Published in final edited form as:

Proteins. 2014 April ; 82(4): 640–647. doi:10.1002/prot.24441.

Prediction and validation of the unexplored RNA-binding protein atlas of the human proteome

Huiying Zhao[1,2], **Yuedong Yang**[1,2,3], **Sarath Chandra Janga**[1], **C. Cheng Kao**[4], and **Yaoqi Zhou**[1,3,*]

[1]School of Informatics, Indiana University Purdue University, Indianapolis, Indiana, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 719 Indiana Ave Ste 319, Walker Plaza Building, Indianapolis, Indiana 46202, USA

[3]Institute for Glycomics and School of Informatics and Communication Technology, Griffith University, Parklands Dr., Southport, QLD4215, Australia

[4]Department of Molecular & Cellular Biochemistry, Indiana University, Bloomington, Indiana, 47405, USA

Abstract

Detecting protein-RNA interactions is challenging both experimentally and computationally because RNAs are large in number, diverse in cellular location and function, and flexible in structure. As a result, many RNA-binding proteins (RBPs) remain to be identified. Here, a template-based, function-prediction technique SPOT-Seq for RBPs is applied to human proteome and its result is validated by a recent proteomic experimental discovery of 860 mRNA binding proteins (mRBPs). The coverage (or sensitivity) is 42.6% for 1,217 known RBPs annotated in the Gene Ontology (GO) and 43.6% for 860 newly discovered human mRBPs. Consistent sensitivity indicates the robust performance of SPOT-seq for predicting RBPs. More importantly, SPOT-seq detects 2,418 novel RBPs in human proteome, 291 of which were validated by the newly discovered mRBP set. Among 291 validated novel RBPs, 61 are not homologous to any known RBPs. Successful validation of predicted novel RBPs permits us to further analysis of their phenotypic roles in disease pathways. The dataset of 2418 predicted novel RBPs along with confidence levels and complex structures is available at http://sparks-lab.org for experimental confirmations and hypothesis generation.

Keywords

RNA binding proteins; DNA binding proteins; KEGG; Human proteome

Introduction

A comprehensive understanding of cellular processes requires identification of RNA-binding proteins (RBP) as well as their ligands. Identification of RBPs is of significant interest because numerous studies have shown that they are key factors associated with cellular processes such as cell cycle checkpoints and genomic stability, and mutations in RBPs are linked to human diseases, including cancer [1]. Recent global analysis indicates that transcripts are not only large in number, but also diverse in localization and function in

*To whom correspondence should be addressed. Tel: 317-278-7674; Fax: 317-278-9201; yaoqi.zhou@griffith.edu.au.

[2]Equal contribution.

Competing Interests: None

cells [2-4]. This implies that underlying post-transcriptional networks are likely larger and more complex than either transcriptional networks or protein-protein interaction networks [5]. However, experimental determination of RNA-binding by every protein is inefficient and impractical, as well as technically challenging and expensive. Attempts at high-throughput biochemical approaches for identifying RBPs progress slowly and are fraught with inaccuracy [5-7]. Thus, computational methods have become a critical component for function annotation and analysis of RBPs [8-16].

Recently, we have developed a template-based technique called SPOT-Seq (RNA) that makes sequence-based prediction of RBPs [16]. In this method, a query sequence is first threaded onto protein template structures of protein-RNA complexes by the fold recognition technique called SPARKS X [17]. The template library contains 1,164 known protein-RNA complex structures on both domain and protein chain levels (95% sequence identity or less). If one of the templates has a good match (according to Z-score) to the query, the structure for the query is predicted and a model complex structure between the predicted structure and the RNA from the template is built. The model complex structure is then employed to predict affinity for protein-RNA-binding using a knowledge-based energy function [15]. If the binding affinity is higher than a threshold, an RBP is predicted. The method achieves a precision of 84% and sensitivity of 47% for a test set of 215 RBPs and 5,765 nonbinding proteins. The precision and sensitivity of SPOT-Seq are more than 10% higher than PSI-BLAST, which uses a sequence-to-profile homology search technique [18]. More importantly, unlike some computational methods, SPOT-Seq (RNA) can distinguish DNA-binding from RNA binding (zero false positives) when applied to 250 DNA binding proteins.

Here, we made a large-scale prediction of RBPs in human proteome using SPOT-Seq and recovered 42.6% for annotated RBPs in human proteome. These predictions, when compared to recently discovered 860 messenger RNA binding proteins in human HeLa cells[19], yielded a consistent sensitivity of 43.6%. More than 2000 novel RBPs are predicted in which 291 proteins are validated by the recently discovered messenger RNA binding proteins. We further showed that some of these novel RBPs are involved in various disease pathways.

Materials and Methods

Fold-recognition and binding-affinity based prediction by SPOT-Seq

SPOT-Seq combines fold recognition and binding affinity prediction for RBP prediction [16]. Each target sequence is aligned to the structures in a template library of 1,164 non-redundant protein-RNA complex structures (95% sequence identity cutoff) by employing the fold recognition method SPARKS X [17]. If the Z-score of the fold recognition is greater than 8.04, a model complex structure between the target protein and template RNA is built by replacing template protein sequence with target protein sequence based on the sequence-to-structure alignment generated from SPARKS X. The model complex structure is then employed to estimate binding affinity according to a statistical energy function based on the distance-scaled finite ideal-gas reference state [20] that was extended to protein-RNA interaction (DRNA) [15]. If the predicted threshold is lower than -0.57, the target protein is predicted as RNA-binding and its complex structure model serves as the basis for the high-resolution prediction of RNA-binding function. The energy and Z-score thresholds (-0.57 and 8.04 respectively) were obtained by optimizing the Matthews correlation coefficient (MCC) based on the leave-homolog-out cross validation with a dataset of 216 RBPs and 5765 non-RNA-binding proteins[16]. We chose to optimize MCC values because MCC is a balanced measure of sensitivity and precision for a training database with an unbalanced number of RNA-binding and non-binding proteins.

Results

Application of SPOT-Seq to human proteome

The human genome dataset from the Uniprot database contains 20,270 unique proteins [21]. The annotations of these genes were obtained from the GO database [22]. We defined an RBP as one whose annotation contains any of the keywords ("RNA binding", "ribosomal", "ribonuclease", or "ribonucleoprotein"). For proteins with keywords "RNA polymerase", we limited to 17 specific GO terms as RBPs (see Table I). This definition leads to 1,217 (6%) proteins annotated as RBPs, while 15,595 proteins are annotated with other functions and 3,458 are unannotated (unknown function). Table I lists the number of proteins found according to the keywords used. Although this definition of RBP is subject to annotation errors/omissions and choices of keywords, it provides a useful reference for analyzing our predicted RBPs.

Application of SPOT-Seq to human proteome identified 2,937 proteins as RNA-binding after removing those proteins whose predicted structures have overlap with predicted transmembrane regions by THUMBUP [23]. This filter is necessary because our method based on protein-RNA complex structures cannot predict the structures of transmembrane proteins. Among 2,937 predicted RBPs, 519 proteins were annotated as RNA-binding and belong to one of the keyword classes shown in Table I. In addition, 1,848 proteins were annotated with functions other than RNA-binding and 570 proteins lack annotations. Fig. 1 shows a bar diagram that indicates the number of predicted RBPs in annotated RBPs, non-RBPs annotated with other function, and proteins with unknown function. The result reveals sensitivity (or coverage) of 42.6% (519/1,217). This sensitivity is consistent with results from our benchmark study [16] despite the latter being based on proteins whose structures were solved in complex with RNA. We noted that the sensitivity strongly depends on specific categories of RBPs. The sensitivity is the highest at 56% for the proteins annotated with the keyword of "RNA binding" and the lowest at 13% with the keyword of "RNA polymerase" as shown in Table I.

Table II lists the top ten templates employed for all predicted RBPs for human proteome. The 60S ribosomal protein L3 encoded by the RPL3 gene (chain C in pdb structure 3o58), is responsible for predicting 1181 proteins with 61 annotated as RNA binding. Four other 60S ribosomal proteins are also in the top-ten list. The unexpected popularity of L3 as a template leads us to examine the accuracy associated with these predictions. SPOT-seq was tested by 215 RBPs and 5,765 non-RBPs [16]. Among these proteins, 11 binding proteins and 15 non-binding targets employed protein chains contained in structure 3o58 as templates. The Matthews correlation coefficient (MCC) for the use of 3o58 chains as templates is 0.64, similar to the overall MCC value of 0.62 when all templates are employed. Thus, the performance for prediction based on 3o58 chains is consistent with the overall performance.

Newly predicted human RBPs have the same non-RNA-binding functions as known RBPs

1,848 predicted novel RBPs were annotated with functions other than RNA-binding. That is, these proteins perform a moonlighting role of RNA-binding. We investigated novel and existing moonlighting RBPs based on their shared molecular functions. In Table III, we tabulate number of proteins and GO terms in molecular function that are unique or shared between predicted and annotated RBPs. More than 90% of predicted novel RBPs [91%, 226/(226+21) for proteins with root annotations only and 98%, 1,238/(1,238+26) for proteins with leaf annotations] shared GO IDs with annotated RBPs. In other words, almost all functions of these predicted moonlighting RBPs are associated with known RBPs. We noted that the entire human proteome has 1,411 leaf GO IDs and annotated RBPs have 288 leaf

GO IDs. That is, 20% of all leaf GO IDs associated with RBPs indicate the extensive association of RBPs with other biological functions.

To illustrate shared functions between predicted and annotated RBPs, we showed four clusters of predicted and annotated RBPs with four GO IDs in Fig. 2. Each GO ID not only contains both predicted and annotated RBPs but also connects with each other through proteins having multiple GO IDs. Top 10 GO IDs (excluding RNA-binding functions) enriched with moonlighting RBPs are listed in Table IV. Many of these 10 GO IDs are associated with transcription regulatory activity, suggesting DNA-binding activity. For example, zinc-ion-binding has an odd ratio of 1.06 enriched with annotated RBPs (fraction of annotated RBPs in a given GO ID in all annotated RBPs versus fraction of all proteins in a given GO ID in all proteins). This odd ratio increases to 1.82 with predicted novel RBPs. Indeed, we found that 350 out of 1,217 annotated RBPs (29%) are also annotated as DNA-binding proteins according to GO annotations. Similarly, 22% (114/519) of predicted and annotated RBPs and 39% (728/1848) of predicted novel moonlighting RBPs are DNA-binding proteins. Thus, a significant fraction of proteins likely interact with DNA and RNA at the same time. It is worth to mention that SPOT-Seq has 100% success rate in discriminating RNA from DNA-binding proteins.[16]

Validation of predicted novel RBPs by proteomic studies of human HeLa cells

Sharing GO IDs between annotated and predicted RBPs supports but does not validate predicted novel RBPs. Direct validation of our predicted RBPs is made possible by a recent proteomic experiment that detected mRNA-binding proteins of HeLa cells [19]. In this study, mRNA-binding proteins (mRBPs) in living HeLa cells were crosslinked by UV irradiation, captured by oligo(dT) magnetic beads after cell lysis, and identified by high resolution nano-LC-MS/MS. They identified 860 mRBPs in which 375 were predicted as RBPs by SPOT-Seq. That is, the sensitivity for this dataset is 43.6%, close to 42.6% sensitivity for all GO annotated RBPs. Similar sensitivity despite significantly different datasets confirms the overall accuracy of SPOT-Seq.

The 860 mRBPs contain many novel RBPs. Using the same GO definition for RBPs as in the human proteome, 746 proteins were novel RBPs in which 291 were predicted as RBPs by SPOT-Seq. Thus, SPOT-Seq has a 39% sensitivity for identification of novel RBPs, close to the sensitivity for all RBPs (42.6%). In these 291 predicted and validated mRNA-binding proteins, the most frequently used templates were chains in PDB ID 3o58 (87 times). This validates the use of 3o58 as a template for predicting RBPs. Moreover, the majority of 291 predicted novel proteins (70%, 203/291) employed a template protein with mRNA binding function.

Castello et al. (2012) also defined a more stringent subset of “previously unknown” RBPs by excluding proteins that were previously experimentally validated, inferable by homology, and/or with a GO annotation containing “RNA” (not just RNA binding). This stringent set of “previously unknown” RBPs contains 315 proteins, 61 of which (19%) are predicted novel RBPs by SPOT-Seq. This large overlap demonstrates the ability of SPOT-Seq to go beyond homology-based inference to uncover novel RBPs.

Disease pathways associated with predicted RBPs

Validation of predicted novel RBPs provides incentive for analyzing their relevance to disease using known disease pathways of Kyoto Encyclopedia of Genes and Genomes (KEGG) database [24]. The KEGG database classified diseases into 11 types: cancer, immune system diseases, nervous system diseases, cardiovascular diseases, digestive diseases, urinary and reproductive diseases, musculoskeletal and skin diseases, respiratory diseases,

congenital disorder of metabolism, and other congenital disorders. These diseases correspond to 176 pathways and 4602 proteins. Among these, 337 are annotated RBPs. 151 (44.8%) of the annotated RBPs are predicted by SPOT-Seq, consistent with the overall sensitivity of 42.6%. In addition to recover known RBPs, SPOT-Seq also predicted 284 novel RBPs. The overall fraction of RBPs, both predicted and annotated, in all proteins involved in disease pathways is about 13%, slightly lower than 18% for all proteins in the human genome. Table V lists the number of annotated and predicted RBPs for the 11 disease pathways. These newly predicted RBPs in disease pathways should be useful for understanding disease mechanisms and generating new hypotheses for experimental testing.

For example, there are 6 diseases such as Charcot-Marie-Tooth disease, progressive myoclonic epilepsy, and pontocerebellar hypoplasia involving in the aminoacyl-tRNA biosynthesis pathway. Eleven annotated RBPs are involved in this pathway, and seven of them were also predicted as RBPs by SPOT-seq. In addition, SPOT-Seq discovered 18 novel RBPs. Most of the predicted novel RBPs (13/18=72%) employed templates that bind with tRNA. Predicted binding with tRNA provide additional supports for our predicted novel RBPs.

Discussion

In this study, a new method for RBP prediction based on known RBP complex structures was applied to human genome. The method uncovered 2,418 proteins that were not previously annotated as RBPs in the GO database. About half of these predicted novel RBPs were annotated as ORFs that lack GO annotations of molecular functions (908), or have only GO root ID (247). Importantly, 284 of these predicted novel RBPs are linked to disease pathways (Table V). Partial validation of this prediction tool includes 12% of these predicted novel proteins (291) that have been identified in a recently study to bind mRNAs in living HeLa cells [19]. The consistent sensitivity (42.6% for annotated RBPs in human genome and 43.6% for mRBPs in HeLa Cells) demonstrates the robustness of SPOTseq in making highly accurate prediction of RBPs.

Among all predicted RBPs, 80.5% have unknown functions or are annotated with functions other than RNA binding. This suggests that many more RBPs exist than those that are currently annotated. If we combine predicted RBPs with annotated RBPs and assume that majority of predicted and annotated RBPs are true, these RBPs would consist of 18% [(1,848+570+1,217)/20,270] of all genes. With the sensitivity of SPOT-Seq being about 43%, the actual number of RBPs is likely to exceed 18% even if when errors were taken into account. The potentially large number of RBPs highlights the scope and significance of the protein-RNA interaction network.

Most of the RBPs predicted here have functions other than RNA-binding. This so-call moonlighting capability of RBPs is consistent with experimental screens of yeast and human proteins. It was found that novel RBPs uncovered in screens often have enzymatic activities as well as RNA-binding architectures [19]. Thus, RBPs that moonlight in other functions is likely to be more common than previously appreciated. In particular, 39% of predicted moonlighting proteins are related to DNA-binding. This is not caused by inability of SPOT-seq to distinguish RNA- from DNA-binding. In fact, the application of SPOT-seq to 250 DNA-binding proteins did not yield any false positive prediction of RBPs [16]. Thus, many proteins can interact with both RNA and DNA.

To prevent potential false positive prediction, we have excluded those predicted RBPs that are transmembrane proteins. This is because our template proteins are all globular proteins. Current implementation, however, requires a separate prediction of transmembrane proteins

and manual comparison. In a future version, we hope to incorporate a transmembrane filter directly inside of SPOT-seq. However, excluding predicted transmemrane regions could lead to removal of some true positive predictions because the accuracy for prediction of transmembrane proteins is 88%[23]. We employed the transmembrane filter in order to further improve the precision of our RBP prediction at a minor cost of fewer predictions.

A surprising result from our template-based technique is that many predicted RBPs employed the templates from 60 S ribosomal proteins, especially L3 (PDB ID 3o58). This is true for both predicted novel and annotated RBPs. We are confident about these predictions because our benchmark test indicates the accuracy of prediction based on 3o58 is the same as that based on other templates. More importantly, 87 novel RBPs based on 3o58 templates are validated as mRNA-binding proteins [19]. The popularity of L3 and other ribosomal protein in predicting RBPs may have its origin in ribosomal proteins being ancient in the tree of life and the potential amplification of genes associated with translation.

To demonstrate the significance of predicted novel RBPs, we employ the serine-protein kinase ATM as an example. ATM was matched to the complex structure between pre-microRNA and chain A of PDB structure 3a6p (exportin-5) by SPOT-Seq despite that the e-value given by PSI-BLAST is 100 (i.e. no homology-inferred function between ATM and exportin-5). Serine-protein kinase ATM is known for DNA-damage induced protein phosphorylation[26] and DNA binding[27]. There is no direct experimental evidence for its binding with RNA. However, a recent paper by Zhang et al[28] found that the ATM kinase is involved in enhancing binding between KH-type splicing regulatory protein (KSRP) and pre-microRNA. The ATM kinase was also found to regulate the interaction between mRNA and HuR[29] and nuclear export of pre-microRNA[30]. Thus, the match between the predicted ATM structure with the binding region of pre-miRNA-binding exportin-5 is likely more than a simple coincidence.

One caveat of the SPOT-seq method is its reliance on known protein-RNA complex structures as templates for predicting complex structures. That is, if no matching template is found, the query protein will be predicted as non-RNA-binding proteins. An *ab initio* structure prediction technique to make structure prediction was not employed because the accuracy of predicted structures by template-free techniques is not yet reliable [31]. The limited number of available templates of protein-RNA complex structures contributes to the sensitivity of our prediction to be approximately 40%. That is, there are a significant number of false negatives. In addition to limited number of templates, inaccurately predicted binding regions due to rigid-body assumption in structural modeling could lead to steric clashes that prevent prediction of high binding affinity and thus lead to a false-negative prediction.

In future, as more protein-RNA complex structures are solved, SPOT-Seq should improve the recovery of known RBPs and uncovering novel ones. Furthermore, it should be possible to increase the sensitivity of SPOT-seq by combing it with other sequence- and structure-based approaches [8-15]. These analyses are in progress, but the ability to double the number of annotated RBPs with SPOT-Seq should generate hypotheses that will impact protein structure/function with relevance to human disease pathways.

Acknowledgments

We would like to thank Bernd Fischer and Professor Matthias Hentze for providing us the list of mRBPs, and Professor Frank Yang for critical reading. This work is supported by National Institutes of Health R01 GM085003 to Y.Z., and 1RO1AI090280 to C.K. SCJ acknowledges support from the School of Informatics at IUPUI in the form of startup funds.

References

1. Lukong KE, Chang KW, Khandjian EW, Richard S. RNA-binding proteins in human genetic disease. Trends Genet. 2008; 24:416–25. [PubMed: 18597886]
2. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 2002; 420:563–73. [PubMed: 12466851]
3. Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, et al. Functional annotation of a full-length mouse cDNA collection. Nature. 2001; 409:685–90. [PubMed: 11217851]
4. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799–816. [PubMed: 17571346]
5. Iioka H, Loiselle D, Haystead TA, Macara IG. Efficient detection of RNA-protein interactions using tethered RNAs. Nucleic Acids Research. 2011; 39:E53. [PubMed: 21300640]
6. Zhang CL, Darnell RB. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. Nature Biotechnology. 2011; 29:607–U86.
7. Galante PAF, Sandhu D, Abreu RD, Gradassi M, Slager N, Vogel C, et al. A comprehensive in silico expression analysis of RNA binding proteins in normal and tumor tissue Identification of potential players in tumor formation. RNA Biology. 2009; 6:426–33. [PubMed: 19458496]
8. Perez-Iratxeta C, Palidwor G, Andrade-Navarro MA. Towards completion of the Earth's proteome. Embo Rep. 2007; 8:1135–41. [PubMed: 18059312]
9. Yu XJ, Cao JP, Cai YD, Shi TL, Li YX. Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. J Theor Biol. 2006; 240:175–84. [PubMed: 16274699]
10. Cai YD, Lin SL. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. Bba-Proteins Proteom. 2003; 1648:127–33.
11. Han LY, Cai CZ, Lo SL, Chung MCM, Chen YZ. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. Rna. 2004; 10:355–68. [PubMed: 14970381]
12. Shao XJ, Tian YJ, Wu LY, Wang Y, Jing L, Deng NY. Predicting DNA- and RNA-binding proteins from sequences with kernel methods. J Theor Biol. 2009; 258:289–93. [PubMed: 19490865]
13. Spriggs RV, Murakami Y, Nakamura H, Jones S. Protein function annotation from sequence: prediction of residues interacting with RNA. Bioinformatics. 2009; 25:1492–7. [PubMed: 19389733]
14. Kumar M, Gromiha MM, Raghava GPS. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J Mol Recognit. 2011; 24:303–13. [PubMed: 20677174]
15. Zhao H, Yang Y, Zhou Y. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucleic Acids Research. 2011; 39:3017–25. [PubMed: 21183467]
16. Zhao H, Yang Y, Zhou Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. Rna Biology. 2011; 8:988–96. [PubMed: 21955494]
17. Yang Y, Faraggi E, Zhao H, Zhou Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics. 2011; 27:2076–82. [PubMed: 21666270]
18. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997; 25:3389–402. [PubMed: 9254694]
19. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, et al. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell. 2012; 149:1393–406. [PubMed: 22658674]

20. ZHOU H, ZHOU Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. PROTEIN SCIENCE. 2002; 11:2714–26. [PubMed: 12381853]
21. Apweiler R, Martin MJ, O'Donovan C, Magrane M, Alam-Faruque Y, Antunes R, et al. The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Research. 2010; 38:D142–D8. [PubMed: 19843607]
22. Sherman BT, Huang DW, Tan QN, Guo YJ, Bour S, Liu D, et al. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics. 2007; 8:426. [PubMed: 17980028]
23. Zhou H, Zhou Y. Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-Markov-model-based method. PROTEIN SCIENCE. 2003; 12:1547–55. [PubMed: 12824500]
24. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Research. 2004; 32:D277–D80. [PubMed: 14681412]
25. Robins P, Pappin DJ, Wood RD, Lindahl T. Structural and functional homology between mammalian DNase IV and the 5′-nuclease domain of Escherichia coli DNA polymerase I. J Biol Chem. 1994; 269:28535–8. [PubMed: 7961795]
26. Cortez D, Wang Y, Qin J, Elledge SJ. Requirement of ATM-dependent phosphorylation of brca1 in the DNA damage response to double-strand breaks. Science. 1999; 286:1162–6. [PubMed: 10550055]
27. Smith GC, Cary RB, Lakin ND, Hann BC, Teo SH, Chen DJ, et al. Purification and DNA binding properties of the ataxia-telangiectasia gene product ATM. Proc Natl Acad Sci U S A. 1999; 96:11134–9. [PubMed: 10500142]
28. Zhang X, Wan G, Berger FG, He X, Lu X. The ATM kinase induces microRNA biogenesis in the DNA damage response. Molecular cell. 2011; 41:371–83. [PubMed: 21329876]
29. Mazan-Mamczarz K, Hagner PR, Zhang Y, Dai B, Lehrmann E, Becker KG, et al. ATM regulates a DNA damage response posttranscriptional RNA operon in lymphocytes. Blood. 2011; 117:2441–50. [PubMed: 21209379]
30. Wan GH, Zhang XN, Langley RR, Liu YH, Hu XX, Han C, et al. DNA-Damage-Induced Nuclear Export of Precursor MicroRNAs Is Regulated by the ATM-AKT Pathway. Cell Rep. 2013; 3:2100–12. [PubMed: 23791529]
31. Zhou YQ, Duan Y, Yang YD, Faraggi E, Lei HX. Trends in template/fragment-free protein structure prediction. Theor Chem Acc. 2011; 128:3–16. [PubMed: 21423322]

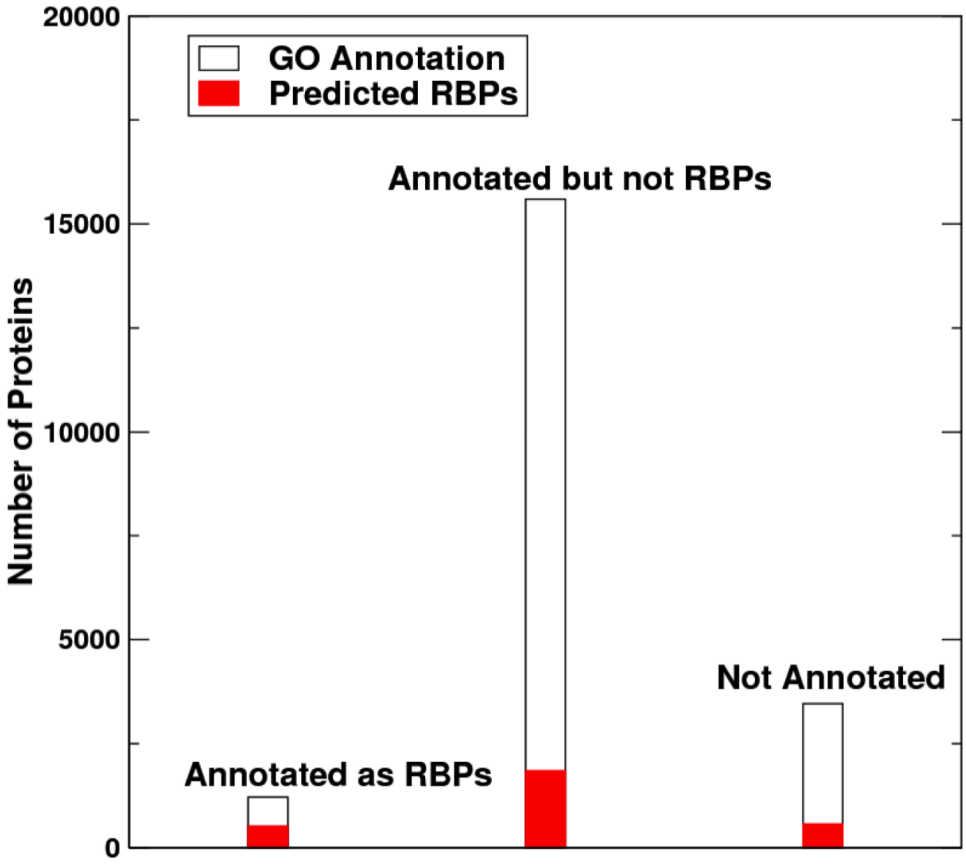

Figure 1.
A bar diagram for annotated RBPs, proteins with functions other than RNA-binding, and proteins with unknown function. All three categories contain predicted RBPs in significant fractions (in red) as shown.

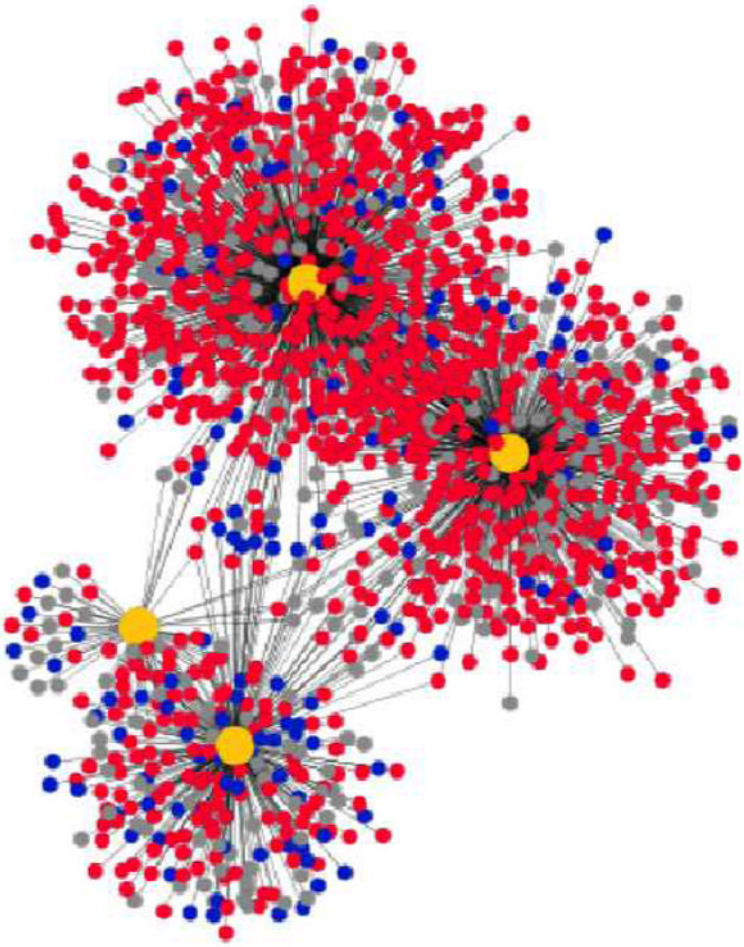

Figure 2.

The connection between proteins with four GO terms (GO:0030528, GO:0008270, GO:0001883 and GO:0000287, in yellow) that are shared by annotated but not predicted (Grey); predicted and annotated (Blue), and predicted and novel (Red) RBPs. Each node represents a protein. One protein can connect to one or more GO terms. This diagram is to illustrate that predicted and annotated RBPs associated with same non-RNA-binding functions

Table I

The number of annotated RBPs according to keywords, compared to the number of those proteins that are predicted as RNA-binding by threading protein sequences onto known protein-RNA complex structures and calculating sequence-structure matching and protein-RNA binding affinity scores(SPOT-seq).

Keywords	Number of Proteins		Sensitivity/ Coverage(%)
	Annotated	Predicted	
RNA binding	722	402	56%
ribosomal	68	37	54%
ribonucleoprotein	240	52	22%
ribonuclease	67	12	18%
RNA polymerase	120	16	13%
Total	1,217	519	43%

GO IDs in RNA polymerase with RNA-binding function. GO:0000428: DNA-directed RNA polymerase complex; GO:0003899: DNA-directed RNA polymerase activity; GO:0003968:RNA-directed RNA polymerase activity; GO:0005665:DNA-directed RNA polymerase II; GO:0005666: DNA-directed RNA polymerase III; GO:0005736:DNA-directed RNA polymerase I complex; GO:0006368:RNA elongation from RNA polymerase II promoter; GO:0006369: termination of RNA polymerase II transcription; GO: 0016591:DNA-directed RNA polymerase II; GO0030880 RNA polymerase complex;GO:0031379:RNA-directed RNA polymerase complex;GO:0031380:nuclear RNA directed RNA polymerase complex;GO:0034062:RNA polymerase activity;GO:0042789:mRNA transcription from RNA polymerase II promoter ;GO: 0042795:snRNA transcription from RNA polymerase II promoter;GO:0042796:snRNA transcription from RNA polymerase III promoter; GO: 0042797:tRNA transcription from RNA polymerase III promoter.

Table II

Top 10 Templates employed for all predicted human RPPs.

PDB ID	Gene Name	Protein Name	# Proteins(#Annotated)	# Nonredudant
3o58C	RPL3	60S ribosomal protein L3	1181(61)	835
1hvuA	gag-pol	Gag-Pol polyprotein	223(12)	177
3o58E	RPL5	60S ribosomal protein L5	180(10)	150
3ciyB	Tlr3	Tol l-like receptor 3	149(2)	54
3o58F	RPL6A	60S ribosomal protein L6A	123(6)	114
3ivkB		Fab light chain	112(0)	17
3a6pA	XPO5	Exportin-5	98(5)	91
3o58b	RPL32	60S ribosomal protein L32	90(5)	82
3o58T	RPL21A	60S ribosomal protein L21A	95(8)	60
1cvjA	PABPC1	Polyadenylate-binding protein	1 58(50)	41

The last letter in PDB ID is the chain ID. # Nonredudant is the number of proteins that are 30% sequence identity or lower among each other.

Table III

GO terms in molecular function that are unique in annotated or predicted RBPs and/or shared between them.

			#of Proteins[b]				# of GO IDs[c]			
			Root		Leaf		Root		Leaf	
Type[a]	Total	None	Unique	Shared	Unique	Shared	Unique	Shared	Unique	Shared
Annotated	1217	118	92	477	47	483	95	189	192	96
A-A∩P	698	102	56	221	29	290	84	178	143	83
A∩P	519	16	36	256	18	193	11	11	39	13
P-A∩P	2418	907	21	226	26	1238	148	189	250	96

[a] A-A∩P (annotated but not predicted RBPs), A∩P (annotated and predicted RBPs), and P-A∩P (predicted but not annotated as RBPs).

[b] The total number of proteins, the number of proteins without GO IDs, with unique GO IDs, and shared GO IDs between predicted and annotated proteins at root and leaf levels.

[c] The number of GO IDs that are unique or shared between predicted and annotated proteins at root and leaf levels.

Table IV

Top 10 GO IDs enriched with annotated and predicted RBPs, ranked according to the number of annotated RBPs

GO-Id	Function	All Proteins	#RBPs			%	
			A	A∩P	P-A∩P	(A/All)	(A+P-A∩P/All)
GO:0008270	zinc ion binding	2307	148	84	604	6%	28%
GO:0030528	transcription regulator activity	1508	138	98	434	24%	35%
GO:0001883	purine nucleoside binding	1599	132	66	136	8%	13%
GO:0005524	ATP binding	1475	129	65	133	8%	13%
GO:0016563	Transcription activator activity	146	44	35	105	30 %	79%
GO:0003702	RNA polymerase II transcription factor activity	245	37	28	67	15%	31%
GO:0000287	magnesium ion binding	454	34	24	32	7%	9%
GO:0003743	translation initiation factor activity	58	29	16	5	50%	31%
GO:0016564	Transcription repressor activity	317	27	19	81	9%	28%
GO:0005525	GTP binding	372	19	14	7	5%	3%

All proteins: all human proteins in the specific GO ID; A: annotated RBPs; A∩P: predicted and annotated RBPs; P- A∩P: predicted but not annotated, novel RBPs; A+P-A∩P: annotated plus predicted novel RBPs.

Table V

Number of proteins and RBPs involved in 11 different phenotypes.

Disease	Pathways	All	Annotated	A∩P	P-A∩P
Cancer	14	372	10	0	41
Immune System	30	1579	53	8	115
Nervous System	30	3740	233	75	253
Cardiovascular	44	2668	157	71	166
Endocrine/Metabolic	24	1603	19	2	106
Digestive	27	2128	41	5	154
Urinary/reproductive	20	1497	14	5	109
Musculoskeletal/skin	61	3152	88	13	225
Respiratory	4	428	0	0	17
Congenital/metabolism	101	3299	103	17	192
Congenital/other	83	3543	198	86	245
Total	176	4602	337	151	284

A∩P: predicted and annotated RBPs; P-A∩P: predicted but not annotated, novel RBPs