

RESEARCH ARTICLE

Open Access

# Sequence search and analysis of gene products containing RNA recognition motifs in the human genome

Sony Malhotra and Ramanathan Sowdhamini\*

## Abstract

**Background:** Gene expression is tightly regulated at both transcriptional and post-transcriptional levels. RNA-binding proteins are involved in post-transcriptional gene regulation events. They are involved in a variety of functions such as splicing, alternative splicing, nuclear import and export of mRNA, RNA stability and translation. There are several well-characterized RNA-binding motifs present in a whole genome, such as RNA recognition motif (RRM), KH domain, zinc-fingers *etc.* In the present study, we have investigated human genome for the presence of RRM-containing gene products starting from RRM domains in the Pfam (Protein family database) repository.

**Results:** In Pfam, seven families are recorded to contain RRM-containing proteins. We studied these families for their taxonomic representation, sequence features (identity, length, phylogeny) and structural properties (mapping conservation on the structures). We then examined the presence of RRM-containing gene products in *Homo sapiens* genome and identified 928 RRM-containing gene products. These were studied for their predicted domain architectures, biological processes, involvement in pathways, disease relevance and disorder content. RRM domains were observed to occur multiple times in a single polypeptide. However, there are 56 other co-existing domains involved in different regulatory functions. Further, functional enrichment analysis revealed that RRM-containing gene products are mainly involved in biological functions such as mRNA splicing and its regulation.

**Conclusions:** Our sequence analysis identified RRM-containing gene products in the human genome and provides insights into their domain architectures and biological functions. Since mRNA splicing and gene regulation are important in the cellular machinery, this analysis provides an early overview of genes that carry out these functions.

**Keywords:** RNA recognition motif, *Homo sapiens*, Genome-wide survey, Domain architecture, Splicing

## Background

The gene expression process in eukaryotes needs to be tightly regulated at every step. Firstly, it is regulated at the transcription level by means such as chromatin structure, DNA sequence elements and binding of transcription factors *etc.* [1,2]. In spite of this tight regulation, post-transcriptional regulation plays an important role in regulating the levels of mRNA that are expressed in all tissues and serves as a supplement control mechanism. The post-transcriptional regulation governs several processes namely alternative splicing, RNA editing, transport of RNA from nucleus to cytoplasm, RNA stability and

translation [1,3]. The aberrations in the regulation of gene expression are also implicated in several human diseases such as Huntington's disease, leukoencephalopathy, cancer *etc.* [4-6].

RNA-binding proteins (RBPs) mediate all the post-transcriptional control events. As there are varied levels of control and targets to be regulated, there exists a wide repertoire of RNA-binding motifs. To achieve the sequence-specific recognition of targets, there are several RNA-binding domains that are well-characterized in RBPs such as RRM (RNA recognition motif) domains, KH domains, pumilio homology domain, zinc fingers, double-stranded RNA binding motifs (dsRBMs) [7,8]. RRM is the most abundant RNA-binding domain in higher vertebrates and is also known as RNA binding domain (RBD) or ribonucleoprotein domain (RNP) [9].

\* Correspondence: mini@ncbs.res.in  
National Centre for Biological Sciences (TIFR), GKVK Campus, Bellary Road,  
Bangalore 560 065, India

RRM is ~80-90 amino acids in length and contain two conserved motifs, RNP1 and RNP2, which are rich in aromatic amino acids. RRM structure possesses  $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$  topology, containing a four-stranded  $\beta$ -sheet which is packed against two  $\alpha$ -helices [7,9]. RNP1 is eight amino acids long and is present on  $\beta_3$  and RNP2 (six amino acids) is present on  $\beta_1$  [9]. Recently, RRM domains are also reported to be associated with the RNA-binding prion candidate proteins [10]. Birney *et al.*, performed an analysis on 125 sequences (possessing 252 RRM) of splicing factors and reported three solvent-exposed aromatic conserved residues in RNP-1 and RNP-2, which are implicated in RNA-binding [11].

Detailed study of RRM domains and their functions in the available sequenced genomes will help to improve our understanding and functions of RBPs. RBPs can be identified by the identification of RNA-binding domains in a given genome of interest. There have been several attempts to perform genome-wide analysis for specific RBPs in various organisms such as in *Drosophila melanogaster*, *Mus musculus*, *Arabidopsis thaliana*, sponge *Amphimedon queenslandica*, *C.elegans*, and yeast genomes [12-16]. This has led to identification of several RBPs in these genomes and 5-8% of genes are reported to encode RBP in yeast and ~2% in *C. elegans*, *D. melanogaster* and mouse. These studies provide insight into the distribution of RBPs and their classes in the genome being examined and their underlying functions.

In the present study, we have performed sequence searches in the human genome. We first studied the RRM-containing protein families in the Pfam database [17-20] for their taxonomic distribution, sequence features (sequence identity, phylogeny) and mapped conserved residues on their structures. We employed the profiles built using the members of these families to perform searches in the *Homo sapiens* genome. The gene products that retain sequence signatures of RRM-domains were next studied for their domain architectures, biological processes, pathways and disease relevance.

## Results and discussion

### RRM families in Pfam

There are seven families in the protein sequence family database (Pfam), which possess RRM domains (Table 1). We studied these families for their sequence features, taxonomic distributions and structural features (Figure 1).

RRM domains are known to be ~80-100 amino acids in their length [9] and Figure 2 shows the length distribution of the members for the seven RRM families in Pfam database. The proteins possessing RRM domains are present mainly in Eukaryota (Metazoa, Viridiplantae, Fungi kingdoms). But there is very little representation of RRM\_1 domain in Bacteria (Additional file 1).

#### (i) Sequence features

We examined the sequence dispersion of members within and across the seven families for the distribution of sequence identities among and across family members and domain lengths.

RRM domains from the members of the same families were observed to be very diverse at the sequence level (Figure 3). The average sequence identity (Figure 3A) of the seven RRM families was <40% for four families (RRM\_1, RRM\_3, RRM\_5 and RRM\_6). The distribution of sequence identities between members of the same family is plotted in form of a Box-Whisker plot (Figure 3B). Across the different RRM families, as expected, the average percentage identity was <20%. However, across families of RRM\_1, RRM\_5 and RRM\_6, some members share high sequence identity, as shown in Additional file 2.

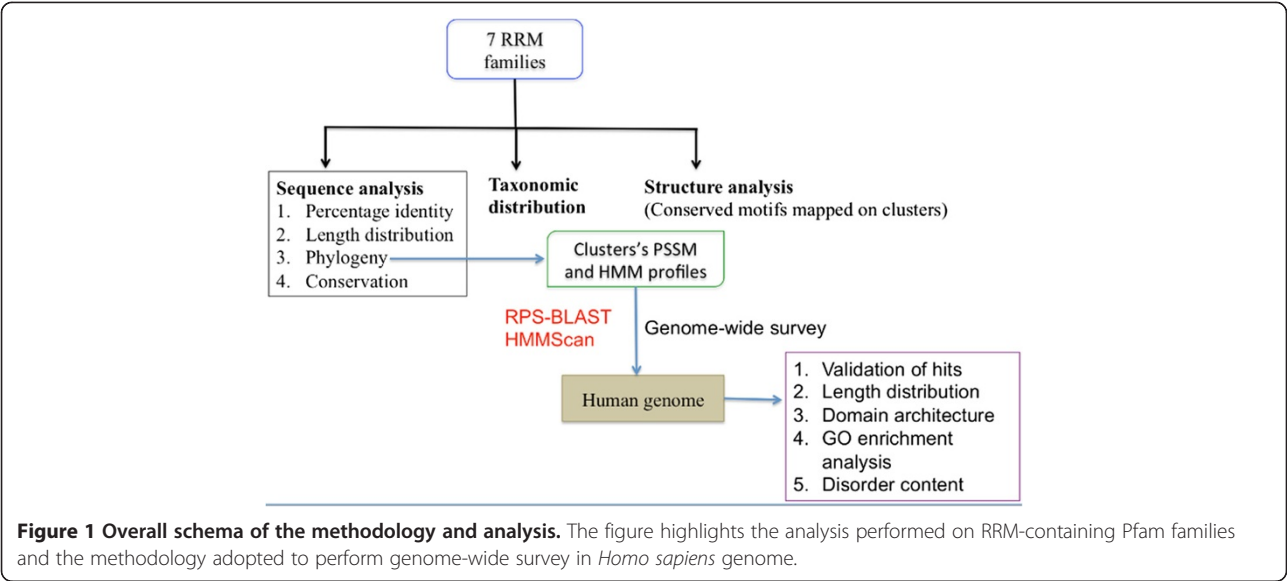
The Pfam seed sequences from all the seven RRM families were employed (full set if seed set contains <50 sequences, Table 1) to generate phylogeny using neighbor-joining method using ClustalW [21,22], Additional file 3A) and maximum-likelihood (PhyML [23], Additional file 3B) method. Sequences from families RRM\_1, RRM\_5 and RRM\_6 were observed to co-cluster, consistent with the observation of distribution of sequence identities (please

**Table 1 RRM families in Pfam database**

Pfam ID	Pfam name	Description	Number of Pfam seed sequences
PF00076	RRM_1	Clan RRM, has splicing factor and GO annotation as nucleic acid binding. Well-characterised	79
PF04059	RRM_2	Clan RRM, Meiosis protein mei2	4* (310)
PF08777	RRM_3	Clan RRM, La protein (RNA chaperone), 5 stranded Beta sheet (atypical RRM)	15* (185)
PF10598	RRM_4	PrP8 protein (Large RNA protein complex of spliceosome)	25* (233)
PF13893	RRM_5	Clan RRM, hnRNP	107
PF14259	RRM_6	Clan RRM	79
PF10378	RRM	Found only in fungi, putative RNA binding domain	7* (74)

There are seven families defined in Pfam based on their gathering threshold values. The families are listed with their description and the number of seed sequences in Pfam.

\*The numbers in brackets represent the full set members when the seed sequences are less than 50 in number.



see above). We refined the alignment of seed sequences using MUSCLE 3.8 [24] and built a new neighbor-joining tree using MEGA 6 [25] (Additional file 3C) using 500 bootstraps. The co-clustering of the members belonging to the three Pfam-defined families (RRM\_1, RRM\_5 and RRM\_6) was still persistent. Therefore, we defined new distinct clusters derived from the phylogeny (Additional file 3C, inner circle). The sequences of these clusters were re-aligned using MUSCLE 3.8 [24] and their PSSM and HMM profiles were created (please see Methods for details) to perform the searches in the human genome.

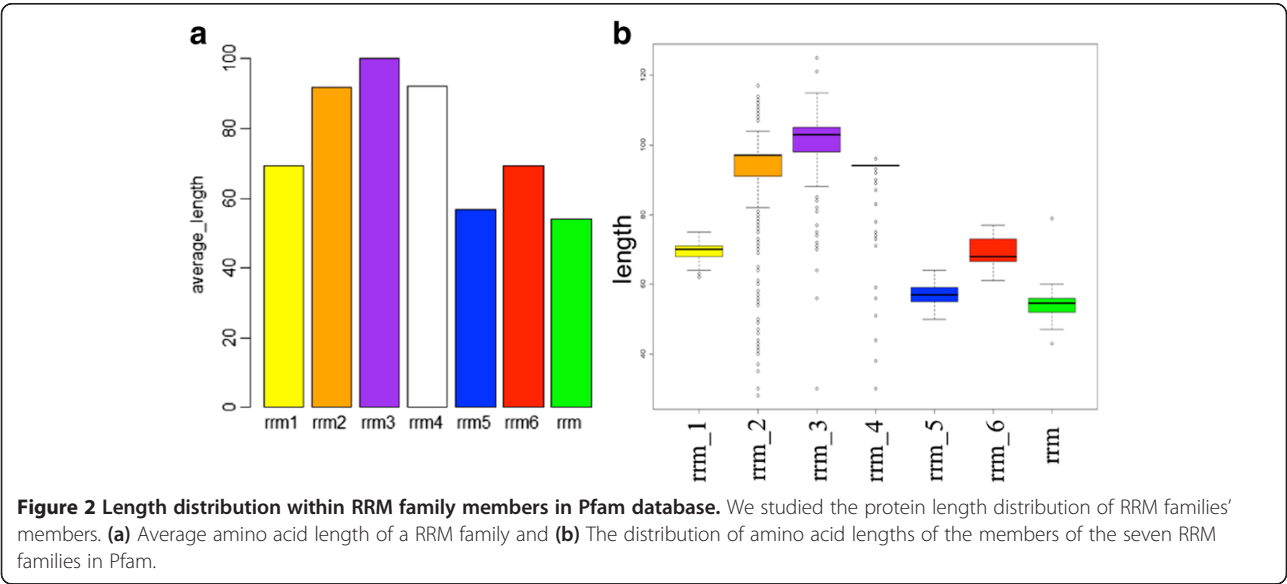
*(ii) Structural features*

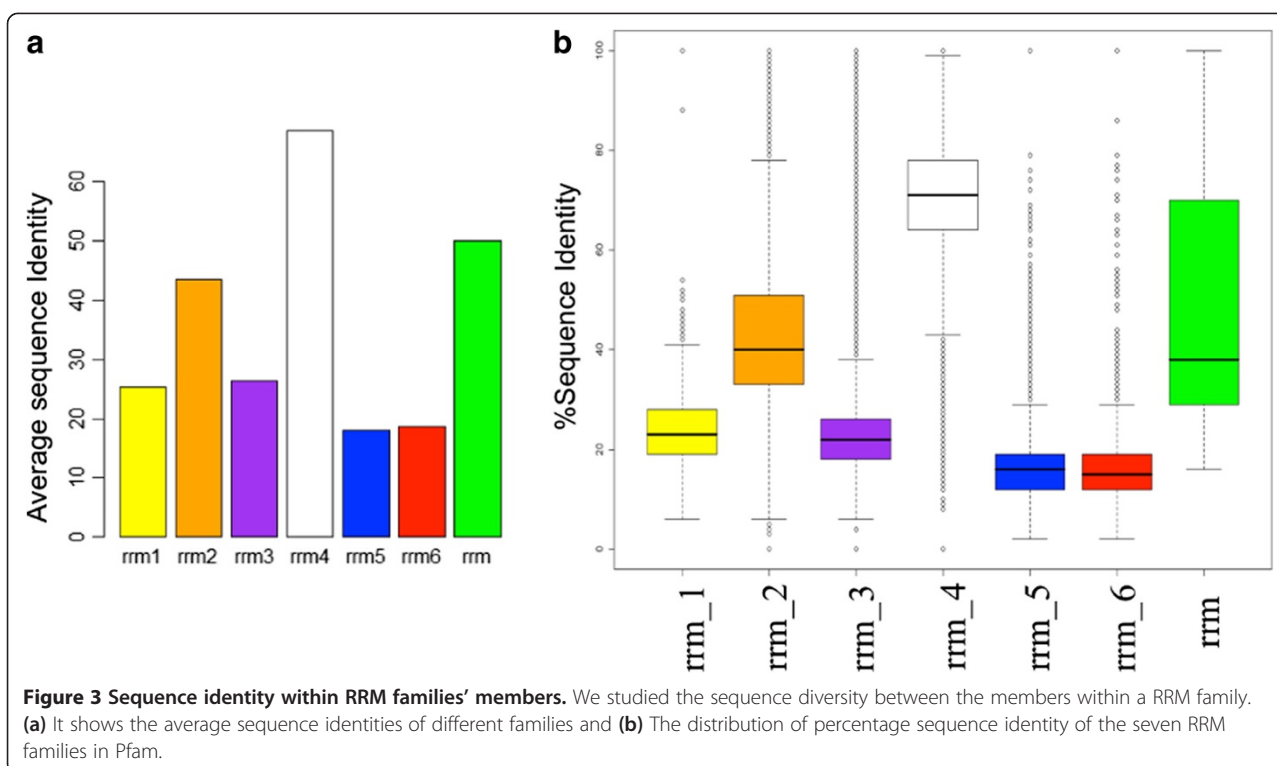
Four of the seven families have structural representation in the Protein Data Bank (PDB) [26]. The alignments for

the seven families were analyzed for the conservation of residues using ConSurf [27]. The conserved residues were mapped on the structures from each of the Pfam RRM family (Figure 4). We observed that the conserved residues map to the same structural region. This observation, together with the percentage similarity plots (across families, Additional file 2) explains the cause of co-clustering observed between the members of the different RRM families. Therefore, as explained above, we made distinct clusters based on sequence identities to perform searches in the human genome.

**Searches in the human genome and validation**

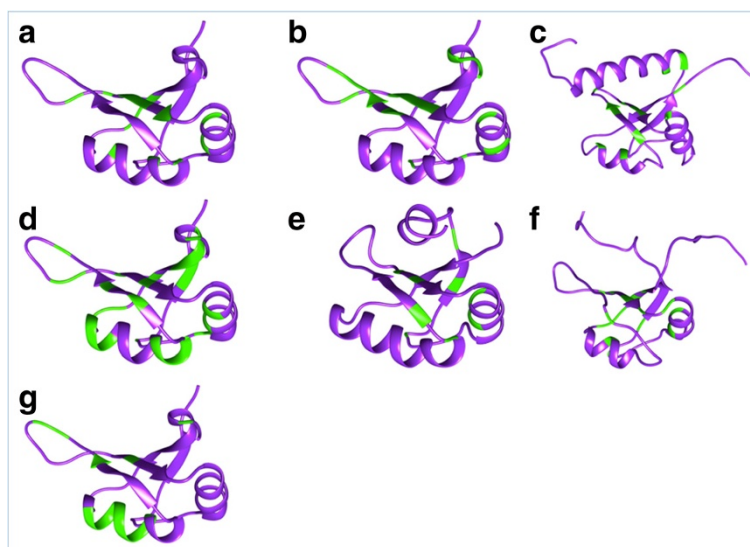
Both PSSM and HMM profiles of the new clusters (please see Methods for details) were employed to





search the human genome for the presence of RRM-containing gene products using RPS-BLAST [28,29] and HMMscan [30] respectively. 928 RRM-containing gene products were thus identified in the human genome purely by sequence searches (Additional file 4). Of these, 50% (452 gene products) are unreviewed proteins and

belong to UniProt/TrEMBL. Subsequent to the clustering using BLASTCLUST [31,32] (at 98% sequence identity), 403 human gene products were retained. 84% of these (340 gene products) are annotated in Gene Ontology (GO) database [33] for their molecular functions. Out of these, 337 gene products were annotated as RNA-binding



**Figure 4 Conservation mapping on RRM structures.** The multiple sequence alignments for each of these families were employed to study sequence conservation and the conserved residues were mapped to the RRM structures (in green). However for the families RRM\_2, RRM\_4 and family-RRM, there is no structural representation and we therefore, used structure of RRM\_1 family for mapping. (a) RRM\_1 (1L3K, chainA), (b) RRM\_2, (c) RRM\_3 (1OWX, chainA), (d) RRM\_4, (e) RRM\_5 (1A9N, chain d), (f) RRM\_6 (1WG5, chain a) and (g) RRM.

(and child terms) and/or nucleotide binding (and child terms) in GO.

### Length distribution

The full-length distribution of the RRM-containing gene products identified by sequence searches was next analyzed. RRM is a small domain of ~80-100 amino acids; however, we observed that most of the RRM-containing human hits are >150 amino acids in length (Figure 5). This implies that there are other co-existing domains or multiple RRM domains or unstructured regions accompanying RRM domains in the full-length human gene products that contain RRM domains.

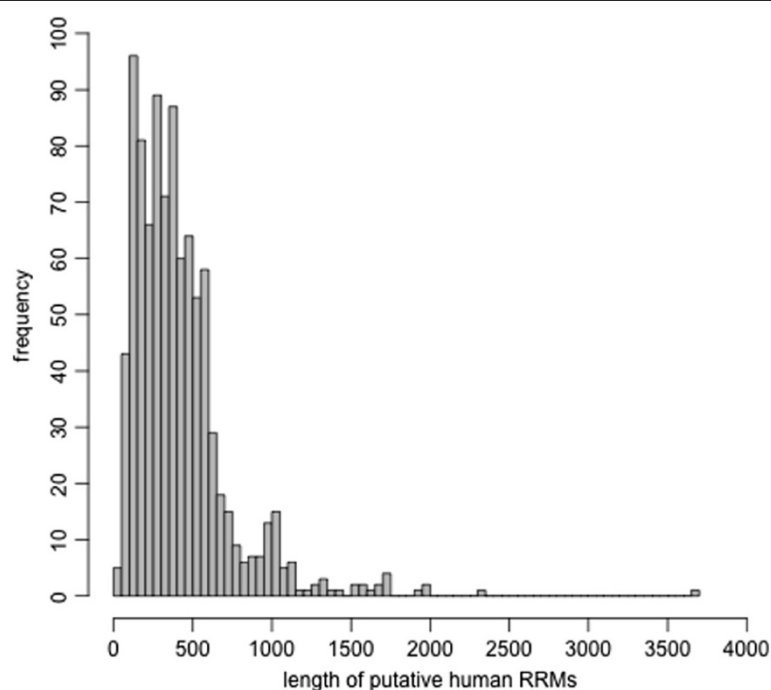
### Domain architectures (co-existing domains)

The full-length RRM-containing human gene products were also analyzed for their complete domain architectures using HMMScan [30] against HMM profiles of Pfam families. Such a search enabled the association of RRM domains identified in the human genome into any one of the known seven families in Pfam database as well. RRM\_1 (PF00076) is the most well-characterized and well-populated RRM family in the Pfam database. 79% of the identified human RRM-containing gene products possess RRM\_1 domain. In Pfam database, RRM\_1 is present in larger fraction of protein sequences of the class Mammalia as compared to other six RRM families (Additional file 1). Out of the seven RRM families in

Pfam, we observed that two families (RRM\_2 and RRM) have no representation in the human genome. There are no sequences from the class mammalia for these two families in the Pfam database also. RRM is a fungal-specific family and RRM\_2 is found only in Viridiplantae and fungi (Additional file 1).

We further analyzed the co-existing domains and observed that in 13 gene products (where isoforms are reported), the number of RRM occurrence and co-existing domains are different (Table 2). This implies that may be during the alternative splicing event full domains are also spliced out. Figure 6 highlights the ten most frequent domain architectures observed in RRM-containing human gene products in a schematic form drawn. 40% of the gene products contain no other recognizable co-existing domain, whereas rest of the sequences possesses either multiple RRM domains or other co-existing domains. The gene products containing a single domain connect to RRM\_1/RRM\_5/RRM\_families and map to their biological processes using GO annotations. 39 of these gene products are annotated with their biological process (Additional file 5).

Genes containing multiple RRM domains were present in 28% of the RRM-containing human hits (Figure 7) and this event is a frequent occurrence almost as a rule [9,34]. There are 56 non-RRM co-existing domains (Additional file 6) which were noted for their functions and are observed to be involved in a variety of cellular



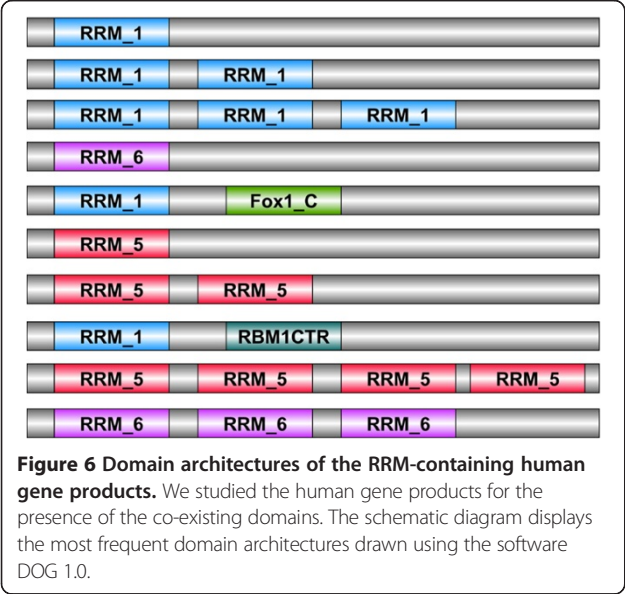
**Figure 5 Length distribution of the RRM-containing human gene products.** We studied the protein length distribution of the human gene products that were identified as RRM-containing.



**Table 2 Domain architectures in isoforms**

Domain architecture	Human protein	Domain architecture	Human protein
RRM_1,RRM_1,RRM_1	O60506	PWI, RRM_5,RRM_5	Q5T8P6
RRM_1,RRM_1	O60506-2	PWI, RRM_5,RRM_5	Q5T8P6-2
RRM_1,RRM_1,RRM_1	O60506-3	PWI, RRM_5,RRM_5	Q5T8P6-3
RRM_1,RRM_1	O60506-4	RRM_5,RRM_5	Q5T8P6-4
RRM_1,RRM_1,RRM_1	O60506-5	RRM_5,RRM_5	Q5T8P6-5
La, RRM_1,RRM_3	Q4G0J3	RRM_1,RRM_1	Q865G3
RRM_3	Q4G0J3-2	RRM_1	Q865G3-2
RRM_5	O95628	RRM_1,RRM_1,RRM_1,RRM_1	Q8IUH3
RRM_5	O95628-2	RRM_1,RRM_1,RRM_1	Q8IUH3-2
RRM_5	O95628-3	RRM_1,RRM_1,RRM_1,RRM_1	Q8IUH3-3
zf-C3HC4_3,RRM_5	O95628-4	RRM_1,RRM_1,RRM_1	Q8N6W0
RRM_5	O95628-5	RRM_1,RRM_1	Q8N6W0-2
RRM_5	O95628-6	zf-CCCH,zf-CCCH	Q8WU68
zf-C3HC4_3,RRM_5	O95628-7	zf-CCCH	Q8WU68-2
zf-C3HC4_3,RRM_5	O95628-8	zf-CCCH,zf-CCCH	Q8WU68-3
RRM_6,RRM_6	P31942	RRM_5,RRM_5,RRM_5	Q8WV9
RRM_6,RRM_6	P31942-2	RRM_5,RRM_5	Q8WV9-2
RRM_6,RRM_6	P31942-3	RRM_5,RRM_5	Q8WV9-3
RRM_6	P31942-4	RRM_5,RRM_5,RRM_5	Q8WV9-4
RRM_6	P31942-5	RRM_1,RRM_1,RRM_1	Q96J87
RRM_6	P31942-6	RRM_1,RRM_1	Q96J87-2
RRM_1,RRM_1,RRM_1	Q9P2K5	RRM_1,RRM_1,RRM_1	Q96J87-3
RRM_1,RRM_1,RRM_1	Q9P2K5-2	RRM_1,RRM_1	Q96J87-4
RRM_1	Q9P2K5-3		

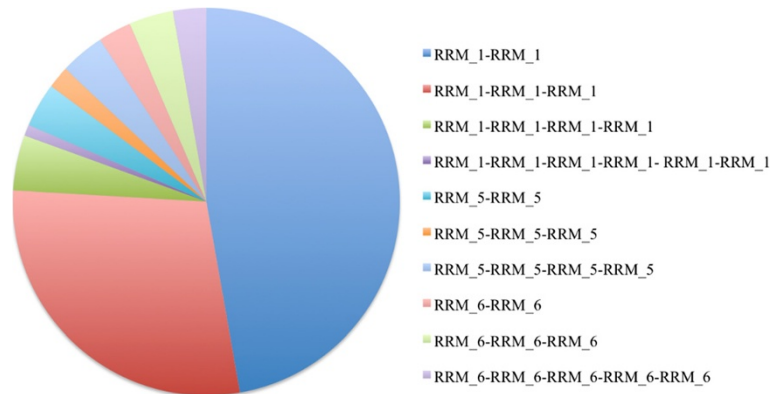
928 RRM-containing gene products were studied for their co-existing domains. We observed that in 13 gene products where isoforms are reported the domain architectures are different. This implies may be during the alternative splicing event full domains are also spliced out.



activities such as developmental signaling, apoptosis, transcriptional regulation, splicing and alternative splicing (Additional file 6). We also noted these co-existing domains for their frequency of occurrence in the human gene products (Additional file 7). As observed by the functions of co-existing domains, RRM containing gene products possess other RNA as well as protein binding domains, which might govern their specificity and affinity towards their RNA targets and assist them in performing their diverse biological functions.

**Enrichment analysis**

We studied the RRM domain-containing human gene products for their involvement in biological processes using DAVID 6.7 [35,36]. They were observed to be involved in various processes involved in RNA metabolism. We further identified the biological processes, which were enriched in the RRM-containing gene products based on normalization using the biological processes performed by all the human gene products as background. Out of the



**Figure 7 Modular nature of RRM.** Within a single polypeptide sequence, RRM domain was observed to occur multiple times in 108 of the human gene products. This figure highlights the distribution of different RRM domains and their multiple occurrences. RRM\_1 is present most frequently (88 of the gene products) and is repeated twice within a single protein sequence.

set of 403 gene products, 173 are annotated with GO biological processes in DAVID 6.7. Upon functional clustering, these belong to 22 clusters. Upon filtering the results based on Bonferroni correction method ( $p < 0.05$ ), 42 gene products were observed to belong to six biological processes performing mRNA processing and RNA splicing (Figure 8, Additional file 8).

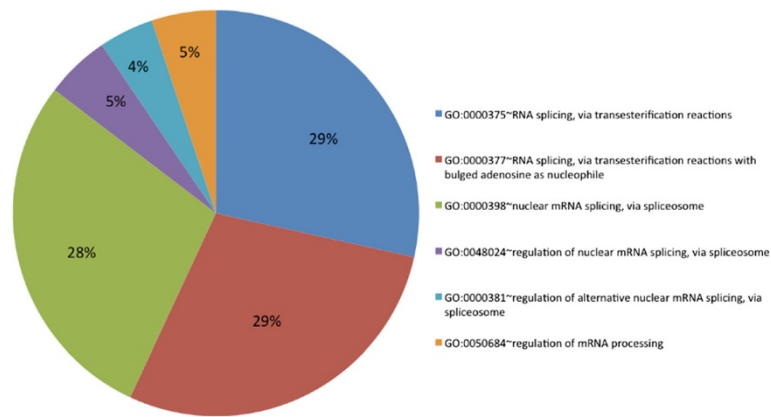
Using DAVID 6.7, we also studied the KEGG pathway [37] enrichment in this set of human RRM-containing gene products. Upon performing functional clustering and using the same filtering parameters as explained above, 33 gene products were observed to be part of the spliceosome machinery (Additional file 9).

**Disease involvement and disorder content**

The set of RRM-containing human gene products were further analyzed for their role in diseases and disorder content. We obtained a comprehensive list of RNA-binding

proteins which are linked to Mendelian diseases in human (as recorded in OMIM database) from a recent review [38]. There are 157 RNA binding Ensembl gene models that are implicated in Mendelian diseases [38]. We mapped these to the RRM-containing gene products identified in our analysis. 14 of these RNA-binding proteins linked with Mendelian diseases contain RRM domain (Additional file 10).

As ageing is reported as a risk factor for neurodegeneration and the role of RNA-binding proteins is implicated in neurodegeneration [39-41], we analyzed the disorder content of the RRM-containing gene products. Their disorder content of was analyzed using DISOPRED [42]. 16% of the gene products (Additional file 11) were high ( $>0.7$ ) in their disorder content (% of disordered residues) and such gene products could be involved in processes such as ageing [39-41].



**Figure 8 Enrichment analysis for biological processes.** The RRM-containing human gene products were studied for their functional clustering based enrichment analysis using DAVID 6.7. The processes, which were observed to be enriched were related to mRNA processing, splicing and its regulation.

## Conclusions

RNA-binding proteins govern gene regulatory events at the post-transcriptional level. There are several well-characterized RNA-binding motifs present in the protein partner. Of these, RRM are the most abundant in higher vertebrates. In the present work, the genome-wide survey for the presence of RRM-containing gene products was performed in the human proteome, employing computational approaches starting from the known RRM-containing sequences present in the Pfam database.

The seven RRM families in Pfam are derived based on the HMM-HMM comparisons using a gathering threshold (GA threshold). GA thresholds are Pfam-bit scores and estimates of significance of hits. We studied these families for their features. We observed:

1. Taxonomic representation: The majority of the sequences belonging to these families are present in Eukaryota, with few bacterial RRM-containing proteins in the family RRM\_1. RRM\_2 and RRM Pfam families are not present in the class Mammalia and are present only in plants and fungi.
2. Sequence features: The members within the families are more similar as compared to other family members, as expected. However, some of the members of RRM\_1, RRM\_5 and RRM\_6 families share high (>50%) sequence identity.
3. The conservation of amino acid residues was studied using ConSurf and mapped on the protein structures from each of the RRM families. The conserved residues were localized on similar structural regions.

We identified 928 gene products (403 gene products at 98% sequence identity), which contain RRM domain in the human genome upon performing the genome-wide scan using profile-based sequence search methods. As documented in existing literature, RRM is an abundant domain in eukaryotes [7,15,43,44] and we also observed that 50% of the reported RNA-binding proteins (860 RNA-binding proteins, experimentally characterized by isolating mRNA interactome) in the human genome from a recent study [45] possess RRM domain. Their full-length sequences were analyzed for domain architectures in order to understand their functional roles. As RNA-binding proteins are known to mediate variety of different interactions and regulatory functions, analyzing the domain architectures of these full-length gene products will provide an insight into understanding of their evolution and biological functions. RRM\_1 domain is present in majority of these human gene products (79%). 60% of the gene products were observed to possess multiple domains (either multiple RRM or non-RRM co-existing domains). RRM-containing proteins are known to possess modular nature (multiple repeats of RRM) [9,34].

The length of the linker between the different RRM domains is known to govern specificity of RNA-binding, since a single RRM domain can bind from only four to eight nucleotides [9]. Therefore, modular nature of these proteins confer specificity to bind the target RNA as the number of nucleotides identified by single RRM domain is too small to define a unique target. The non-RRM co-existing domains were observed to be involved in functions like developmental signaling, apoptosis, transcriptional regulation, splicing and alternative splicing suggesting fundamental cellular roles of such genes.

The set of RRM-containing human gene products were mapped for their biological processes and pathways. The biological processes which were enriched in these gene products were related to mRNA splicing and its regulation. 33 of gene products were involved in the spliceosomal pathway. There are more than 100 gene products that are known to be part of the spliceosome. It is recently reported that more than half of the proteins in the spliceosome are intrinsically disordered (when proteins are considered in isolation) [46,47]. These intrinsically disordered proteins are also implicated in age related neurodegenerative diseases [39-41]. One of the RRM-containing human gene product that encodes for FUS protein (fused in sarcoma, 546 amino acids long), is predicted to possess only 50 amino acids, which form a folded structure [39]. We calculated the fraction of disorder residues (low sequence complexity, rich in hydrophilic and aromatic residues) in the human RRM-containing gene products and observed that 16% of these are rich in disordered regions. It will be interesting to follow if these intrinsically disordered regions become structured upon binding to their RNA targets.

Also, recently RRM domain proteins have been implicated in several Mendelian diseases [38] and are observed to possess prion-forming ability [10,48]. Therefore, we mapped these gene products to OMIM database and 14 of the human gene products were linked to Mendelian diseases. This study will help in characterization of RRM-containing gene products in the human genome and to provide early bioinformatics view of their functions.

## Methods

### Human proteome

The entire proteome of *Homo sapiens*, comprising of both reviewed and unreviewed entries, was downloaded from UniProt FTP website (<http://www.uniprot.org/downloads>). This set of human sequences was used to perform the genome-wide survey.

### RRM families

To perform searches in the human genome, we collated the known RRM sequences from the protein family database [17-20] (Pfam). Pfam clusters sequences on the



basis of their sequence similarities (HMM based) into seven different families. We studied these families for their sequence-based features like sequence identities and length distributions. The family alignments were employed to identify conserved sequence motifs using ConSurf [27] and these motifs were mapped on the structures of the RRM families. The sequences belonging to these families were also studied for their taxonomic distributions in various kingdoms and classes.

### Search protocol and its validation

The RRM-containing sequences belonging to the seven PFam families were used to perform searches in the human genome. Based on the phylogenetic tree analysis, these families were clustered into 10 new distinct clusters. The profiles of the new clusters were employed to perform searches in the human genome.

The multiple sequence alignment for all the members for each of the clusters was performed using MUSCLE 3.8 [24]. All the phylogenetic trees were visualized using FigTree 1.4.0 [49]. Subsequently, the human genome was searched using sensitive profile-based sequence search methods, RPS-BLAST [28,29] and HMMScan [30].

1. We built PSI-BLAST profiles (position-specific substitution matrix) for each cluster, using the alignment of cluster members as an input against NR database at an Eval =  $10^{-10}$ . A database of profiles of all the clusters was generated. The human gene products were searched against this database of profiles using RPS-BLAST at Eval =  $10^{-3}$ .
2. For each of the cluster, we also generated HMM profiles using the alignment as an input. The entire human proteome was also searched against the HMM profiles of all clusters to identify putative RRM-containing gene products using HMMScan and an Eval =  $10^{-2}$ .

The RRM-containing gene products identified in the human genome were further subjected to validation using GO annotations available for the human proteome [33]. We filtered them based on GO terms, RNA binding and nucleotide binding including their child terms.

### Analysis of RBP identified in human genome

These full-length gene products that contain sequence signature for at least one RRM domain were analyzed for their length distributions.

1. Domain architectures  
The identified gene products were further filtered to remove isoforms, fragments and highly similar sequences, by clustering them using BLASTCLUST [31,32] at 98% sequence similarity over an area

covering 50% of the length. We then studied the domain architectures using HMMScan against a database of entire Pfam HMM profiles at  $10^{-5}$ . The domain architectures were observed for RRM-repeats and non-RRM co-existing domains were noted for their functions. The schematic for domain architectures was drawn using the software DOG 1.0 [50].

2. Biological processes and pathways  
The identified RRM-containing human gene products were also mapped to their biological processes and the enrichment study for these processes was performed using DAVID 6.7 [35,36]. Upon functional clustering, the results were filtered based on Bonferroni correction ( $p < 0.05$ ). We also studied these gene products for their pathway mapping in the KEGG database [37] using DAVID 6.7 [35,36].
3. Disease implications and disorder content  
The gene products were further mapped to OMIM [51] database using DAVID 6.7 [35,36]. The disorder analysis was performed using DISOPRED [42]. All the residues were analyzed for their disorder and the disorder content (fraction of disordered residues) for these gene products was calculated.

### Additional files

**Additional file 1:** Is a table listing the taxonomic representation of RRM families.

**Additional file 2:** Is a figure, which highlights the percent sequence identity across different RRM families (In the Additional file 2, r1 stands for RRM\_1, r2 for RRM\_2, r3 for RRM\_3, r4 for RRM\_4, r5 for RRM\_5, r6 for RRM\_6 and r for RRM family. R1\_r2 implies percent identity distribution between the members of RRM\_1 and RRM\_2 families and likewise for other combinations).

**Additional file 3:** Is a figure that shows the co-clustering between members belonging to different RRM families (A. Neighbor joining tree-using ClustalW, B. Maximum-likelihood tree using PhyML and C. Neighbor joining tree using MEGA 6 and employing alignment derived from MUSCLE 3.8). The color code followed is: RRM\_1: Blue, RRM\_2: Brown, RRM\_3: Red, RRM\_4: Pink, RRM\_5: Yellow, RRM\_6: Green and RRM: Cyan.

**Additional file 4:** Is a table listing the RRM-containing gene products identified in the human genome.

**Additional file 5:** Is a table listing the single domain RRM-containing human gene products and their molecular functions.

**Additional file 6:** Is a table listing the non-RRM co-existing domains with their functions present in the set of human RRM-containing gene products.

**Additional file 7:** Is a figure that shows the frequency of co-existing domains in the RRM-containing human gene products.

**Additional file 8:** Is a table listing the enriched GO functions in RRM-containing human gene products.

**Additional file 9:** Is a figure that highlights the gene products involved in spliceosome pathway (In red the spliceosome components that contain the gene products we identified upon genome-wide survey are marked). The figure displays all the components that are known to be part of spliceosome pathway (as in

KEGG). The gene products, which were identified using our search strategy, are marked with red stars.

**Additional file 10:** Is a table listing the 14 human RRM-containing gene products with implications in Mendelian diseases.

**Additional file 11:** Is a table listing the predicted disorder content in the RRM-containing gene products identified in the human genome.

# Competing interests

The authors declare that they have no competing interests.

# Authors' contributions

RS conceived the idea and designed the experiments. SM carried out all the work on scripting and analysis. RS and SM interpreted the results and improved strategies. SM wrote first draft of the manuscript and RS improved it. Both authors have read the manuscript.

# Acknowledgements

SM is supported by a fellowship from Department of Biotechnology. The authors thank NCBS (National Centre for Biological Sciences) for infrastructure and other facilities.

Received: 29 July 2014 Accepted: 12 December 2014

Published: 22 December 2014

# References

- Latchman D: *Gene Regulation*, Garland Science. ; 2007.
- Le Jeune E, Ladurner AG: *Analysing gene expression*, edited by S. Lorkowski and P. Cullen. *Protein Sci Publ Protein Soc* 2004, **13**:1950–1952.
- Jackson DA, Pombo A, Iborra F: *The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells*. *FASEB J Off Publ Fed Am Soc Exp Biol* 2000, **14**:242–254.
- Ambrose CM, Duyao MP, Barnes G, Bates GP, Lin CS, Srinidhi J, Baxendale S, Hummerich H, Lehrach H, Altherr M: *Structure and expression of the Huntington's disease gene: evidence against simple inactivation due to an expanded CAG repeat*. *Somat Cell Mol Genet* 1994, **20**:27–38.
- Aerts S, Cools J: *Cancer: Mutations close in on gene regulation*. *Nature* 2013, **499**:35–36.
- Madhamshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA: *Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets*. *Genome Med* 2012, **4**:41.
- Cléry A, Blatter M, Allain FH-T: *RNA recognition motifs: boring? Not quite*. *Curr Opin Struct Biol* 2008, **18**:290–298.
- Burd CG, Dreyfuss G: *Conserved structures and diversity of functions of RNA-binding proteins*. *Science* 1994, **265**:615–621.
- FROM STRUCTURE TO FUNCTION OF RNA BINDING DOMAINS. [http://www.ncbi.nlm.nih.gov/books/NBK63528/.
- King OD, Gitler AD, Shorter J: *The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease*. *Brain Res* 2012, **1462**:61–80.
- Birney E, Kumar S, Krainer AR: *Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors*. *Nucleic Acids Res* 1993, **21**:5803–5816.
- Gamberi C, Johnstone O, Lasko P: *Drosophila RNA binding proteins*. *Int Rev Cytol* 2006, **248**:43–139.
- Kerner P, Degnan SM, Marchand L, Degnan BM, Vervoort M: *Evolution of RNA-binding proteins in animals: insights from genome-wide analysis in the sponge Amphimedon queenslandica*. *Mol Biol Evol* 2011, **28**:2289–2303.
- Tamburino AM, Ryder SP, Walhout AJM: *A compendium of Caenorhabditis elegans RNA binding proteins predicts extensive regulation at multiple levels*. *G3 Bethesda Md* 2013, **3**:297–304.
- Lorković ZJ, Barta A: *Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA-binding proteins from the flowering plant Arabidopsis thaliana*. *Nucleic Acids Res* 2002, **30**:623–635.
- McKee AE, Minet E, Stern C, Riahi S, Stiles CD, Silver PA: *A genome-wide in situ hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain*. *BMC Dev Biol* 2005, **5**:14.
- Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL: *The Pfam protein families database*. *Nucleic Acids Res* 2000, **28**:263–266.
- Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, Studholme DJ, Yeats C, Eddy SR: *The Pfam protein families database*. *Nucleic Acids Res* 2004, **32**(suppl 1): D138–D141.
- Finn RD, Mistry J, Tate J, Coghill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A: *The Pfam protein families database*. *Nucleic Acids Res* 2009, **38**(Database):D211–D222.
- Punta M, Coghill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD: *The Pfam protein families database*. *Nucleic Acids Res* 2012, **40**:D290–D301.
- Thompson JD, Higgins DG, Gibson TJ: *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*. *Nucleic Acids Res* 1994, **22**:4673–4680.
- Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: *Clustal W and Clustal X version 2.0*. *Bioinformatics* 2007, **23**:2947–2948.
- Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O: *New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0*. *Syst Biol* 2010, **59**:307–321.
- Edgar RC: *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. *Nucleic Acids Res* 2004, **32**:1792–1797.
- Tamura K, Stecher G, Peterson D, Filipski A, Kumar S: *MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0*. *Mol Biol Evol* 2013, **30**:2725–2729.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: *The protein data bank*. *Nucleic Acids Res* 2000, **28**:235–242.
- ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. [http://nar.oxfordjournals.org/content/38/suppl\\_2/W529.short](http://nar.oxfordjournals.org/content/38/suppl_2/W529.short).
- Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH: *CDD: a database of conserved domain alignments with links to domain three-dimensional structure*. *Nucleic Acids Res* 2002, **30**:281–283.
- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. *Nucleic Acids Res* 1997, **25**:3389–3402.
- Eddy SR: *Accelerated profile HMM searches*. *PLoS Comput Biol* 2011, **7**:e1002195.
- BLASTclust. <http://toolkit.tuebingen.mpg.de/blastclust#>.
- NCBI News: Spring 2004|BLASTLab. NCBI News: Spring 2004|BLASTLab.
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: *Gene ontology: tool for the unification of biology*. The Gene Ontology Consortium. *Nat Genet* 2000, **25**:25–29.
- Shamoo Y, Abdul-Manan N, Williams KR: *Multiple RNA binding domains (RBDs) just don't add up*. *Nucleic Acids Res* 1995, **23**:725–728.
- Huang DW, Sherman BT, Lempicki RA: *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*. *Nat Protoc* 2009, **4**:44–57.
- Huang DW, Sherman BT, Lempicki RA: *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists*. *Nucleic Acids Res* 2009, **37**:1–13.
- Kanehisa M, Goto S: *KEGG: kyoto encyclopedia of genes and genomes*. *Nucleic Acids Res* 2000, **28**:27–30.
- Castello A, Fischer B, Hentze MW, Preiss T: *RNA-binding proteins in Mendelian disease*. *Trends Genet TIG* 2013, **29**:318–327.
- Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT: *The DISOPRED server for the prediction of protein disorder*. *Bioinformatics* 2004, **20**:2138–2139.
- Gray DA, Woulfe J: *Structural disorder and the loss of RNA homeostasis in aging and neurodegenerative disease*. *Front Genet* 2013, **4**:149.
- Vanderweyde T, Youmans K, Liu-Yesuiev L, Wolozin B: *Role of stress granules and RNA-binding proteins in neurodegeneration: a mini-review*. *Gerontology* 2013, **59**:524–533.

42. Wolozin B: **Regulated protein aggregation: stress granules and neurodegeneration.** *Mol Neurodegener* 2012, **7**:56.
43. Daubner GM, Cléry A, Allain FH-T: **RRM-RNA recognition: NMR or crystallography...and new findings.** *Curr Opin Struct Biol* 2013, **23**:100–108 [*Folding and Binding / Protein-Nucleic Acid Interactions*].
44. Maris C, Dominguez C, Allain FH-T: **The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression.** *FEBS J* 2005, **272**:2118–2131.
45. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, Krijgsvelde J, Hentze MW: **Insights into RNA biology from an atlas of mammalian mRNA-binding proteins.** *Cell* 2012, **149**:1393–1406.
46. Tompa P, Csermely P: **The role of structural disorder in the function of RNA and protein chaperones.** *FASEB J* 2004, **18**:1169–1175.
47. Korneta I, Bujnicki JM: **Intrinsic disorder in the human spliceosomal proteome.** *PLoS Comput Biol* 2012, **8**:e1002641.
48. Lukong KE, Chang K, Khandjian EW, Richard S: **RNA-binding proteins in human genetic disease.** *Trends Genet TIG* 2008, **24**:416–425.
49. **FigTree.** <http://tree.bio.ed.ac.uk/software/figtree/>.
50. Ren J, Wen L, Gao X, Jin C, Xue Y, Yao X: **DOG 1.0: illustrator of protein domain structures.** *Cell Res* 2009, **19**:271–273.
51. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005, **33**(suppl 1):D514–D517.

doi:10.1186/1471-2164-15-1159

**Cite this article as:** Malhotra and Sowdhamini: Sequence search and analysis of gene products containing RNA recognition motifs in the human genome. *BMC Genomics* 2014 **15**:1159.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

