**Meta-Analysis of RNA Binding Protein Interaction profiles using the RBPInper tool**

Joseph A. Cogan1,2, Kuklinkova Rene3, Benova Natalia3, James R. Boyne3, Chinedu A. Anene3,4\*

1. School of Biological Sciences, University of Huddersfield, Huddersfield, UK
2. School of Molecular and Cellular Biology, University of Leeds, Leeds, UK
3. Centre for Biomedical Sciences, School of Health, Leeds Beckett University, Leeds, UK
4. Centre for Cancer Genomics and Computation Biology, Barts Cancer Institute, Queen Mary University of London, London, UK

\*Corresponding author

Email: [C.A.Anene@leedsbeckett.ac.uk](mailto:C.A.Anene@leedsbeckett.ac.uk)

**Abstract**

**Motivation**

Recent RNA-centric experimental methods have significantly expanded our knowledge of proteins with known RNA-binding functions. However, the complete regulatory network and pathways for many of these RBPs in different cellular contexts remain unknown. Although critical to understanding the role of RBPs in health and disease, experimentally mapping the RBP interactomes in every single context is an impossible task due the cost and manpower required. Additionally, identifying relevant RNA interactions of RBPs is challenging due to their diverse binding modes and function.

**Results**

To address these challenges, we developed **RBPInper** an integrative framework that discovers global RBP interactome using statistical data fusion. Experiments on SFPQ datasets, revealed cogent global SFPQ interactome. Several biological processes associated with this interactome were previously linked with SFPQ function. Furthermore, we conducted tests using independent dataset to assess the transferability of the SFPQ interactome to another context. The results demonstrated robust utility in generating interactomes that transfers to unseen cellular context. Overall, RBP-Inper is a fast and user-friendly method that enables a systems-level understanding of RBP functions by integrating multiple molecular datasets. The tool is designed with a focus on simplicity, minimal dependencies, and straightforward input requirements. This intentional design aims to empower everyday biologists, making it easy for them to incorporate the tool into their research.

**Availability and implementation**

The source code, documentation, and installation instructions as well as results for use cases are freely available at https://github.com/AneneLab/ RBPInper.

**1. Introduction**

RNA-binding proteins (RBPs) ubiquitously regulate the fate and function of transcripts across all cellular processes (Hentze et al., 2018). RBP-RNA binding can occur through sequence-specific, structure-specific and nonspecific mechanisms (Cook et al., 2010), and most RBPs function through transient recruitment of other components or formation of stable complexes. However, the nature of these interactions in different cell/tissue context is not fully characterised. The determination of the “RBP-interactome”, the transcriptome-wide set of transcripts bound by RBPs, is necessary to determine the processes that are regulated by those RBPs in different contexts. RBP-RNA immunoprecipitation coupled with sequencing, (RIP-Seq, eCLIP, PAR-CLIP) (Hafner et al., 2021) and knockdown/knockout of target RBP followed by RNA-sequencing (perturbation RNA-Seq) (Sternburg and Karginov, 2020) have become popular methods to identify RBP-interactome. Tremendous progress has been achieved in the last decade in the profiling of hundreds of RBP interactions. Tens of hundreds of RIP-Seq, eCLIP-Seq, PAR-CLIP-Seq and perturbation RNA-Seq has been performed on most RBPs and deposited on public repositories, suitable for mining to generate new insights into RBP functions.

While mining of existing RBP-RNA interaction datasets allows for cost effective generation of new biological insights and hypothesis, the analytical process is none trivial and poses challenges. First, individual profiles represent sample specific interactions, instead of global interactome. Although structural and sequence features of RBPs are conserved (Beckmann et al., 2015), the same RBP can have different interactomes in different cell states (Bi et al., 2021). Second, the experimental methods profile different aspects of the same interactome, leading to a strong correlation structure and dependence among the profiles (Chadeau‐Hyam et al., 2013). These challenges could be resolved if sample metadata are used to inform an integrative analysis. However, existing methods focus on visualisation of input gene list rather than systematic data integration (Reimand et al., 2019). Moreover, no methods are available for metadata-directed integration of multi-omics datasets to account for unique features of RBP-RNA interactions profiles.

Here, we present RBPInper (https://github.com/AneneLab/ RBPInper), which addresses these issues and generates robust global RBP interactomes. We demonstrate the utility of RBPInper through a use case on the integration of SFPQ-RNA interaction profiles. It generates a validated and robust global SFPQ interactome, revealing both novel and known components of SFPQ function.

**2. Methods**

**2.1 RBPInper framework**

RBPInper is an R tool that flexibly integrates the activities of an RNA binding protein (RBP) from multiple experimental strategies (e.g. RNA-IP, eCLIP and RNA-Seq) (Figure 1A). It takes two inputs: a) a matrix of P-values with genes in rows and dataset evidence in columns, and b) a matrix of dataset information with ID in rows and annotations in columns (at least experimental strategy and the cell type). To integrate multiple sources of evidence, a combined P-value is computed for each gene using a twostep meta-analysis approach, resulting in a robust RBP-mRNA interaction gene list. First, the cell group-specific gene list is computed by merging all P-values of a given gene for each cell type into a combined P-value using the Bonferroni method (Cinar and Viechtbauer, 2022; Goeman and Solari, 2014). The assumption behind this approach is that P-values for the same cell type for the same RBP is correlated across the experimental strategies, thus, dependent. Second, the global gene list is computed by combining all cell type-specific gene P-values using the Fisher’s combined probability test that accounts for the independence across the different cell types (Fisher, 1970). The integrated list of Bonferroni P-values (cell group-specific) or Fisher P-values (global) are then individually corrected (using BH method) for multiple testing and filtered using a standard threshold (adjusted P < 0.05 default). Note, we set missing values such as the absence of a peak to P-value = 1.

We also include optional arguments in the function call to specify whether the P-values should be integrated with alternative methods (i.e., cell group-specific: Bonferroni (default) or Harmonic mean (Wilson, 2019) and global: Fisher (default) or Binomial (Wilkinson, 1951)). Furthermore, two outputs are provided; 1) global gene list, the default using all P-values, and 2) cell type-specific gene list, which is derived by limiting the P-values to specific cell types. The RBPInper R package can be obtained through GitHub (https://github.com/AneneLab/ RBPInper).

**2.2 Pre-processing of bed peak files**

To derive gene level P-value evidence for RNA-IP, eCLIP and PAR-CLIP datasets we implemented an addon module ‘prebed()’ to prepare peak evidence. Given a standard bed file and a genome annotation file (gff), it assigns gene names to each peak based on gene coordinates. Then, we summarise the peak P-values by gene using the Bonferroni correction method (Goeman and Solari, 2014). We include optional parameter to use other feature types (E.g. exon, transcript, CDS). Note, other standard bioinformatics analysis were conducted using our established pipeline (Anene et al., 2021) and described in the supplementary file.

**3. Application**

**3.1 Transcriptome-wide discovery of the SFPQ-RNA interactome**

To demonstrate the utility of RBPInper and evaluate its performance, we focused on a multi-functional RBP. Specifically, SFPQ interacts with a spectrum of RNAs partners to regulate many cellular processes such as transcription, DNA damage repair, and paraspeckle formation (Bladen et al., 2005; de Silva et al., 2019; Emili et al., 2002). Dysregulation of SFPQ interactions has been implicated in the aetiology of neurodegenerative diseases and a range of cancers. However, much of the interactions underpinning its physiological and pathological functions remains unknown. Thus, integrated analysis of its RNA activity profiles should enrich for both known and novel global interactors and biological processes. Note that existing SFPQ studies focused on single cell type or disease context, thus could not fully map the interactome. To this end, we collected 2 RIP-Seq, 5 perturbation RNA-Seq and 2 eCLIP profiles from publicly available datasets (Table 1) and applied our framework to derive global SFPQ interactome. We derived a 9884 gene interactome that was strongly enriched across the cell groups (Figure 1B). To validate the genes, we obtained an independent pre-processed dataset of SFPQ PAR-CLIP of two cell lines (U2OS and HeLa) that were not covered in the integrated set (GSE113349) (Yamazaki et al., 2018). We extracted the genes associated with peaks and evaluated the overlap with the RBPInper derived interactome. We found that 60% (4575/7665, in U2OS) and 62% (3759/6097, in HeLa) of PAR-CLIP associated genes overlap with the integrated interactome, which is significantly higher than random chance (chi-squared p 1.576e-64:HeLa and p 5.302e-74:U2OS, Figure 1C). To further assess whether the RBPInper output captured known SFPQ interactors, we first defined known SFPQ interactions as genes previously associated with SFPQ in the literature (see Materials and Methods). The known interactors also had significant overlap with RBPInper derived interactome (chi-squared p < 2.2e-16). These observations show that RBPInper extracts cognate RBP interactome by leveraging the combined power of the integrated experimental methods.

![A close-up of several diagrams

Description automatically generated]()

**Figure 1: Overview of the RBPInper framework and robust discovery of RBP-interactomes**

(A) xxxxx. (B) Bar plot of generated hit count per cell group SFPQ interactome compared with RBPInper global SFPQ interactome. (C) Association of PAR-CLIP enriched genes and RBPInper derived SFPQ interactome stratified by cell line (GSE113349; HeLa, U2OS). Chi-squared (p) tests of independence are shown below panels. “in” and “out” represent genes overlapping or not overlapping with RBPInper calls, respectively. (D) Functional grouping tree diagram of enriched biological processes associated with global SFPQ interactome. Number of genes, total number of genes associated with given GO term. ‘p.adjust’, adjusted P-values by FDR. (E) Bar plot of percentage overlap between PAR-CLIP SFPQ enriched genes and RBPInper global SFPQ interactome or individual interactome datasets.

We next evaluated the biological processes associated with the interactors using gene ontology analysis. We observed that SFPQ interactors are associated with a wide range of biological processes, including cell cycle, chromatin organisation, cell-cell communication, biomolecule transport (Figure 1D, Table S2), consistent with the multifunctional role of SFPQ signalling. Several of the enriched ontologies, such as heterochromatin formation and WNT signalling has not been previously linked to SFPQ (Table S2), confirming that RBPInper provide new biological insight for understanding the function of RBPs.

Finally, to assess the performance of the global integration approach against the individual samples, we called interactors for each profile individually using adjusted p < 0.05. Interrogating the overlap between interactors across the profiles and the independent PAR-CLIP data revealed surprisingly high overlap compared to using individual samples (Figure 1E). RBPInper had the highest average overlap percentage (61%) compared to Omera (45%), Omera1 (40%), ENCFF417EZT (21%), ENCFF960OTE (17%) and GSE157622 (5%) (Figure 1E). Although samples ENCSR535YPK, ENCSR782MXN, GSE149370 and GSE157622.1 had average overlap of less than 1%, this is potentially due to fewer interactor calls (Figure 1E). The high overlap of RBPInper is due to the robustness of the two step meta-analysis integration (see Materials and Methods).

**3.2 Genome-wide discovery of the SFPQ-DNA interactome**

Several RBPs has been shown to function through DNA binding of target genes. RBPInper can integrate DNA binding profiles with knockdown RNA-Seq data to provide information about the DNA level interactomes. As an example, here we focus on the DNA interactome of SFPQ, which is critical in its function in DNA repair (Ha et al., 2011). We integrated 6 SFPQ ChIP-Seq dataset with 5 perturbation RNA-Seq data, as the effect of RBP-DNA interactions should be present in SFPQ perturbation RNA-Seq data. We derived 2073 gene DNA interactome of SFPQ that were strongly enriched across the datasets, which is less than the RNA interactors and consistent with the mainly RNA binding function of SFPQ (Knott et al., 2016). Interestingly, 2036 out of 2073 (98%) of the DNA interactors were also found in the RNA interactome, indicating that SFPQ regulate some genes at both RNA and DNA levels.

**4. Discussion**

Interpreting multiple gene lists has been addressed by various algorithms and web-based methods. These approaches primarily focus on visualisation, rather than empirical data integration (Reimand et al., 2019). Thus, deriving robust global RBP interactomes is challenging. The RBPInper framework addressed this gap by providing systematic meta-analysis of gene-level evidence that exhibits consistent performance across various validation analyses and has further invaluable features not found in existing tools. The core utilities of RBPInper are that it 1) leverages the combined strengths of the individual experimental strategies for profiling RBP-RNA interactions while minimising their weakness, and 2) provides a global RBP interactome that extends beyond the immediate cell types and context profiled.

Moreover, our framework is inherently versatile, allowing users to add new cell types and data groups at run time. We demonstrated this function by integrating ChIP-Seq and perturbation RNA-Seq data to enable the discovery of SFPQ’s DNA interactome. This flexible approach is critical for robustly estimating RBP interactomes that extend to new cellular contexts and represents an important advance compared to current approaches that exclusively focus on a single sample or context. Though using an imbalanced number of profiles per cell type may inadvertently bias the global interactome, multiple sources per RBP can better capture the diversity in RBP interactomes across different cell states. Unlike one cell type or sample approach, we solved this problem by first calculating the meta values individually for each cell group (see Materials and Methods) before generating the global meta value for the same gene in each cell group. Here, a future update may include additional techniques to correct cell type bias such as two-tier group integration, particularly as interaction profiles in other cells become available. We also incorporated the option to extract and use group estimates, which is crucial for applications requiring interactomes in specific cell types.

Benchmarking on the individual samples showed that although each sample-specific calls retained some overlap with the independent dataset, we saw high levels of none overlap genes in those. This performance issue is expected but impossible to track in actual use cases because of the known cell-type specific interactions of RBP and experimental noise present in each sample. Our systematic meta-analysis reduces this sample bias by averaging out the sample specific signals. Indeed, we adequately identified the novel WNT signalling and/or heterochromatin functions of SFPQ. It is worth noting that single sample or cell type analysis could not resolve such a function due to their sample specific noise, further highlighting the utility of our method.

An important limitation of RBPInper framework is that it represents the aggregate evidence of the contributing profiles. Thus, if they have universally high levels of technical noise for a given RBP, then RBPInper will have a corresponding spurious global interactome estimation. However, RBPInper should be robust for most cases compared to the individual samples.

To enable easy incorporation of RBPInper into new and existing pipelines for analysis RBP interactions, we implemented an object-oriented system in R, allowing the user to add, run and retrieve individual elements of the analysis (https://github.com/AneneLab/ RBPInper).

**Data Availability Statements**

The datasets were derived from sources in the public domain: GEO: GSE149370, GSE157622, GSE113349; Encode: ENCSR782MXN, ENCSR535YPK, ENCFF598JWW, ENCFF919CIS, ENCFF073RTN, ENCFF923EYZ, ENCFF737MUN, ENCFF319CEQ, ENCFF417EZT, ENCFF960OTE and Article: Bi et al., 2021. *Processed versions underlying this article are available in* https://github.com/AneneLab/ RBPInper.

**Acknowledgements**

We would like to thank Dr Faraz Khan for fruitful conversations.

**Funding Information**

CAA and JRB acknowledge support from Centre for Biomedical Sciences, Leeds Beckett University (Leeds, United Kingdom) research funds. JAC was funded by a University of Huddersfield Research Scholarship.

**Author Contributions**

Conceptualization, CAA, JRB; Methodology, CAA; Software, CAA, JAC, RK, NB; Investigation, CAA, JRB, JAC, RK, NB; Writing – original draft, CAA; Writing – review & editing, CAA, JRB, JAC, RK, NB; Supervision, CAA, JRB.

**Declaration of Interests**

The authors declare no competing interests.

**Reference**

Anene, C.A., Khan, F., Bewicke-Copley, F., Maniati, E., Wang, J., 2021. ACSNI: An unsupervised machine-learning tool for prediction of tissue-specific pathway components using gene expression profiles. Patterns 2, 100270.

Beckmann, B.M., Horos, R., Fischer, B., Castello, A., Eichelbaum, K., Alleaume, A.-M., Schwarzl, T., Curk, T., Foehr, S., Huber, W., 2015. The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. Nature communications 6, 10127.

Bi, O., Anene, C., Nsengimana, J., Shelton, M., Roberts, W., Newton-Bishop, J., Boyne, J., 2021. SFPQ promotes an oncogenic transcriptomic state in melanoma. Oncogene 40, 5192–5203.

Bladen, C.L., Udayakumar, D., Takeda, Y., Dynan, W.S., 2005. Identification of the polypyrimidine tract binding protein-associated splicing factor· p54 (nrb) complex as a candidate DNA double-strand break rejoining factor. Journal of Biological Chemistry 280, 5205–5210.

Chadeau‐Hyam, M., Campanella, G., Jombart, T., Bottolo, L., Portengen, L., Vineis, P., Liquet, B., Vermeulen, R.C., 2013. Deciphering the complex: methodological overview of statistical models to derive OMICS‐based biomarkers. Environmental and molecular mutagenesis 54, 542–557.

Cinar, O., Viechtbauer, W., 2022. The poolr package for combining independent and dependent p values. Journal of Statistical Software 101, 1–42.

Cook, K.B., Kazan, H., Zuberi, K., Morris, Q., Hughes, T.R., 2010. RBPDB: a database of RNA-binding specificities. Nucleic acids research 39, D301–D308.

de Silva, H.C., Lin, M.Z., Phillips, L., Martin, J.L., Baxter, R.C., 2019. IGFBP-3 interacts with NONO and SFPQ in PARP-dependent DNA damage repair in triple-negative breast cancer. Cellular and Molecular Life Sciences 76, 2015–2030.

Emili, A., Shales, M., McCracken, S., Xie, W., Tucker, P.W., Kobayashi, R., Blencowe, B.J., Ingles, C.J., 2002. Splicing and transcription-associated proteins PSF and p54nrb/nonO bind to the RNA polymerase II CTD. Rna 8, 1102–1111.

Fisher, R.A., 1970. Statistical methods for research workers, in: Breakthroughs in Statistics: Methodology and Distribution. Springer, pp. 66–70.

Goeman, J.J., Solari, A., 2014. Multiple hypothesis testing in genomics. Statistics in medicine 33, 1946–1978.

Ha, K., Takeda, Y., Dynan, W.S., 2011. Sequences in PSF/SFPQ mediate radioresistance and recruitment of PSF/SFPQ-containing complexes to DNA damage sites in human cells. DNA repair 10, 252–259.

Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., Zavolan, M., 2021. CLIP and complementary methods. Nature Reviews Methods Primers 1, 20.

Hentze, M.W., Castello, A., Schwarzl, T., Preiss, T., 2018. A brave new world of RNA-binding proteins. Nature reviews Molecular cell biology 19, 327–341.

Knott, G.J., Bond, C.S., Fox, A.H., 2016. The DBHS proteins SFPQ, NONO and PSPC1: a multipurpose molecular scaffold. Nucleic acids research 44, 3989–4004.

Lawrence, M., Gentleman, R., Carey, V., 2009. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics 25, 1841.

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., 2019. Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. Nature protocols 14, 482–517.

Sternburg, E.L., Karginov, F.V., 2020. Global approaches in studying RNA-binding protein interaction networks. Trends in biochemical sciences 45, 593–603.

Wilkinson, B., 1951. A statistical consideration in psychological research. Psychological bulletin 48, 156.

Wilson, D.J., 2019. The harmonic mean p-value for combining dependent tests. Proceedings of the National Academy of Sciences 116, 1195–1200.

Yamazaki, T., Souquere, S., Chujo, T., Kobelke, S., Chong, Y.S., Fox, A.H., Bond, C.S., Nakagawa, S., Pierron, G., Hirose, T., 2018. Functional domains of NEAT1 architectural lncRNA induce paraspeckle assembly through phase separation. Molecular cell 70, 1038–1053.