Supplementary Data

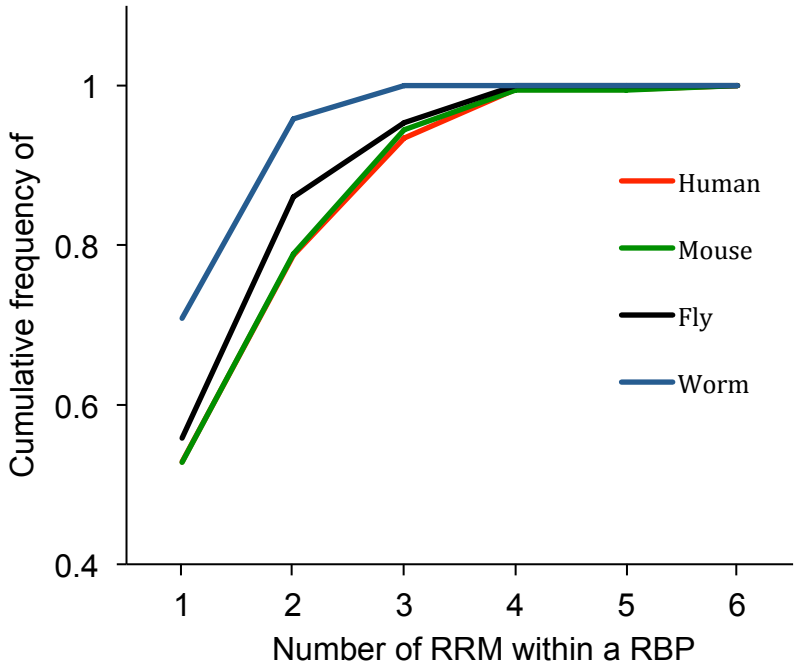


Figure S1. Increased numbers of RRMs within a single RBP in mammals. Cumulative frequency of RBP with different numbers of RRM was plotted in four species (*H. sapien*, *M. musculus*, *D. melanogaster* and *C. elegans*).

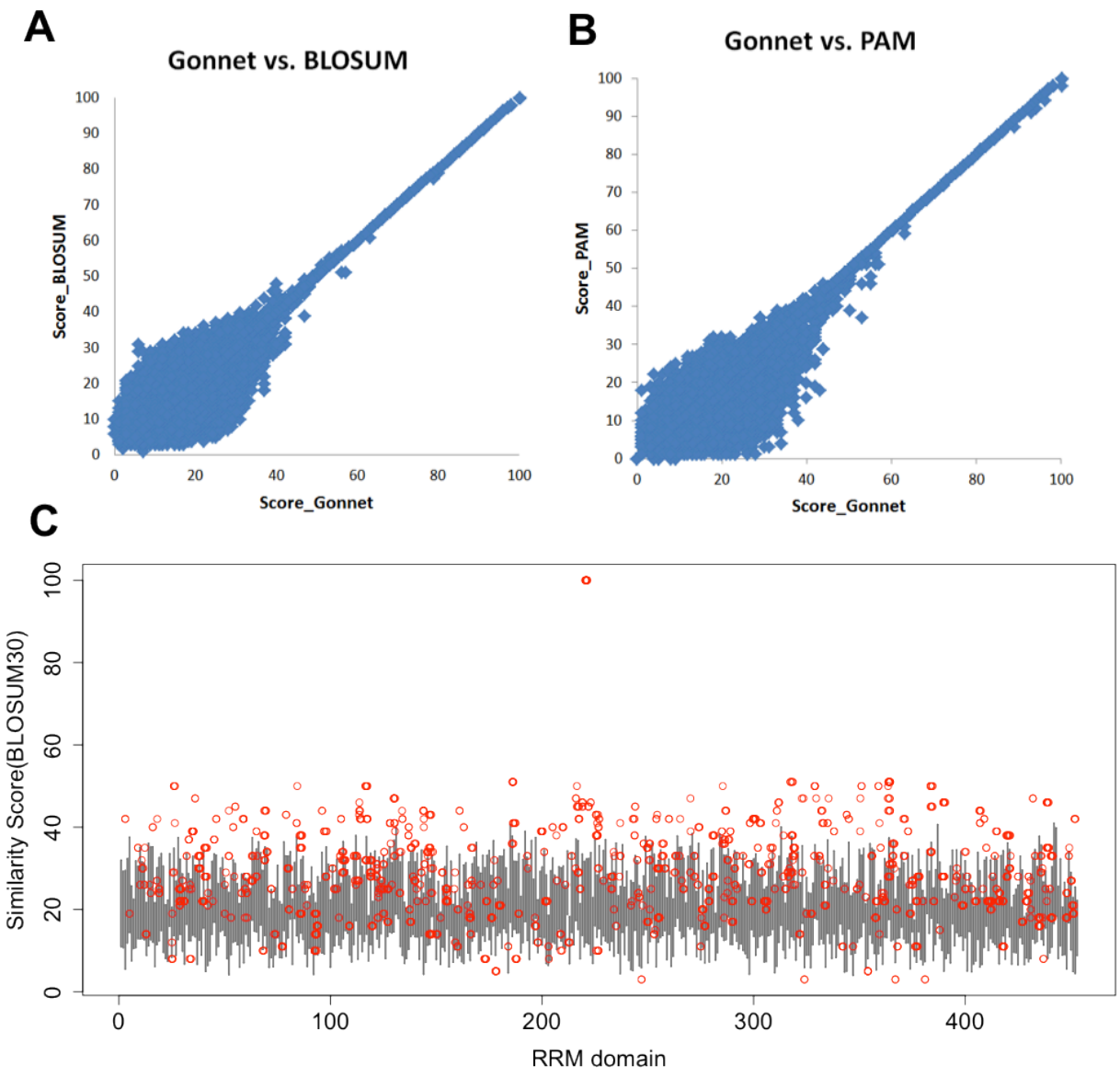


Figure S2. The similarity scores of RRMs measured with different scoring matrices.

To control for possible variations in our scoring schemes, we used different scoring matrices (Gonnet250, BLOSUM30 and PAM350) to measure similarity scores between all 453 RRM pairs in RBPs. (A) Scatter plot of scores measured by Gonnet250 vs. BLOSUM30. (B) Scatter plot of scores measured by Gonnet250 vs. PAM350. (C) Similarity scores between all RRM pairs in human RBPs. The data is plotted like Figure 1B, except the BLOSUM30 was used as the scoring matrix instead of Gonnet250.

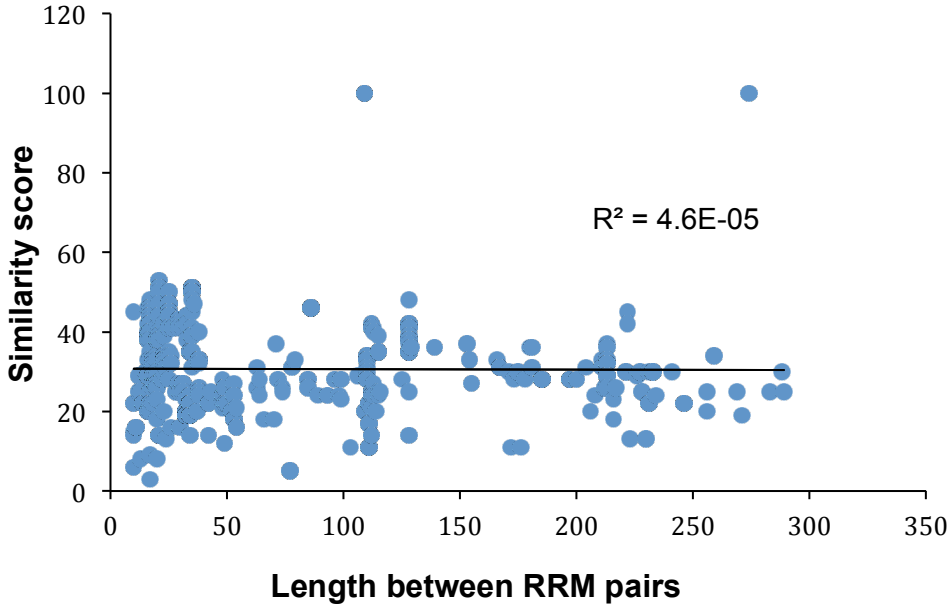


Figure S3. The lengths between sibling RRMs do not affect the similarity.

The similarity scores between sibling RRMs is plotted against the length (i.e. number of amino acid) between sibling RRMs. No correlation is found between the length and similarity scores.

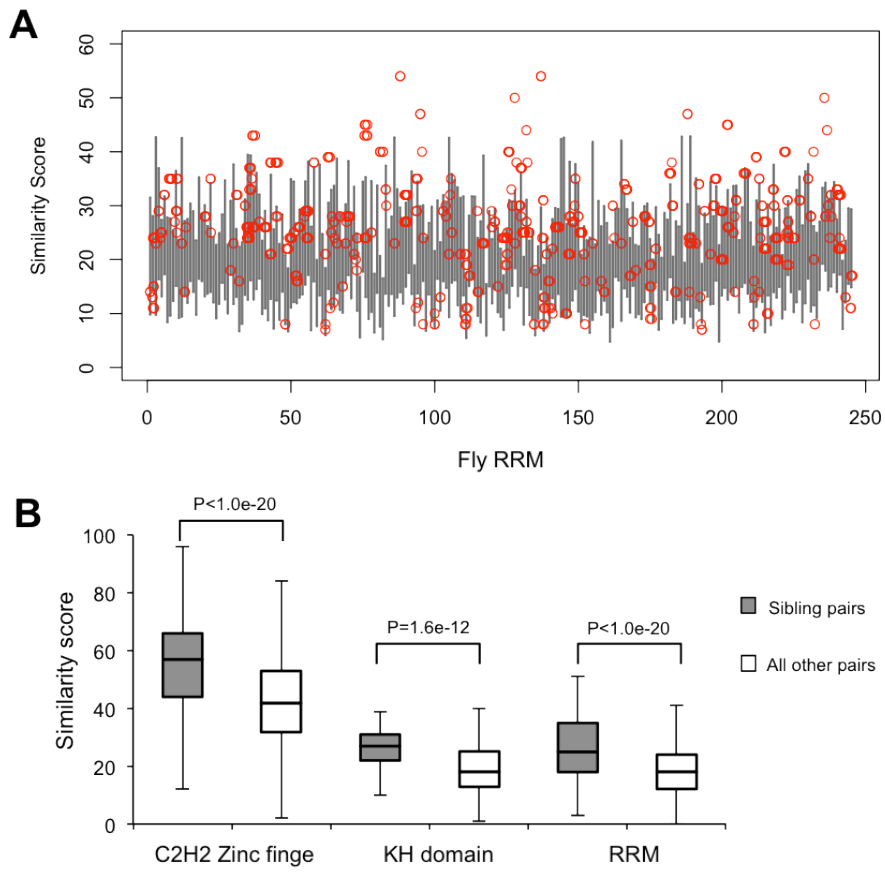
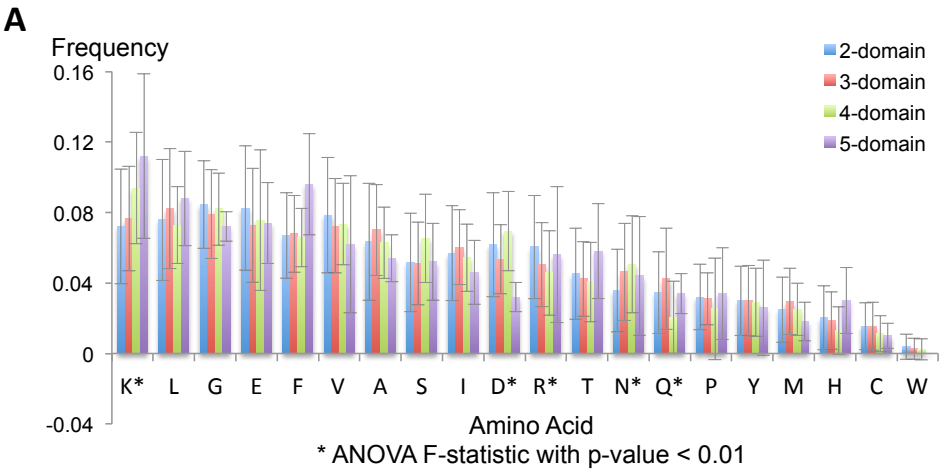


Figure S4. Similarity scores between RRM pairs in *D. melanogaster* RBPs and pairs of KH domain in human RBPs.

(A) Each RRM was aligned with all other RRMs in *D. melanogaster*, where the distribution of similarity score is represented by a grey vertical line spanning the mean ± 1× standard derivation. For proteins with multiple RRMs, the similarity score between the sibling RRMs was represented as a red circle. The order of RRM along the x-axis is arbitrary. (B) Boxplots of sequence similarities for both sibling domain pairs (gray box) and all other non-sibling domain pairs (white box) were shown in the three types of RNA binding domains. The sequences of C2H2 Zinc finger, KH domains and RRMs were extracted from human proteins.



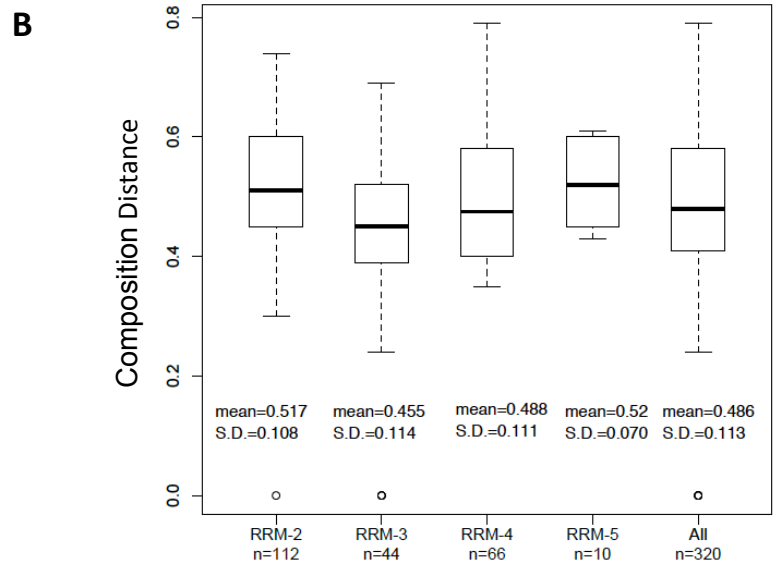


Figure S5. Amino acid composition frequency and composition distance in real RRM-pairs. (A) Amino acid frequencies of each RRM were calculated. We plotted the mean ±1 S.D. for RRMs in each group and compare the difference between groups using ANOVA test. Compositions that are significantly different between groups are denoted with *. (B) Composition distances between real RRM-pairs were calculated in each group and box plot of the distribution were plotted. We also listed the mean and standard deviation (S.D.) of the distribution.

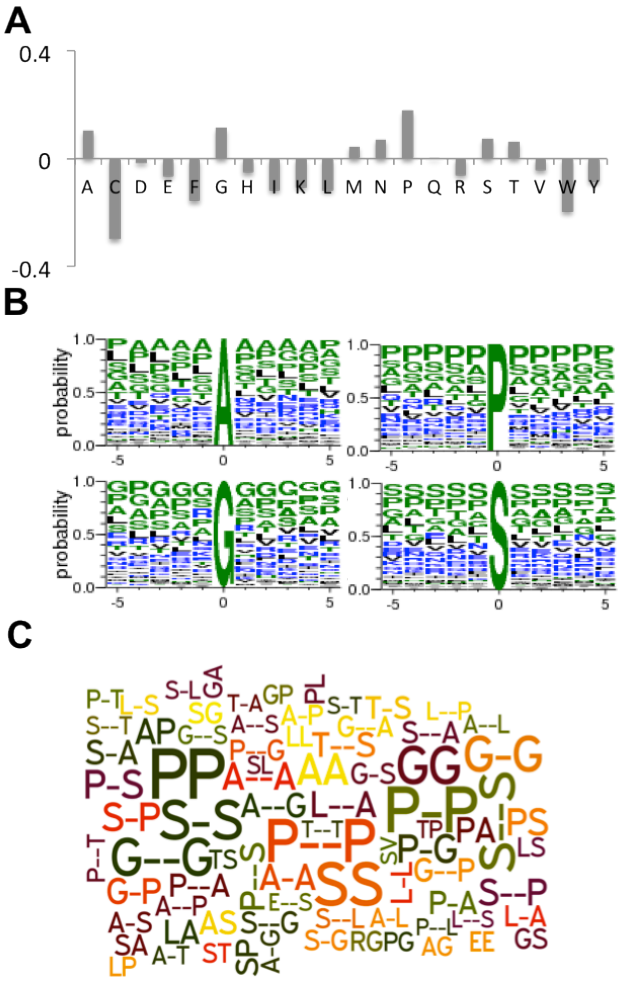


Figure S6. Sequence motifs enriched in human RBPs containing KH domain(s).

(A) We removed the KH sequences from the RBPs and analyzed the remaining sequence for amino acid propensities. For all 20 amino acids, their frequencies in non-KH regions were compared to other proteins in the human proteome and the relative ratio plotted. (B) Sequence logos around the most enriched amino acid residues in RBPs. The height of each single-letter amino acid code corresponds to the probability of occurrence at each position. (C) Repetitive sequence patterns that significantly co-occur with KH in all human proteins. The size of each pattern corresponds to the number of occurrence. The top 80 motifs are shown.

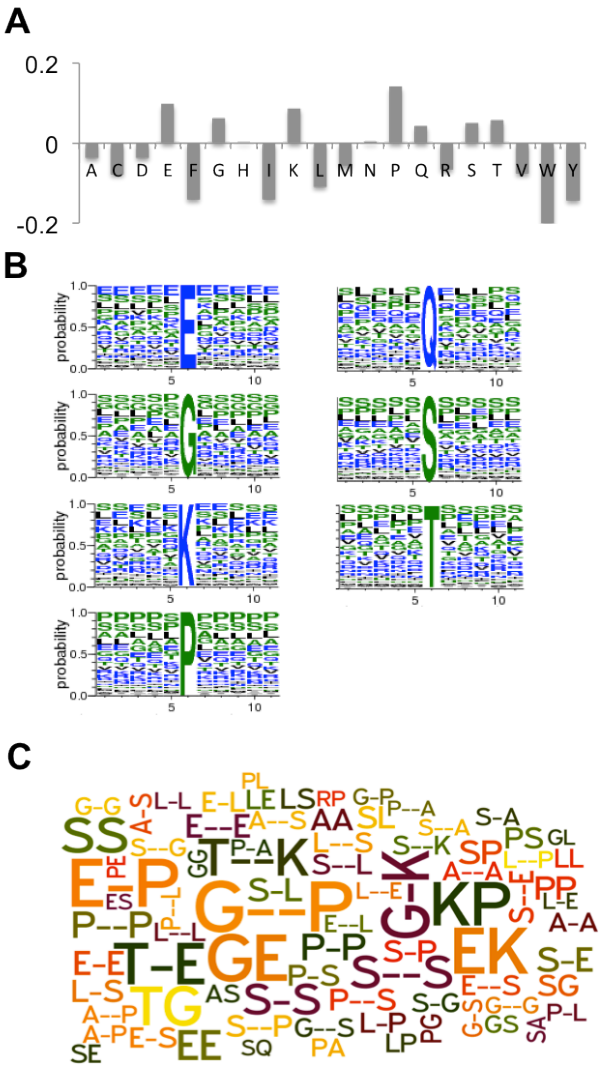


Figure S7. Sequence motifs enriched in human RBPs containing Zinc finger C2H2 domain(s).

(A) We removed the Zinc finger C2H2 sequences from the RBPs and analyzed the remaining sequence for amino acid propensities. For all 20 amino acids, their frequencies within non-Zinc finger C2H2 regions were compared to other proteins in the human proteome and the relative ratio plotted. (B) Sequence logos around the most enriched amino acid residues in RBPs. The height of each single-letter amino acid code corresponds to the probability of occurrence at each position. (C) Repetitive sequence patterns that significantly co-occur with Zinc finger C2H2 in all human proteins. The size of each pattern corresponds to the number of occurrence. The top 80 motifs are shown.