatology*, **65**, 1612–1627.
42. Su,W., Feng,S., Chen,X., Yang,X., Mao,R., Guo,C., Wang,Z., Thomas,D.G., Lin,J., Reddy,R.M. *et al.* (2018) Silencing of long noncoding RNA MIR22HG triggers cell Survival/Death signaling via oncogenes YBX1, MET, and p21 in lung cancer. *Cancer Res.*, **78**, 3207–3219.

43. Liu,Y., Zhao,J., Zhang,W., Gan,J., Hu,C., Huang,G. and Zhang,Y. (2015) lncRNA GAS5 enhances G1 cell cycle arrest via binding to YBX1 to regulate p21 expression in stomach cancer. *Sci. Rep.*, **5**, 10159.

44. Dimartino,D., Colantoni,A., Ballarino,M., Martone,J., Mariani,D., Danner,J., Bruckmann,A., Meister,G., Morlando,M. and Bozzoni,I. (2018) The long non-coding RNA lnc-31 interacts with Rock1 mRNA and mediates its YB-1-dependent translation. *Cell Rep.*, **23**, 733–740.