In-silico Prediction and Validation of Novel RNA Binding Proteins and Residues in the Human Proteome

Shomeek Chowdhury[1,2#], Jian Zhang[2,3#], Lukasz Kurgan[2,*]

[1] Dr. Vikram Sarabhai Institute of Cell and Molecular Biology, The Maharaja Sayajirao University of Baroda, Gujarat, India, 390005
[2] Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA, 23284
[3] School of Computer and Information Technology, Xinyang Normal University, Xinyang, China, 464000

Emails SC: shomeekchowdhury26@gmail.com
JZ: zhangj943@nenu.edu.cn
LK: lkurgan@vcu.edu

# These authors contributed equally
* Corresponding Author
Phone: +1-804-827-3986; Fax: +1-804-828-2771; email: lkurgan@vcu.edu
401 West Main Street, Room E4225, Richmond, Virginia 23284-3019

Short Title: RNA Binding Proteins in Human Proteome

Keywords:
human proteome
prediction
protein-RNA interactions
RNA binding proteins
RNA binding residues

Abstract

Deciphering a complete landscape of protein-RNA interactions in the human proteome remains an elusive challenge. We computationally elucidate RBPs using an approach that complements previous efforts. We employ two modern complementary sequence-based methods that provide accurate predictions from the structured and the intrinsically disordered sequences, even in the absence of sequence similarity to the known RBPs. We generate and analyze putative RNA-binding residues on the whole proteome scale. Using a conservative setting that ensures low, 5% false positive rate, we identify 1,511 putative RBPs that include 281 known RBPs and 166 RBPs that were previously predicted. We empirically demonstrate that these overlaps are statistically significant. We also validate the putative RBPs based on two major hallmarks of their RNA binding residues: high levels of evolutionary conservation and enrichment in charged amino acids. Moreover, we show that the novel RBPs are significantly under-annotated functionally which coincides with the fact that they were not yet found to interact with RNAs. We provide two examples of our novel putative RBPs for which there is recent evidence of their interactions with RNAs. The dataset of novel putative RBPs and RNA binding residues for the future hypothesis generation is provided in the supplement.

Significance

RNA binding proteins interact with many types of RNAs, are diverse in their subcellular locations, and are involved a wide range of cellular functions. The considerable diversity of the transcripts and RNA binding proteins contributes to claims that the corresponding protein-RNA networks could be larger than transcriptional networks and protein-protein interaction networks. Recent works suggest that the landscape of protein-RNA interactions is largely unexplored and that many more RNA binding proteins remain to be discovered. The discovery of a complete landscape of protein-RNA interactions in human also remains elusive. We use two complementary computational methods that specialize in the predictions from the structured and the disordered regions in protein sequences to elucidate novel RNA binding proteins in the whole human proteome. We identify and computationally validate 1,511 putative RNA binding proteins including 1,230 novel interactors. Moreover, we are the first to provide and analyze putative RNA-binding residues in these proteins. We also demonstrate that these putative novel RNA binding proteins are currently significantly under-annotated functionally. The dataset of the putative RNA binding proteins and residues for the future hypothesis generation is provided with this article.

1 Introduction

RNA binding proteins (RBPs) are known to interact with many types of RNAs including mRNAs, tRNAs, rRNAs, lncRNAs and miRNAs [1]. Studies suggest that these transcripts are abundant and diverse in their subcellular locations and functions [2]. Correspondingly, RBPs are involved a wide range of cellular functions including protein synthesis, post-transcriptional modifications and regulation, RNA transport, packing, stabilization and replication, and mediation of RNA interactions with other macromolecules [3, 4]. They have been also associated with genetic, neurological, neuro-muscular diseases [5]. The considerable diversity of the transcripts and RBPs contributes to a recent claim that the corresponding protein-RNA networks could be larger than transcriptional networks and protein-protein interaction networks [6]. The progress to experimentally identify RBPs is relatively slow and expensive, and suffers inaccuracies and incompleteness [6, 7]. In spite of these difficulties, hundreds of novel RBPs are being discovered every year [8-10]. One way to quantify this growth is to measure the number of non-redundant protein-RNA complexes deposited to the Protein Data Bank (PDB) [11]. Figure 1 reveals that the number of complexes deposited to PDB has grown by close to 6 folds, from 79 that were added between 2000 and 2001 to over 463 that were included between 2016 and 2017. Recent works suggest that the landscape of protein-RNA interactions is largely unexplored and that many more RBPs remain to be discovered [8, 12].

We focus on the computational identification of a complete set of RBPs in the human proteome. This choice is motivated by the obvious importance of this organism and the fact that many human RBPs were already identified and can be used to validate the computational results. More specifically, about 1,500 experimentally annotated human RBPs were recently catalogued by Gerstberger *et al.* [13]. In a follow up article, RNA-binding protein domains collected from across different organisms using the SCOP [14] and Pfam [15] resources were mapped into the human proteome to estimate completeness of these current experimental data [16]. The authors managed to annotate 2,625 RBPs based on their high sequence similarity to the known RNA-binding domains. They showed that approximately 40% of these putative RBPs overlap with the 1,500 experimentally validated RBPs the were assembled by Gerstberger *et al.* [13]. These computational results demonstrate that we are still far from knowing a complete set of human RBPs. Moreover, the approach taken in this study is limited to the proteins that include domains that share high sequence similarity with the already known RNA-binding domains in other organisms. Consequently, it does expand the current knowledgebase of the RNA-binding domains.

A viable alternative to support and extend experimental and sequence-similarity based approaches to elucidate putative RBPs is to use computational methods [17, 18, 19-21]. A recent review summarizes two classes of methods that produce predictions from protein sequences and from protein structures [17]. These methods not only identify RBPs but many of them also predict RNA-binding residues and, in the case of the structure-based methods, atomic levels details of these interactions. In 2014, a protein structure-based predictor, SPOT-Seq [22, 23], was used to identify putative RBPs in the human proteome [24]. This method aligns sequences of human proteins to the library of structures of proteins that are in complex with RNAs. Next, the structures of the input proteins that are sufficiently similar are modelled and their binding affinity to RNA is estimated. This type of an approach is likely to result in a higher coverage than the

annotations that rely on high sequence similarity to the known RNA binding domains. Using SPOT-Seq, the authors generated a list of 2,935 putative RBPs. Analysis by Ghosh and Sowdhamini [16] has revealed that these putative human RBPs share 6.1% proteins in common with the 1,500 known RBPs that were collected by Gerstberger *et al.* [13]. Moreover, about 80% of these putative RBPs lack functional annotations or are annotated with functions other than RNA binding [24]. The combined list of these putative RBPs and the known RBPs was shown to constitute about 18% of the human proteome, pointing to the incompleteness of the current list of human RBPs. The authors also asserted that the actual number of RBPs is likely even higher due to the relatively low sensitivity of SPOT-Seq, which they estimate to be at around 40% [24]. This stems from the fact that SPOT-Seq relies on the limited number of the known protein-RNA complex structures to make predictions.

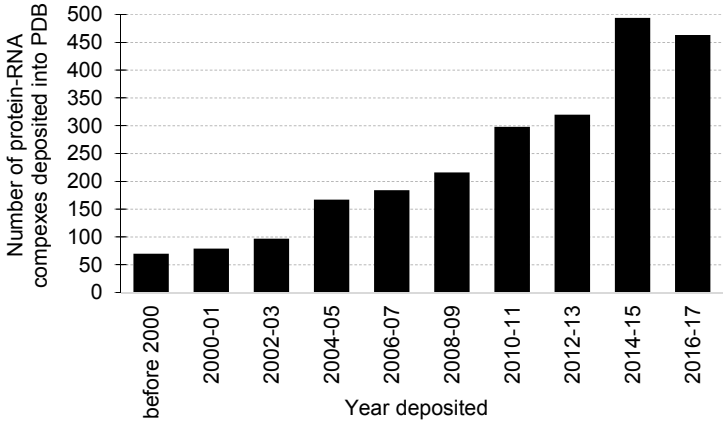

Figure 1. Number of structures of protein–RNA complexes deposited into the PDB. The data were collected on January 31, 2018 and includes non-redundant protein sequences that share identity < 90%.

Our objective is to generate and analyze putative human RBPs using the protein sequence-based methods, which are complimentary to the structure-based SPOT-Seq. The sequence-based methods are designed to provide accurate results in the absence of sequence similarity, even when tested on proteins that share below 30% similarity [25, 26-28]. This will likely lead to a higher coverage of RBPs when compared to the current annotations based on the similarity to the known RNA-binding domains. The sequence-based methods can be used to generate proteome-wide predictions as opposed to the structure-based methods that are limited to proteins for which either native or putative structure is available. Recent analyses shows that only up to 28% of human proteins have either experimental or predicted structures [29]. We note that an older generation of the sequence-based methods was empirically shown to cross-predict DNA and RNA interactions [19, 20]. This substantially reduced predictive quality of these methods that over-predicted the DNA-binding proteins as RBPs. However, recent research resulted in the development of a new generation of predictors that accurately identify RNA binding protein and residues from the protein sequences [26-28].

The defining aspects of our study is that we combine two modern sequence-based methods to comprehensively predict RBPs in the human proteome. The first method, DRNApred [26], focuses on the predictions for structured regions in protein sequences, while the other, DisoRDPbind [28], concentrates on the intrinsically disordered regions. This is motivated by an

observation that many RBPs are intrinsically disordered [3, 10, 30, 31]. The inclusion of these predictions complements the results from the structure-based approaches, which by definition could not be applied to analyze the disordered proteins. Moreover, we are the first to provide and analyze putative RNA-binding residues, while such analysis was not pursued [24] or possible [16] in the previous studies.

2 Materials and methods

2.1 Sequence-based predictors

The sequence-based predictors that can be used to identify putative RBPs can be divided into two groups: methods that directly predict RNA binding proteins vs. methods that predict RNA binding residues [17, 19, 21, 32, 33]. We focus on the latter group of methods since they can be easily used to annotate putative RBPs (i.e., proteins that include putative RNA binding residues) and since the putative RNA binding residues that they generate provide additional information that we use to validate these predictions.

We apply two modern computational methods to predict RNA binding residues in the human protein sequences. The first method, DRNApred [26], was designed to accurately and in high-throughput predict RNA and DNA binding residues and to discriminate between these two types of interactions. DRNApred is based on a logistic regression model. This model makes predictions using a rich set of relevant to the identification of the nucleic acid binding properties of the input protein that are derived directly from the sequence. These properties include empirically selected set of eight physicochemical characteristics of amino acids, evolutionary profiles, and putative intrinsic disorder, secondary structure, and solvent accessibility. The regression maps the numerically quantified properties into real-valued propensities that quantify likelihood for RNA binding for each amino acid in the input protein chain. The mapping was optimized using a large dataset of proteins annotated based on structures of protein-nucleic acid complexes. This dataset is also characterized by a substantially improved coverage of the nucleic acid binding residues when compared to the datasets used in the past [19, 26]. Given the origin of the dataset, DRNApred is predisposed to make predictions for the structured regions in protein sequences. This predictor benefits from several novel design strategies that specifically aim to reduce the cross-prediction of the DNA and RNA interactions. As a result, it was empirically shown to outperform several other representative sequence-based predictors when considering prediction of the RNA binding residues and RBPs [26]. Moreover, DRNApred requires on average only about 15 seconds to predict a single protein and about 2 months to predict the whole human proteome using a single modern CPU [26].

The second method, DisoRDPbind [28], is the first and only method that specifically targets prediction of the RNA binding residues located in the intrinsically disordered regions [34]. DisoRDPbind is implemented using a computationally efficient design that, similar to DRNApred, relies on a comprehensive set of relevant sequence-derived properties of the input protein. The properties include seventeen empirically selected physiochemical properties of amino acids, sequence complexity, putative (derived from the sequence) secondary structure and intrinsic disorder, and sequence alignment. The underlying logistic regression model was optimized using experimentally annotated disordered RNA binding regions collected from the

DisProt resource [35], the database of manually annotated disordered regions. DisoRDPbind is very fast. On average, it generates prediction for a single protein in about a second and can be used to process the complete human proteome in approximately 2 days [28].

Importantly, both DRNApred and DisoRDPbind offer relatively good predictive performance (AUC = 0.67), especially considering the fact that it was tested on a benchmark set of proteins that share below 30% similarity with the training proteins used to optimize these model [26, 28]. Proteins at such low levels of similarity cannot be accurately predicted based on sequence similarity/alignment. Moreover, both of these methods were also shown to generate low false positive rates when tested on proteins that do not interact with RNA, i.e., low rates of incorrectly predicted RNA binding residues in these proteins. Specifically, these false positive rates of DRNApred and DisoRDPbind were empirically estimated to be at 5% [26] and 1% [28], respectively. This suggests that these tools are unlikely to produce false RNA-binding proteins. The ability to accurately predict the low similarity chains together with the low false positive rates and computational cost have motivated application of these tools on the full proteome scale in this project. Moreover, selection of these tools is also driven by the fact that they were already embraced by the end users. The webservers for DRNApred and DisoRDPbind, which are publically available since July 2015 and February 2017, respectively, were already utilized by 962 unique users coming from 60 countries and 435 cities (source: Google Analytics as of February 7, 2018).

2.2 Annotation of sequence-based putative RNA binding residues and proteins

The manually curated human proteome that includes 20,101 proteins was collected from Uniprot [36]. Putative RNA binding was annotated using the two sequence-based methods, DRNApred and DisoRDPbind. We use both methods to generate real-values propensities for the RNA binding for each residue in the input protein chains. The real-values propensities were converted into a binary annotation of RNA binding (RNA binding vs. non-binding residues) using a method-specific cutoff, i.e., residues with propensities above the cutoff were assumed to bind RNA. Based on the published results for these two methods that utilize benchmark datasets that include both RNA binding and non-RNA binding proteins, we select the cutoff value that results in a low false positive rate (i.e., low rate of incorrectly predicted RNA binding residues) equal 5% [26, 28]. This is to ensure that we work with a conservative set of accurate predictions of protein-RNA interactions. The corresponding sensitivity values of DRNApred and DisoRDPbind at this low false positive rate are 18.9% and 19.5%, respectively. Given that these two methods specialize in the prediction of different types of RNA binding regions (structured vs. intrinsically disordered), the expected combined sensitivity should be at about 38%. This is similar to the 40% sensitivity of the previously used structure-based approach SPOT-Seq [24]. We use these putative annotations of the RNA binding residues to generate putative RBPs. Given the 5% residue-level false positive rate of the two predictors, we assume that a given protein binds RNA if the number of putative RNA binding residues in this protein exceeds 5%.

2.3 Annotation of native RNA binding proteins and residues

We secured 1,556 RBPs after mapping the sequences of the native human RBPs collected from the recent study by Gerstberger *et.al*. [13] into the UniProt's human proteomes. This comprehensive set of known human RBPs was established by aggregating results generated with a variety of experimental approaches including next generation sequencing, gel electrophoresis and protein mass spectrometry. The authors have also cross-checked these native RBPs via analysis of evolutionary conservation, tissue-specific expression levels, and interactions for different classes of RNA. However, the original dataset does not provide annotations of the RNA binding residues for these RBPs. We used the BioLiP resource, a semi-manually curated database of protein-ligand interactions extracted from structures of protein-ligand complexes [37], to annotate the binding residues. BioLiP uses all currently available and updated weekly complex structures, collected primarily from PDB, to generate list of residues that interact with a wide range of over 22 thousand ligands including RNAs. We managed to annotate 5,408 RNA binding residues in 124 of the 1556 native RBPs using BioLiP. The coverage is relative low since most of the currently known human RBPs were not yet solved structurally in complex with the RNA. We compare certain key characteristics of these native RNA binding residues with the putative RNA binding residues that were generated with the sequence-based methods.

2.4 Characterization of hallmarks of RNA-binding residues

A recent survey has characterized major hallmarks of the RNA, DNA and protein binding residues that were discussed in the literature and validated empirically [33]. It shows that the RNA binding residues tend to have higher than expected evolutionary conservation and are enriched in the positively charged amino acids (Arg and Lys). The latter is a consequence of the ionic interactions between the positively charged residues of the protein and the phosphate group of RNA [38]. We assess these two hallmarks for the native RNA binding residues and compare them with the corresponding values for the putative RNA binding residues.

We quantify the evolutionary conservation for each residues in the considered human protein chains using relative entropy (RE) [39]:

$$\mathrm{RE} = \sum_{i=1}^{20} p_i log_2(\frac{p_i}{p_{ib}})$$

where i iterates over the 20 amino acid types, p_i is the emission frequency of the i^{th} amino acid type in the alignment profile computed based on the multiple sequence alignments generated with HHblits (HMM-HMM–based lightning-fast iterative sequence search) [40], and p_{ib} is the background frequency of the i^{th} amino acid found in naturally occurring protein sequences We compute the latter using the *nr* dataset. We utilize HHblits since it was shown to produce more accurate alignments while also being faster and more sensitive when compared to other popular multiple alignment tools like PSI-BLAST and HMMER [40]. We have run HHblits with the default parameters and uniprot20 database. Moreover, the relative entropy was empirically demonstrated to be more sensitive to detect functional sites (such as RNA binding residues) when compared to entropy (which does not utilize the p_{ib} values) and several other evolutionary conservation measures [39].

Table 1. Summary of the putative human RNA binding proteins and residues generated with the two sequence-based predictors.

Method used	Number of putative RNA binding proteins	Number of putative RNA binding residues	Total number of residues for the putative RNA binding proteins
DisoRDPbind	251	10,892	91,344
DRNApred	1,260	168,465	576,754
Total	1,511	179,357	668,098

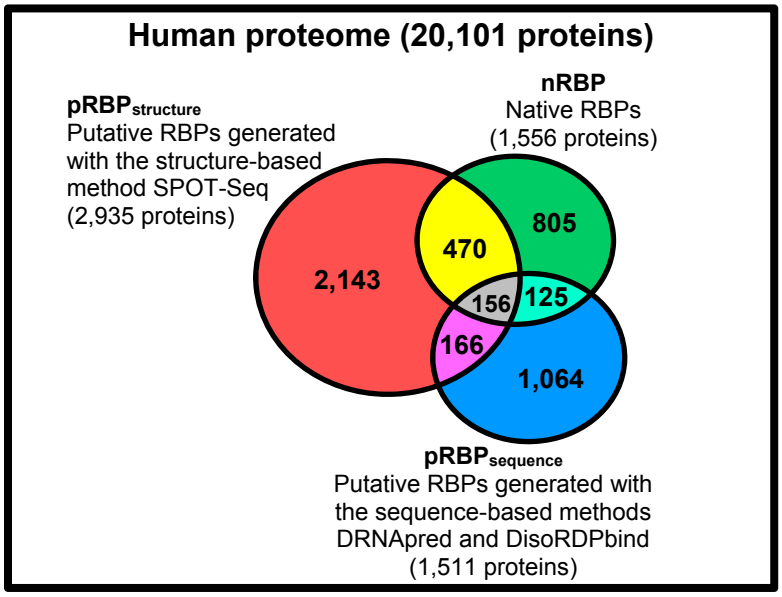

Figure 2. Venn diagram of the overlap between the set of 1,556 native RBPs (nRBP), 1,511 RBPs predicted with the sequence-based methods ($pRBP_{sequence}$) and 2,935 RBPs predicted with the structure-based methods ($pRBP_{structure}$). The outside box is at the scale of the size of the complete human proteome. The geometrical proportions of the plot were generated using eulerAPE application [41].

3 Results and discussion

3.1 Sequence-based putative human RBPs

We use the putative RNA binding residues generated by the two complementary methods, DRNApred and DisoRDPbind, to annotate the putative RBPs (see Section 2.2 for details). Table 1 summarizes these results. In total, we found 1511 putative RBPs, with 251 and 1260 having disordered and structured RNA binding regions, respectively. Since the two methods target different types of proteins (structured vs. disordered), two sets of RBPs are disjoint. This support our claim that the combined sensitivity of the two predictors should be around 38%. We predict close to 180 thousand RNA binding residues in these 1511 proteins, for an overall per sequence content of RNA binding residues at 26.8%. The putative RBPs constitute about 7.5% of the

human proteome (20,101 proteins). The 1511 RBP are provided in the supporting file S1 in the Supplement. For each protein we include its UniProt accession number, annotation whether this is an already experimentally validated RBP (based on the list of 1556 native RBPs from Gerstberger *et al.* [13]) or a novel RBP, name of the method that was used to generate this RBP and annotation of the putative RNA binding residues.

We analyze relation between the three sets of RBPs: 1,556 native RBPs collected from Gerstberger *et al.* [13] (nRBP), 1,511 putative RBPs generated by our sequence-based predictors ($pRBP_{sequence}$) and 2,935 putative RBPs generated by the structure-based predictor SPOT-Seq[24] ($pRBP_{structure}$). Figure 2 shows Venn diagram that visualizes overlap between these RBP sets. The $pRBP_{sequence}$ set includes 281 native RBPs (gray- and teal-colored areas in Figure 2) and 1230 novel RBPs. The latter set also includes 166 putative RBPs that were predicted with SPOT-Seq (pink area in Figure 2). The amount of the overlap between the native RBPs and each of the two sets of the putative RBPs is similar. About 18.6% of the sequence-based putative RBPs are in common with the set of native RBPs, compared to 21.4% of the structure-based predictions. Figure 3 evaluates statistical significance of these overlap values. We compare overlap between the nRBP set and a set of 500 randomly selected human protein vs. the overlap between the nRBPs and a set of 500 randomly selected putative RBPs produced by one of the methods. We repeated this experiment 100 times. Figure 4 shows the corresponding averages and standard deviations over these repetitions. As expected, the average overlaps with the $pRBP_{sequence}$ and $pRBP_{structure}$ sets equal 18.8% and 21.4%, respectively. The average overlap with a generic set of human proteins is much smaller and equals 7.7%. This number stems from the fact that there are 1556 native RBPs among the 20,101 human proteins. The latter overlap is significantly smaller than the overlap with ether set of the putative RBPs (p-value < 0.001). This suggests that both sets of putative RBPs are significantly biased toward inclusion of the native RBPs.

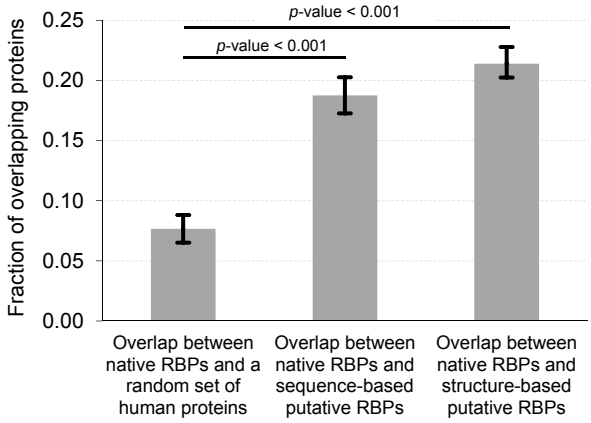

Figure 3. **Comparison of overlap between the putative and native RBPs.** The calculation of overlap was sampled 100 times, each time using the complete set of 1556 native RBPs and selecting at random 500 human proteins (left-most bar) or 500 putative RBPs (the other two bars). The bars and error bars give the average overlap values and the corresponding standard deviations, respectively, computed over the 100 tests. Since the distributions of the overlap value are normal (based on the Anderson-Darling test at 0.05 significance) *t*-test was used to evaluate statistical significance of differences between the overlap of native RBPs and random human proteins and between overlap of native and putative RBPs. Results are shown above the bars; we include only the statistically significant differences with *p*-values < 0.05.

We further analyze 322 proteins that overlap between the sequence-based predictions and the structure-based predictions (gray and pink regions in Figure 2). They include 277 proteins that were predicted by both DRNApred and SPOT-Seq and 45 that were predicted by both DisoRDPbind and SPOT-Seq. The latter set covers 45 out of 251 = 18% of the putative RBPs generated by DisoRDPbind and includes proteins with disordered RNA-binding regions that likely undergo binding-induced folding [31].

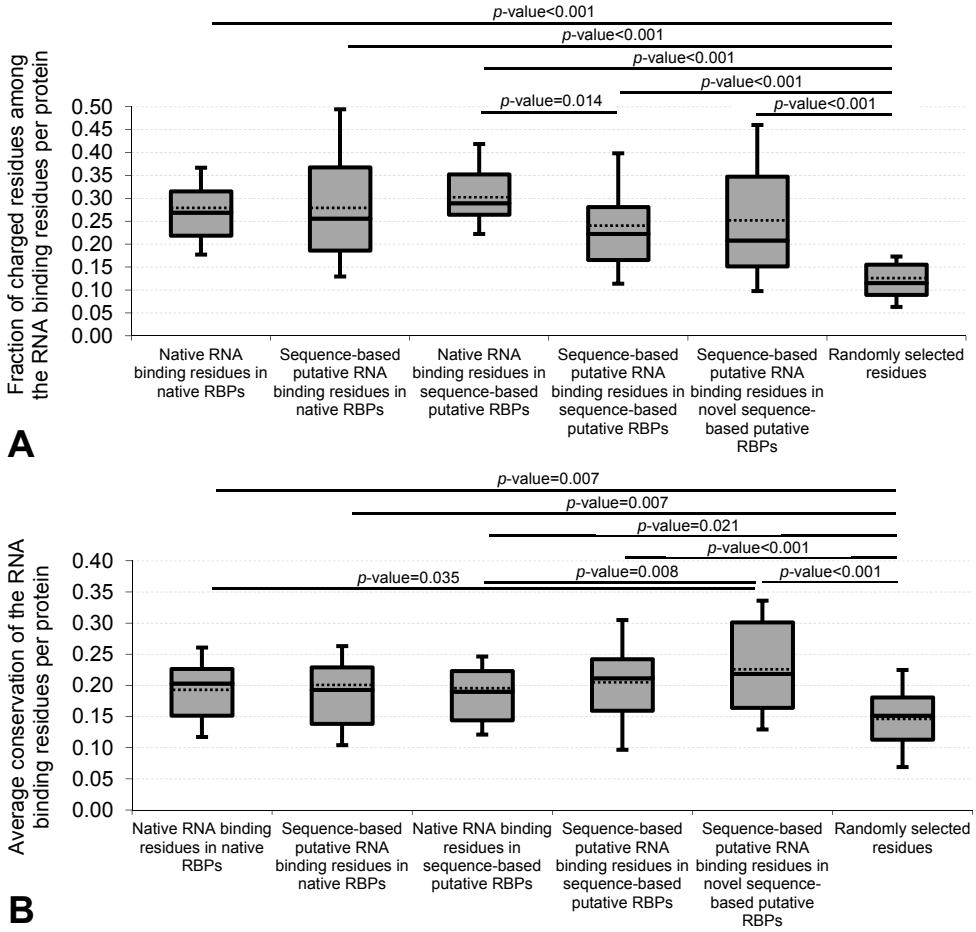

Figure 4. **Average content of charged residues and evolutionary conservation for the RNA binding residues in the native RBPs, sequence-based putative RBPs, and a generic set of human proteins.** Panel A compares the fraction of charged residues (Arg and Lys) while panel B focuses on the evolutionary conservation. We consider native RBPs, sequence-based putative RBPs and novel sequence-based putative RBPs, which are annotated with the native RNA binding residues and sequence-based putative RNA binding residues. We also compute a baseline results that is based on a random set of residues selected from a randomly drawn set of human proteins such that the number of protein and residues equals to the number of native RBPs and native RNA binding residues, respectively. The plots show the 10th centile (bottom whisker), first quartile (bottom of box), median (horizontal line inside the box), third quartile (top of the box), and 90th centile (top whisker) of the per protein fraction of charged residues and average conservation. Mean values are shown using dotted horizontal lines. To ensure robustness of the statistical analysis we randomly choose the same number of proteins for each analysis, which is equal to the smallest set of the sequence-based putative RBPs that have native RNA binding residues. Since these distributions are normal (based on the Anderson-Darling test at 0.05 significance), *t*-

test was used to evaluate statistical significance of differences between distributions of the charge content and conservation for every pair of the six protein sets. Results are shown above the bars; we include only the statistically significant differences with *p*-values < 0.05.

3.2 Validation of the sequence-based putative human RBPs based on the annotations of RNA binding residues

We validate the annotations of the putative RBPs based on the annotations of the RNA binding residues. We compare the values of the two main hallmarks of RNA binding residues between the native RNA binding residues and the putative RNA binding residues. The latter were used to annotate the putative RBPs. The hallmarks include the enrichment in charged amino acids (Arg and Lys) and high levels of evolutionary conservation.

Figure 4A summarizes the results of analysis of the content of charged amino acids, defined as the fraction of the charged amino acids among all considered (either putative or native) RNA binding amino acids. We compute the content values for the native RNA binding residues that are annotated in the native RBPs as well as in the sequence-based putative RBPs. We compare these values to the content of the sequence-based putative RNA binding residues predicted in the native RBPs, putative RBPs, and novel putative RBPs. As a baseline, we compute the content of the charged residues among randomly selected residues. More precisely, we select at random the same number of proteins and residues as the number of the native RNA binding residues in the native RBPs to calculate the baseline content. The mean content of the charged residues for a generic (randomly selected) residues is 0.122. This in agreement with the overall abundance of Arg and Lys. A recent study list the Arg and Lys content in the eukaryotic proteomes at 0.057 and 0.056, respectively [42]. As expected, Figure 4A reveals that the content of charged residues among the native RNA binding residues in the native RBPs and in the sequence-based putative RBPs is significantly higher and equals 0.278 and 0.303, respectively (p-value < 0.001). More importantly, the content for the sequence-based putative RNA binding residues is also similarly high and equals 0.280 for the native RBPs, 0.237 for all sequence-based putative RBPs, and 0.254 for the 1230 novel sequence-based putative RBPs. These content values are significantly higher than the baseline content (p-value < 0.001). This means that the putative RNA binding residues, which we use to annotate putative RBPs, share the same levels of enrichment in charged residues as the native RNA binding residues.

Figure 4B focuses on the analysis of the evolutionary conservation. Similar to the analysis of the charged residues, we compare the average value of conservation for the native and putative RNA binding residues, and random/generic residues, which are computed per protein. The average conservation for the randomly chosen (baseline) residues equals 0.155. Consistent with other studies [33], we show that the average conservation among the native RNA binding residues in the native RBPs and in the sequence-based putative RBPs is significantly higher than the baseline. It equals 0.193 (*p*-value = 0.007) and 0.185 (*p*-value = 0.021), respectively. Similarly, the conservation of the sequence-based putative RNA binding residues is also significantly higher. Figure 4B show that this is true for the native RBPs (average conservation = 0.196, *p*-value = 0.007), for all sequence-based putative RBPs (average conservation = 0.213, *p*-value < 0.001) and for the novel putative RBPs (average conservation = 0.229, *p*-value < 0.001).

The significantly higher content of the charged residues and evolutionary conservation among the putative RNA binding residues, which are on par with the values of these hallmarks for the native RNA binding residues, suggest that the putative residue-level annotations are likely to be correct. This in turn validates the release of the putative RBPs that were annotated based on these putative binding residues.

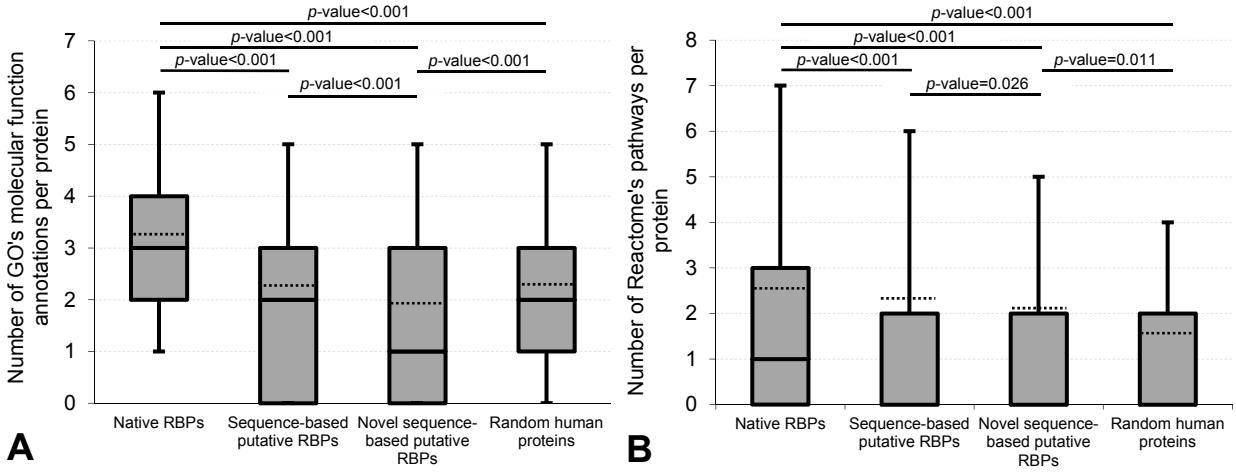

Figure 5. **Comparison of the counts of functional annotations between the native RBPs, putative RBPs, and a generic set of human proteins.** Panel A compares the number of molecular functions annotated using GO computed per protein. Panel B contrasts the number of pathways extracted with Reactome calculated per protein. The plots show the 10th centile (bottom whisker), first quartile (bottom of box), median (horizontal line inside the box), third quartile (top of the box), and 90th centile (top whisker). Mean values are shown using dotted horizontal lines. Since the distributions of the counts are not normal (based on the Anderson-Darling test at 0.05 significance), Wilcoxon rank sum test was used to evaluate statistical significance of differences between distributions of counts for every pair of the four protein sets. Results are shown above the bars; we include only the statistically significant differences with p-values < 0.05.

3.3 Novel sequence-based putative RBPs are functionally under-annotated

We hypothesize that the reason why the sequence-based putative RBPs were not yet annotated is that these proteins are in general under-annotated functionally. To study that, we compare the per-protein rates of functional annotations in the native RBPs, all putative RBPs, novel putative RBPs and a generic (randomly chosen) set of human proteins. Figure 5 presents the results when considering two types of functional annotations: GO molecular function terms [43] collected from Uniprot [36] (Figure 5A), and pathways that were obtained from the Reactome resource [44] (Figure 5B). We count the number of these functional annotations per protein and the figure summarizes distributions of these counts for the four protein sets, including the mean and median values.

Figure 5 reveals that the native RBPs have on average 3.2 GO molecular functions and are assigned to 2.7 pathways. Moreover, only 4.9% of these proteins lack molecular function terms and 43.5% are not associated with pathways. The other three protein sets are characterized by

lower functional annotation counts. Compared to the native RBPs, the sequence-based putative RBPs have significantly less GO molecular functions (average = 2.2, *p*-value < 0.001) and pathways (average = 2.3, *p*-value < 0.001). Moreover, a much larger fraction of 26.7% and 54.7% of these proteins do not have any GO molecular function terms and pathway terms, respectively. Similar trend is observed when comparing the 1230 novel putative RBPs with the native RBPs. They also have significantly fewer GO terms (average = 1.96, *p*-value < 0.001) and associated pathways (average = 2.0, *p*-value < 0.001) and much more of them have no annotations (32.1% and 58.3%, respectively). Interestingly, the generic set of human proteins is also relatively sparsely annotated. These proteins have significantly fewer GO molecular function terms when compared to the native RBPs (average = 2.25, *p*-value < 0.001). However, they actually have significantly more of these terms than the novel putative RBPs (*p*-value < 0.001). Also, fewer human protein have no GO terms compared to the novel putative RBPs (24.1% vs. 32.1%). These observation are also consistent when considering the pathway annotations. The human proteins are associated with significantly fewer pathways compared to the native RBPs (average = 1.7, *p*-value < 0.001) but also fewer of them lack pathway annotations when compared to the novel putative RBPs (50.5% vs. 58.3%). Moreover, the novel putative RBPs have significantly fewer GO terms (*p*-value < 0.001) and pathways (*p*-value = 0.025) when compared to the set of all putative RBPs that were generated by our sequence-based predictors and which include some of the native RBPs.

Overall, these results demonstrate that the sequence-based putative novel RBPs are significantly under-annotated when compared to native RBPs. We believe that this is linked to the fact that they are yet to be recognized as interacting with RNAs. The novel RBPs also have significantly fewer molecular function terms when contrasted against generic human proteins. That suggest that the completeness of their functional annotations is below the expected levels.

3.4 Sequence-based predictions of RBPs have low false positive rate

We also estimate false positive rates of the sequence-based predictions with DRNApred and DisoRDPbind. We evaluate the rate of putative RBPs generated by these tools on a set of human proteins that are unlikely to interact with RNA. Inspired by a procedure by Zhou *et al.* [22], we generate this dataset in three steps. In the first step we cluster the human proteome with BLASTCLUST at 25% similarity and remove all clusters that include any of the 1556 native RBPs that we collected from Gerstberger *et al.* [13]. Next, among the resulting 13,101 clusters we remove proteins that have functional annotations that may suggest interactions with RNA and those that are lacking functional annotations. More specifically, we remove proteins with the UniProt names and GO terms that include RNA, ribosome, ribosomal keywords, as well as proteins that do not have molecular function and biological process GO terms in UniProt. In the third step, we uniformly sample the remaining 10,484 proteins based on their sequence similarity to avoid biasing the results to certain protein families. We cluster with BLASTCLUST at 25% similarity and select one protein from each of the resulting 7332 clusters. The 7332 non-RNA binding human proteins together with a list of false positive predictions from the DisoRDPbind, DRNApred and SPOT-Seq that overlap with these proteins are included in the supporting file S2 in the Supplement. DisoRDPbind and DRNApred predict 48 and 372 RBPs in this dataset. This corresponds to 48/7332 = 0.6% and 5.1% false positive rates, respectively, and is in good agreement with their previously estimated false positive rates of 1% and 5%[26, 28]. To compare,

the structure-based SPOT-seq identifies 714 RBPs among the 7332 non-RNA binding human proteins, which corresponds to also a relatively low 9.7% false positive rate.

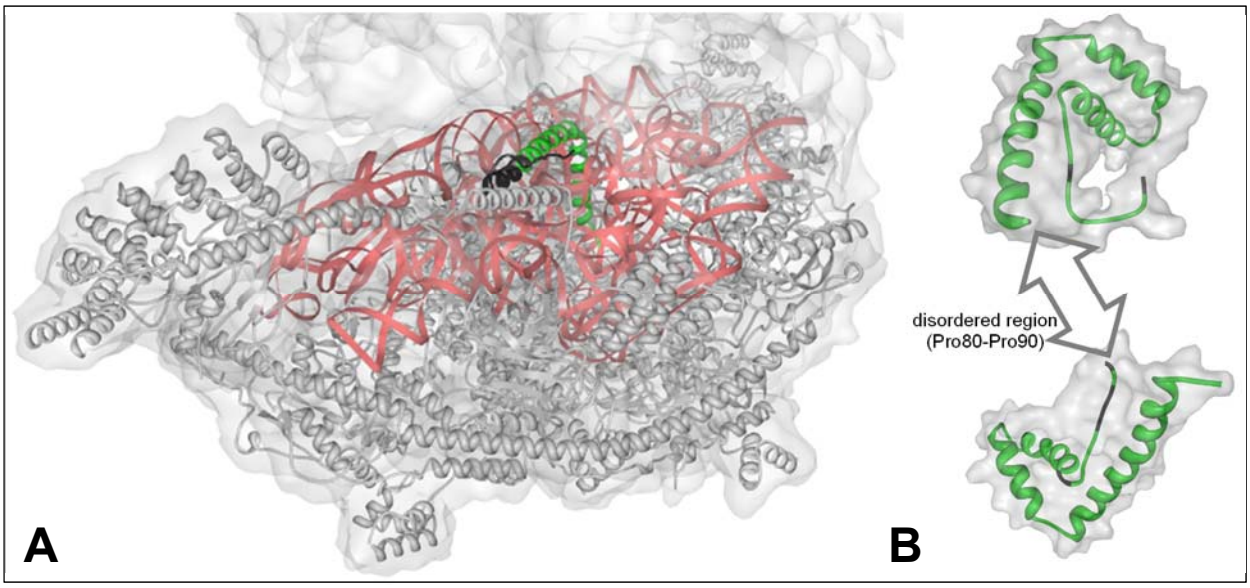

Figure 6. **Structures of the two case study proteins.** Panel A shows structure of the Aurora kinase A-interacting protein (AURKAIP1; UniPRot accession number: Q9NWT8; PDB identifier: 3J9M chain A3). Structure of this protein, which covers positions 128 to 196, is shown as a part of the small subunit of the human mitochondrial ribosome; the surface of the lower part of large subunit is visible at the top of this panel. Gray and red structures denote proteins and RNAs, respectively. Panel B shows structure of the high mobility group protein B3 (HMGB3; UniPRot accession number: O15347; PDB identifiers: 2EQZ (top structure) and 2YQI (bottom structure)). The structures of the tw proteins are shown with green and black ribbons where black denotes putative RNA binding residues predicted by DisoRDPbind (Q9NWT8) and DRNApred (O15347). The structures were visualized using Protein Workshop [45].

3.5 Case studies

We further investigate two putative novel RBPs to explain and validate these predictions. We selected based on three criteria. First, they must have at least a partial structure in PDB, so we can visualize the putative RNA binding residues. Second, one protein should be predicted with DisoRDPbind and the other with DRNApred. Third, they should target different types of RNA.

Aurora kinase A-interacting protein [46] (AURKAIP1; UniProt accession number: Q9NWT8; PDB identifier: 3J9M chain A3) has 199 amino acids and is annotated in UniProt as ribonucleoprotein. AURKAIP1 is a component of mitochondrial small ribosomal subunit, a large protein-RNA complex that includes 12S ribosomal RNA and 30 proteins [47]. Figure 6A shows the structure of the small subunit where the protein of interest is visualized using green and black color; the black denotes RNA binding residues that were predicted with DisoRDPbind. This protein has elongated conformation and thus is likely to be disordered outside of the complex, which is why it was predicted by DisoRDPbind. The structure of AURKAIP1 covers the C-terminus of this protein (positions 128 to 196). Figure 6A reveals that the entire C-terminus interacts with RNA, including the part that was correctly predicted by DisoRDPbind. This

protein was identified as a part of a mammalian mitochondrial ribosome in 2013 [46] and more recently this finding was confirmed when structures of the small subunit [48] and the entire mitochondrial ribosome [47] were released. These studies have occurred after the set of the native RBPs was developed, which is why AURKAIP1 was not included. However, our sequence-based prediction was able to correctly identify this protein as RBP.

High mobility group protein B3 (HMGB3; UniPRot accession number: O15347; PDB identifiers: 2EQZ and 2YQI) has 200 amino acids. This multifunctional protein associates with chromatin, regulates B-cell and myeloid cell differentiation, and was suggested (via sequence similarity) to act as a cytoplasmic immunogenic DNA/RNA sensor. It includes two HMG-box domains, first spanning positions 1 to 79 (PDI identifier: 2EQZ) and the second between positions 91 and 164 (PDI identifier: 2YQI) [49]. The short Pro80 to Pro90 region between these two domains is intrinsically disordered based on the results produced using the MobiDB resource [50]. Figure 6B shows the two domains and the annotation of the putative RNA binding residues generated with DRNApred. We computed solvent accessibility for the putative RNA binding residues using the DSSP software [51]. Six out of the seven putative RNA binding residues that are located in these two structures (Met8 and Lys17 in 2EQZ; Asn98, Lys101, Arg102 and Phe107 in 2YQI) are solvent exposed (relative solvent accessibility > 25%) and thus they would be accessible to interact with RNA. Our putative annotation of HMGB3 as RBP is in agreement with a recent study that has shown that mouse HMGB3 interacts with both RNAs and DNA *in vitro* [52]. Another more recent experiment that aimed at finding RBPs that are involved in the innate immune response further confirms that HMGB3 indeed interacts with RNA [53].

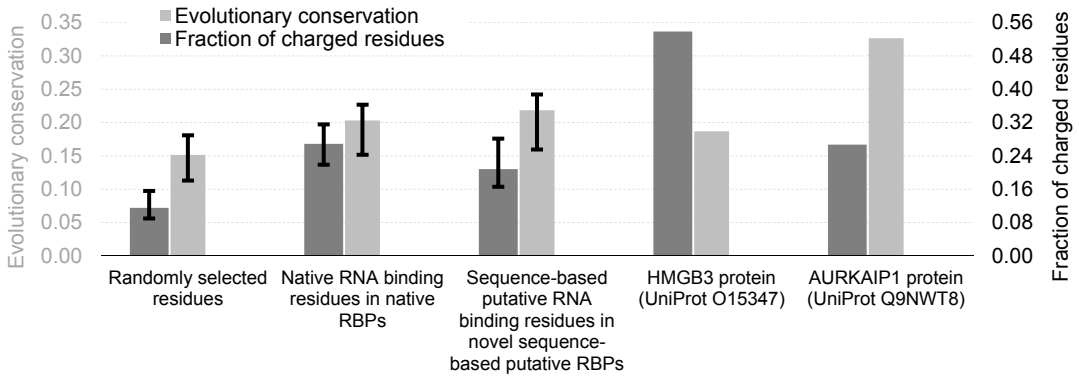

Figure 7. Content of charged residues and evolutionary conservation. We compare these two hallmarks for the putative RNA binding residues in the AURKAIP1 and HMGB3 proteins against the per protein averages for the randomly selected residues in a generic set of human proteins, native RNA binding residues in the native RBPs, and sequence-based putative RBPs in the novel putative RBPs. Dark gray bars and the *y*-axis on the right compare the fraction of charged residues (Arg and Lys) while light gray bars and the *y*-axis on the left correspond to the evolutionary conservation. The results for the first three sets of bars include the median (bar) and the first and third quartiles (error bars) over the per-protein values in a given protein sets, while the last two sets of bars are for a single protein.

Figure 7 compares the average evolutionary conservation and fraction of charged residues computed for the putative RNA binding residues in the AURKAIP1 and HMGB3 proteins versus the corresponding per protein averages for the generic (randomly chosen) set of human residues, native RNA binding residues, and sequence-based putative RNA binding residues in the novel

RBPs. The putative RNA binding residues in the two proteins have much higher conservation levels and are enriched in the charged residues when compared to the generic set of residues. Moreover, these two hallmarks for the putative RNA binding residues for AURKAIP1 and HMGB3 are on par with the native RNA binding residues and the sequence-based putative RNA binding residues. These observation further confirm that these two proteins indeed interact with RNAs and that these hallmarks are indicative of the protein-RNA interactions. Altogether, we conclude that our sequence-based approach has correctly identified the two putative novel RBPs.

4 Concluding remarks

Recent years have seen a strong push to identify RBPs in the human proteome. This problem was addressed with the help of experimental [13] and computational [16, 24] methods. However, we are still a long way from deciphering the complete landscape of protein-RNA interactions in human. This work complements the previous efforts by exploring a different class of computational methods. We employ methods that predict propensity of protein sequences to interact with RNAs, in contrast to the previously utilized computational methods that generated the putative RNA-binding proteins from the protein structures. We use two complementary sequence-based methods that specialize in the predictions from the structured and the disordered regions in protein sequences. These recently released sequence-based methods were shown to give accurate results even in the absence of sequence similarity to the known RNA-binding proteins [26, 28], allowing us to comprehensively scan the whole human proteome. Moreover, we are the first to provide and analyze putative RNA-binding residues.

We identified 1,511 putative RBPs including 281 proteins that are already known to interact with RNA and 1230 novel RBPs. These novel RBPs include 166 proteins that were previously predicted with the structure-based approach [24]. We show that the overlap between our predictions and the other native and putative RBPS is statistically significant. We also validate the putative RBPs based on the annotations of the putative RNA binding residues that we use to annotate the putative RBPs. We show that the values of the two main hallmarks of the RNA binding residues, high levels of evolutionary conservation and enrichment in charged amino acids, are similar between the known native RNA binding residues and the putative RNA binding residues. Moreover, these values of the two hallmarks are significantly higher than expected, which we assessed by comparing them against conservation and content of charged residues in generic human proteins. Finally, we empirically show that the computational tools that we use are characterized by low false positive rates in human proteins. These observations provide support for the release of the putative RBPs that we generated.

We also attempt to explain why these putative RBPs were not yet identified. Our empirical evaluation of the current functional annotations of these proteins has revealed that they are significantly under-annotated when compared to the native RBPs and typical human proteins. The lower than expected completeness of the functional annotations for the putative RBPs suggests that they were so far under-studied, which coincides with the fact that they were not yet found to interact with RNAs. We also provide examples of two putative RBPs for which we found recent evidence that demonstrates that they in fact interact with RNAs.

Our study is parallel to a number of other projects that aim at a comprehensive, genome wide identification of the DNA binding proteins. This projects similarly utilize a mixture of experimental and computational methods [54]. An interesting intersection of these efforts are the proteins that bind both RNA and DNA. Recent article suggests that as many as 2% of the human proteins may in fact interact with both nucleic acids [55]. Thus, besides providing a high quality set of putative novel RBPs, our analysis will also contribute to the identification of these functionally important proteins.

Acknowledgments

This research was supported in part by the Qimonda Endowment to and the National Science Foundation grant 1617369 to Lukasz Kurgan.

The authors have declared no conflict of interest.

References

[1] G. Giudice, F. Sanchez-Cabo, C. Torroja, E. Lara-Pezzi, Database (Oxford) 2016, 2016; K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, T. R. Hughes, Nucleic Acids Res 2011, 39, D301.

[2] J. Kawai, A. Shinagawa, K. Shibata, M. Yoshino, M. Itoh, Y. Ishii, T. Arakawa, A. Hara, Y. Fukunishi, H. Konno, J. Adachi, S. Fukuda, K. Aizawa, M. Izawa, K. Nishi, H. Kiyosawa, S. Kondo, I. Yamanaka, T. Saito, Y. Okazaki, T. Gojobori, H. Bono, T. Kasukawa, R. Saito, K. Kadota, H. Matsuda, M. Ashburner, S. Batalov, T. Casavant, W. Fleischmann, T. Gaasterland, C. Gissi, B. King, H. Kochiwa, P. Kuehl, S. Lewis, Y. Matsuo, I. Nikaido, G. Pesole, J. Quackenbush, L. M. Schriml, F. Staubli, R. Suzuki, M. Tomita, L. Wagner, T. Washio, K. Sakai, T. Okido, M. Furuno, H. Aono, R. Baldarelli, G. Barsh, J. Blake, D. Boffelli, N. Bojunga, P. Carninci, M. F. de Bonaldo, M. J. Brownstein, C. Bult, C. Fletcher, M. Fujita, M. Gariboldi, S. Gustincich, D. Hill, M. Hofmann, D. A. Hume, M. Kamiya, N. H. Lee, P. Lyons, L. Marchionni, J. Mashima, J. Mazzarelli, P. Mombaerts, P. Nordone, B. Ring, M. Ringwald, I. Rodriguez, N. Sakamoto, H. Sasaki, K. Sato, C. Schönbach, T. Seya, Y. Shibata, K. F. Storch, H. Suzuki, K. Toyo-oka, K. H. Wang, C. Weitz, C. Whittaker, L. Wilming, A. Wynshaw-Boris, K. Yoshida, Y. Hasegawa, H. Kawaji, S. Kohtsuki, Y. Hayashizaki, Nature 2001, 409, 685; E. P. C. The, Nature 2007, 447, 799.

[3] Z. Peng, C. J. Oldfield, B. Xue, M. J. Mizianty, A. K. Dunker, L. Kurgan, V. N. Uversky, Cell Mol Life Sci 2014, 71, 1477.

[4] D. N. Wilson, K. H. Nierhaus, Crit Rev Biochem Mol Biol 2005, 40, 243; Z. Li, P. D. Nagy, RNA Biol 2011, 8, 305; G. Dreyfuss, V. N. Kim, N. Kataoka, Nat Rev Mol Cell Bio 2002, 3, 195; A. Cassola, G. Noe, A. C. Frasch, RNA Biol 2010, 7, 339; B. M. Lunde, C. Moore, G. Varani, Nat Rev Mol Cell Biol 2007, 8, 479; C. G. Burd, G. Dreyfuss, Science 1994, 265, 615; P. Lasko, Sci STKE 2003, 2003, RE6.

[5] K. E. Lukong, R. E. Fatimy, in *eLS*, John Wiley & Sons, Ltd, 2001; S. Gerstberger, M. Hafner, M. Ascano, T. Tuschl, Adv Exp Med Biol 2014, 825, 1; K. E. Lukong, K. W. Chang, E. W. Khandjian, S. Richard, Trends Genet 2008, 24, 416; F. B. Gao, J. P. Taylor, Brain Res 2012, 1462, 1; L. De Conti, M. Baralle, E. Buratti, Wiley Interdiscip Rev RNA 2017, 8; C. F. Sephton, G. Yu, Cell Mol Life Sci 2015, 72, 3621.

[6] H. Iioka, D. Loiselle, T. A. Haystead, I. G. Macara, Nucleic Acids Research 2011, 39, e53.

[7] C. Zhang, R. B. Darnell, Nat Biotechnol 2011, 29, 607; J. König, K. Zarnack, N. M. Luscombe, J. Ule, Nature Reviews Genetics 2012, 13, 77.
[8] N. G. Tsvetanova, D. M. Klass, J. Salzman, P. O. Brown, PLoS ONE 2010, 5, e12671.
[9] T. Scherrer, N. Mittal, S. C. Janga, A. P. Gerber, PLoS One 2010, 5, e15499.
[10] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, Benedikt M. Beckmann, C. Strein, Norman E. Davey, David T. Humphreys, T. Preiss, Lars M. Steinmetz, J. Krijgsveld, Matthias W. Hentze, Cell 2012, 149, 1393.
[11] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, Nucleic Acids Res 2000, 28, 235; H. M. Berman, G. J. Kleywegt, H. Nakamura, J. L. Markley, Structure 2012, 20, 391.
[12] C. A. McHugh, M. Guttman, in *RNA Detection: Methods and Protocols*, (Ed: I. Gaspar), Springer New York, New York, NY 2018, 473.
[13] S. Gerstberger, M. Hafner, T. Tuschl, Nat Rev Genet 2014, 15, 829.
[14] A. Andreeva, D. Howorth, J. M. Chandonia, S. E. Brenner, T. J. Hubbard, C. Chothia, A. G. Murzin, Nucleic Acids Res 2008, 36, D419.
[15] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, M. Punta, Nucleic Acids Res 2014, 42, D222.
[16] P. Ghosh, R. Sowdhamini, Mol Biosyst 2016, 12, 532.
[17] H. Zhao, Y. Yang, Y. Zhou, Mol Biosyst 2013, 9, 2417.
[18] O. Fornes, J. Garcia-Garcia, J. Bonet, B. Oliva, Adv Protein Chem Struct Biol 2014, 94, 77; R. R. Walia, C. Caragea, B. A. Lewis, F. Towfic, M. Terribilini, Y. El-Manzalawy, D. Dobbs, V. Honavar, BMC bioinformatics 2012, 13.
[19] J. Yan, S. Friedrich, L. Kurgan, Brief Bioinform 2016, 17, 88.
[20] Z. Miao, E. Westhof, PLoS Comput Biol 2015, 11, e1004639.
[21] T. Puton, L. Kozlowski, I. Tuszynska, K. Rother, J. M. Bujnicki, J Struct Biol 2012, 179, 261.
[22] H. Zhao, Y. Yang, Y. Zhou, Nucleic Acids Res 2011, 39, 3017.
[23] H. Zhao, Y. Yang, Y. Zhou, RNA Biol 2011, 8, 988.
[24] H. Zhao, Y. Yang, S. C. Janga, C. C. Kao, Y. Zhou, Proteins 2014, 82, 640.
[25] Y. Wang, Z. Xue, G. Shen, J. Xu, Amino Acids 2008, 35, 295; J. Tong, P. Jiang, Z. H. Lu, Comput Meth Prog Bio 2008, 90, 148; M. Kumar, A. M. Gromiha, G. P. S. Raghava, Proteins-Structure Function and Bioinformatics 2008, 71, 189; R. V. Spriggs, Y. Murakami, H. Nakamura, S. Jones, Bioinformatics 2009, 25, 1492; Y. Murakami, R. V. Spriggs, H. Nakamura, S. Jones, Nucleic Acids Res 2010, 38, W412; Y. F. Huang, L. Y. Chiu, C. C. Huang, C. K. Huang, Bmc Genomics 2010, 11; T. Zhang, H. Zhang, K. Chen, J. S. Ruan, S. Y. Shen, L. Kurgan, Curr Protein Pept Sc 2010, 11, 609; Z. P. Liu, L. Y. Wu, Y. Wang, X. S. Zhang, L. N. Chen, Bioinformatics 2010, 26, 1616; M. Fernandez, Y. Kumagai, D. M. Standley, A. Sarai, K. Mizuguchi, S. Ahmad, BMC Bioinformatics 2011, 12 Suppl 13, S5; C. C. Wang, Y. P. Fang, J. M. Xiao, M. L. Li, Amino Acids 2011, 40, 239; X. Ma, J. Guo, J. S. Wu, H. D. Liu, J. F. Yu, J. M. Xie, X. A. Sun, Proteins-Structure Function and Bioinformatics 2011, 79, 1230; R. R. Walia, L. C. Xue, K. Wilkins, Y. El-Manzalawy, D. Dobbs, V. Honavar, PloS one 2014, 9, e97725; M. Terribilini, J. H. Lee, C. H. Yan, R. L. Jernigan, V. Honavar, D. Dobbs, Rna 2006, 12, 1450; M. Terribilini, J. D. Sander, J. H. Lee, P. Zaback, R. L. Jernigan, V. Honavar, D. Dobbs, Nucleic Acids Res 2007, 35, W578; U. Muppirala, B. A. Lewis, C. M. Mann, D. Dobbs, Pac Symp Biocomput 2016, 21, 445.

[26] J. Yan, L. Kurgan, Nucleic Acids Res 2017, 45, e84.
[27] Z. Peng, C. Wang, V. N. Uversky, L. Kurgan, Methods Mol Biol 2017, 1484, 187.
[28] Z. Peng, L. Kurgan, Nucleic Acids Res 2015, 43, e121.
[29] Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano, B. Honig, Nature 2012, 490, 556; M. J. Mizianty, X. Fan, J. Yan, E. Chalmers, C. Woloschuk, A. Joachimiak, L. Kurgan, Acta Crystallogr D Biol Crystallogr 2014, 70, 2781.
[30] Z. Peng, J. Yan, X. Fan, M. J. Mizianty, B. Xue, K. Wang, G. Hu, V. N. Uversky, L. Kurgan, Cell Mol Life Sci 2015, 72, 137; Z. Wu, G. Hu, J. Yang, Z. Peng, V. N. Uversky, L. Kurgan, FEBS Lett 2015, 589, 2561; C. Wang, V. N. Uversky, L. Kurgan, Proteomics 2016, 16, 1486; Z. Peng, B. Xue, L. Kurgan, V. N. Uversky, Cell Death Differ 2013, 20, 1257.
[31] S. Basu, R. P. Bahadur, Cell Mol Life Sci 2016, 73, 4075; H. J. Dyson, Mol Biosyst 2012, 8, 97.
[32] J. Si, J. Cui, J. Cheng, R. Wu, Int J Mol Sci 2015, 16, 26303; T. Zhang, H. Zhang, K. Chen, J. Ruan, S. Shen, L. Kurgan, Curr Protein Pept Sci 2010, 11, 609.
[33] J. Zhang, Z. Ma, L. Kurgan, Brief Bioinform 2017.
[34] F. Meng, V. N. Uversky, L. Kurgan, Cell Mol Life Sci 2017, 74, 3069; P. Lieutaud, F. Ferron, A. V. Uversky, L. Kurgan, V. N. Uversky, S. Longhi, Intrinsically Disordered Proteins 2016, 4, e1259708.
[35] D. Piovesan, F. Tabaro, I. Micetic, M. Necci, F. Quaglia, C. J. Oldfield, M. C. Aspromonte, N. E. Davey, R. Davidovic, Z. Dosztanyi, A. Elofsson, A. Gasparini, A. Hatos, A. V. Kajava, L. Kalmar, E. Leonardi, T. Lazar, S. Macedo-Ribeiro, M. Macossay-Castillo, A. Meszaros, G. Minervini, N. Murvai, J. Pujols, D. B. Roche, E. Salladini, E. Schad, A. Schramm, B. Szabo, A. Tantos, F. Tonello, K. D. Tsirigos, N. Veljkovic, S. Ventura, W. Vranken, P. Warholm, V. N. Uversky, A. K. Dunker, S. Longhi, P. Tompa, S. C. Tosatto, Nucleic Acids Res 2016, D1, D219; M. Sickmeier, J. A. Hamilton, T. LeGall, V. Vacic, M. S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. N. Uversky, Z. Obradovic, A. K. Dunker, Nucleic Acids Res 2007, 35, D786.
[36] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi, L. S. Yeh, Nucleic Acids Res 2004, 32, D115; C. UniProt, Nucleic Acids Res 2015, 43, D204.
[37] J. Yang, A. Roy, Y. Zhang, Nucleic Acids Res 2013, 41, D1096.
[38] J. J. Ellis, M. Broom, S. Jones, Proteins 2007, 66, 903; S. L. Li, K. Yamashita, K. M. Amada, D. M. Standley, Nucleic Acids Res 2014, 42, 10086.
[39] K. Wang, R. Samudrala, BMC Bioinformatics 2006, 7, 385.
[40] M. Remmert, A. Biegert, A. Hauser, J. Soding, Nat Methods 2012, 9, 173.
[41] L. Micallef, P. Rodgers, PLoS One 2014, 9, e101717.
[42] L. P. Kozlowski, Nucleic Acids Res 2017, 45, D1112.
[43] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Nat Genet 2000, 25, 25.
[44] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, P. D'Eustachio, Nucleic Acids Res 2018, 46, D649.

[45] J. L. Moreland, A. Gramada, O. V. Buzko, Q. Zhang, P. E. Bourne, BMC Bioinformatics 2005, 6, 21.
[46] E. C. Koc, H. Cimen, B. Kumcuoglu, N. Abu, G. Akpinar, M. E. Haque, L. L. Spremulli, H. Koc, Front Physiol 2013, 4, 183.
[47] A. Amunts, A. Brown, J. Toots, S. H. W. Scheres, V. Ramakrishnan, Science 2015, 348, 95.
[48] P. S. Kaushal, M. R. Sharma, T. M. Booth, E. M. Haque, C. S. Tung, K. Y. Sanbonmatsu, L. L. Spremulli, R. K. Agrawal, Proc Natl Acad Sci U S A 2014, 111, 7284.
[49] Y. Xu, W. Yang, J. Wu, Y. Shi, Biochemistry 2002, 41, 5415.
[50] D. Piovesan, F. Tabaro, L. Paladin, M. Necci, I. Micetic, C. Camilloni, N. Davey, Z. Dosztanyi, B. Meszaros, A. M. Monzon, G. Parisi, E. Schad, P. Sormanni, P. Tompa, M. Vendruscolo, W. F. Vranken, S. C. E. Tosatto, Nucleic Acids Res 2018, 46, D471.
[51] W. Kabsch, C. Sander, Biopolymers 1983, 22, 2577.
[52] H. Yanai, T. Ban, Z. C. Wang, M. K. Choi, T. Kawamura, H. Negishi, M. Nakasato, Y. Lu, S. Hangai, R. Koshiba, D. Savitsky, L. Ronfani, S. Akira, M. E. Bianchi, K. Honda, T. Tamura, T. Kodama, T. Taniguchi, Nature 2009, 462, 99.
[53] A. Liepelt, I. S. Naarmann-de Vries, N. Simons, K. Eichelbaum, S. Fohr, S. K. Archer, A. Castello, B. Usadel, J. Krijgsveld, T. Preiss, G. Marx, M. W. Hentze, D. H. Ostareck, A. Ostareck-Lederer, Molecular & cellular proteomics : MCP 2016, 15, 2699.
[54] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, N. M. Luscombe, Nat Rev Genet 2009, 10, 252; H. Zhao, J. Wang, Y. Zhou, Y. Yang, PLoS One 2014, 9, e96694; B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, R. A. Young, Science 2000, 290, 2306.
[55] W. H. Hudson, E. A. Ortlund, Nat Rev Mol Cell Bio 2014, 15, 749.

Supporting Information

The supporting file S1.doc includes the set of 1511 putative RBPs along with the predicted RNA binding residues. For each protein we provide UniProt accession number, annotation whether this is an already experimentally validated RBP (Native) or a novel prediction (Novel), name of the predictor used to identify this protein, and protein sequence with annotated putative RNA binding residues.

The supporting file S2.doc includes the set of 7332 non-RNA binding human proteins that are used to estimate false positive rates. It also provides a list of false positive predictions from the DisoRDPbind, DRNApred and SPOT-Seq that overlap with these proteins. We provide UniProt accession number and sequence for all proteins.