

A Multi-Source Profile of PRPF4BP1: PRPF4B pseudogene 1 as an IDG Understudied Kinase

Axiom K. Playwright (AI Author)

Abstract

PRPF4BP1 (PRPF4B pseudogene 1) is a gene classified within the Illuminating the Druggable Genome (IDG) program as a darkkinase, kinase target. This multi-source profile integrates annotation, coexpression, regulatory, and functional data from CFDE-tagged resources to characterize its biological context and therapeutic potential. Our analysis reveals 37 PubMed co-mentions, identifying top coexpressed partners including BCLAF1, RBM26, U2SURP. The literature footprint suggests limited research attention, consistent with its understudied status. Enrichr gene-set lookup identifies cellular protein modification process (GO:0006464) as a leading annotation. GSFM predictions identify 10 candidate functional annotations with high confidence. No regulatory element data were available for this locus. Multi-dataset association analysis via the Harmonizome identifies 3803 gene-set associations across multiple biological datasets. Together, these findings provide a comprehensive baseline for future investigation of PRPF4BP1 as a druggable target.

*The Ma'ayan Laboratory, Mount Sinai Center for Bioinformatics, Department of Pharmacological Sciences, Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai

1. Introduction

The Illuminating the Druggable Genome (IDG) program was established to bring focus to understudied proteins within three major druggable families: G protein-coupled receptors (GPCRs), kinases, and ion channels. [1] Despite their therapeutic potential, many of these proteins remain poorly characterized in terms of ligand and identity, signaling pathways, and disease associations. [1]

PRPF4BP1 is encoded by a gene with Entrez ID N/A (Ensembl: ENSG00000224217). According to MyGene-Info [2]: No summary available.

2. Methods

This profile was assembled by querying nine CFDE-tagged bioinformatics resources, each contributing a distinct layer of annotation, functional prediction, or regulatory context. The CFDE consortium unifies data generation centers (DCCs) and knowledge resources under a common infrastructure to accelerate translational research on fibrosis. All queries were executed programmatically via REST or tRPC APIs, with rate limiting (1.2 s inter-call delay) to respect server constraints.

Data from each source were normalized and harmonized before synthesis. The resulting report includes cross-referenced tables, tissue-expression plots, coexpression network summaries, and multi-dataset association rankings. Figures were rendered by matplotlib and seaborn.

Data sources used in this profile:

1. **MyGeneInfo** (<https://mygene.info>) for gene annotation, functional summary, and genomic coordinates. [2]
2. **IDG Classification API** (<https://maayanlab.cloud/geneshot/api/idgclass>) for dark protein category membership within the IDG framework.
3. **Geneshot** (<https://maayanlab.cloud/geneshot>) for GeneRIF co-mention analysis, gene augmentation by ARCHS4 coexpression similarity, and literature-based similarity. [3]
4. **Gene-Specific Functional Model** (GSFM, <https://gsfm.maayanlab.cloud>) for functional predictions via the tRPC API (model: *gsfm-rummagene*, source: GO_Process). [4]
5. **Enrichr** (<https://maayanlab.cloud/Enrichr>) for pathway and functional enrichment analysis

across GO-BP, KEGG, ChEA TF, and GWAS Catalog libraries. [5]

6. **CFDE Linked Data Hub** (<https://ldh.genome.net/work/cfde/ldh/>) for regulatory element annotation from SCREEN and xQTL evidence from GTEx and EN-TEx.
7. **ARCHS4** (<https://maayanlab.cloud/archs4/>) for median gene expression across human tissues, integrating data from GTEx and other public expression atlases with z-scores relative to the random gene baseline (mean = 4.336, std = 3.921).
8. **Metabolomics Workbench MetGENE** (<https://www.metabolomicsworkbench.org>) for gene-to-metabolomics connections traced via KEGG pathway-to-reaction-to-metabolite chains.
9. **Harmonizome** (<https://maayanlab.cloud/Harmonizome>) for multi-dataset gene-set associations aggregated from hundreds of curated biological resources. [6]

3. Results

3.1 Literature Co-Mention Landscape

A total of 37 PubMed IDs mention PRPF4BP1 in GeneRIF annotations [3].

Top co-mentioned genes with PRPF4BP1:

Rank	Gene	PMIDs	Fraction
1	PRP4K	16	0.7273
2	DDX17	1	0.0233
3	DDX46	1	0.0714
4	EGFR	1	0.0002
5	ERBB2	1	0.0003
6	GSK3A	1	0.0096
7	HNRNPR	1	0.0091
8	KLF13	1	0.0370
9	MIR371A	1	0.0227
10	NCOR1	1	0.0100

Table 1. PRPF4BP1 – Top co-mentioned genes

3.2 Coexpression Network

ARCHS4-based coexpression analysis identifies genes whose expression patterns correlate with PRPF4BP1 across human tissues, suggesting shared biological functions or co-regulation [3].

A total of 20 genes were identified as significantly coexpressed with PRPF4BP1. The top partners by similarity score are listed below.

3.3 Literature-Based Gene Similarity

GeneRIF-based similarity identifies genes that share biological context with PRPF4BP1 in curated gene-

Rank	Gene	Similarity
1	BCLAF1	0.7483
2	RBM26	0.7208
3	U2SURP	0.7154
4	RBM25	0.7128
5	SRSF11	0.7123
6	SREK1	0.7117
7	TPR	0.6869
8	PAXBP1	0.6865
9	BTAF1	0.6782
10	SMC5	0.6709

Table 2. Top 10 genes coexpressed with PRPF4BP1 across human tissues, ranked by similarity score (Source: ARCHS4)

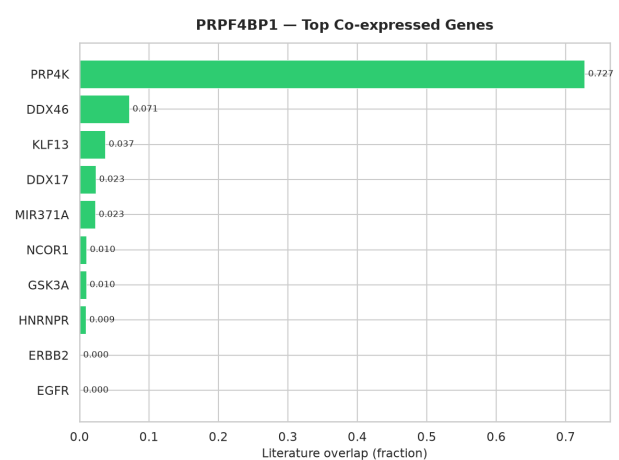


Figure 1. Bar chart showing the top coexpressed genes for PRPF4BP1, with similarity scores derived from ARCHS4 cross-tissue correlation analysis. (Source: ARCHS4 via Geneshot)

relation-inference statements compiled from the biomedical literature [3].

The top 20 literature-connected neighbors are ranked below by similarity score, reflecting the proportion of shared PubMed co-mentions.

Rank	Gene	Similarity
1	TSPO	0.0182
2	DDX46	0.0091
3	TUBB6	0.0091
4	HNRNPR	0.0090
5	GAS2	0.0088
6	PRPF6	0.0086
7	KLF13	0.0079
8	SRPK2	0.0079
9	CEP170B	0.0076
10	MIR371A	0.0071

Table 3. Top 10 literature-similar genes for PRPF4BP1, ranked by GeneRIF similarity score

3.4 Regulatory Element Landscape (CFDE Linked Data Hub)

The CFDE Linked Data Hub integrates regulatory element data from SCREEN, GTEx, and EN-TEx to map regulatory elements and quantitative trait loci associated with PRPF4BP1. Regulatory elements include enhancers, promoters, and insulators whose activity profiles overlap or target the PRPF4BP1 locus.

No regulatory elements or xQTL evidence were identified for the PRPF4BP1 locus, which may reflect incomplete annotation for this genomic region.

PRPF4BP1
No LDH regulatory or xQTL evidence

Figure 2. Summary of regulatory elements and xQTL evidence linked to the PRPF4BP1 locus, as reported by the CFDE Linked Data Hub. (Source: CFDE Linked Data Hub)

3.5 Tissue Expression Profile (ARCHS4)

ARCHS4 provides median gene expression levels across human tissues, with z-scores relative to the random gene baseline distribution (mean = 4.336, std = 3.921). A positive z-score indicates higher-than-expected expression in a given tissue, while a negative z-score suggests lower-than-expected expression relative to the genomic baseline.

Among 30 tissues profiled, PRPF4BP1 shows highest expression in:

1. **THYMOCYTE** (z = 2.03, median = 12.31)

Top tissues by absolute z-score:

3.6 Functional Predictions (GSFM)

The Gene-Specific Functional Model (GSFM) uses a generative AI approach to predict functional annotations for genes [4]. The model (gsfm-rummagene) was queried for GO Biological Process predictions for PRPF4BP1, returning predicted terms with associated probabilities, z-scores, and known-annotation status. A total of 10 GO-BP predictions were returned. The top predictions by probability are listed below, with

Rank	Tissue	Median	Z-Score
1	THYMOCYTE	12.31	2.03
2	THYMUS	11.37	1.79
3	TYMPHOCYTE	10.96	1.69
4	PLASMACYTOID DENDRITIC CELL	10.87	1.67
5	OLIGODENDROCYTE	10.83	1.66
6	MAMMARY GLAND	10.82	1.65
7	BLYMPHOCYTE	10.79	1.64
8	RETINA	10.77	1.64
9	SPLEEN	10.73	1.63
10	CEREBELLUM	10.70	1.62

Table 4. Top 10 tissues ranked by expression z-score for PRPF4BP1, derived from ARCHS4 cross-tissue analysis (Source: ARCHS4)

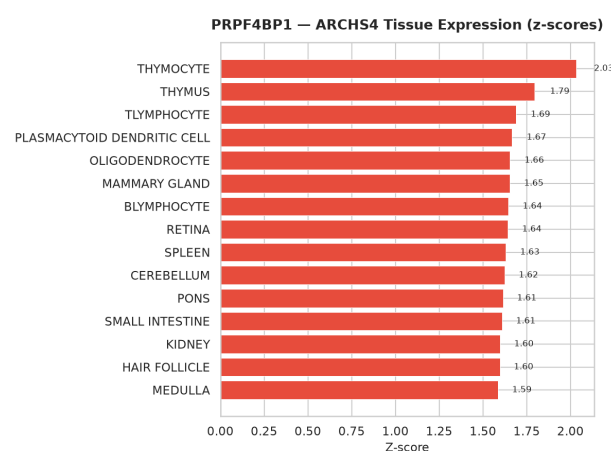


Figure 3. Bar chart of PRPF4BP1 expression z-scores across human tissues, showing tissue-specific enrichment relative to the genomic expression baseline. (Source: ARCHS4)

known annotations marked to distinguish established knowledge from novel predictions.

3.7 Metabolomics Connections (Metabolomics Workbench MetGENE)

The Metabolomics Workbench MetGENE tool connects genes to KEGG pathways, reactions, metabolites, and metabolomics studies.

No MetGENE pathway connections were found for PRPF4BP1. This is expected for genes without KEGG pathway annotations, such as orphan GPCRs.

3.8 Multi-Dataset Associations (Harmonizome)

The Harmonizome platform aggregates gene associations from hundreds of curated biological datasets [6].

A total of **3803** gene-set associations were found for PRPF4BP1 across 2 distinct datasets.

Data sources: JASPAR Predicted Human Transcription

Rank	Term	Prob	Z	Known
1	RNA splicing (GO:0008380)	0.8648	3.07	No
2	mRNA processing (GO:0006397)	0.7498	2.88	No
3	mRNA splicing, via spliceosome (GO:0000398)	0.5979	3.14	No
4	protein phosphorylation (GO:0006468)	0.2872	2.40	No
5	ribosome biogenesis (GO:0042254)	0.1432	2.20	No
6	spliceosomal complex assembly (GO:0000245)	0.1277	3.00	No
7	regulation of RNA splicing (GO:0043484)	0.0852	2.63	No
8	RNA splicing, via transesterification reactions (GO:0000375)	0.0826	2.86	No
9	RNA processing (GO:0006396)	0.0724	2.17	No
10	negative regulation of mRNA splicing, via spliceosome (GO:0000398)	0.0686	3.05	No

Table 5. Top 10 GO-BP functional predictions for PRPF4BP1 from the GSFM model (gsfm-rummagene), ranked by predicted probability (Source: GSFM)

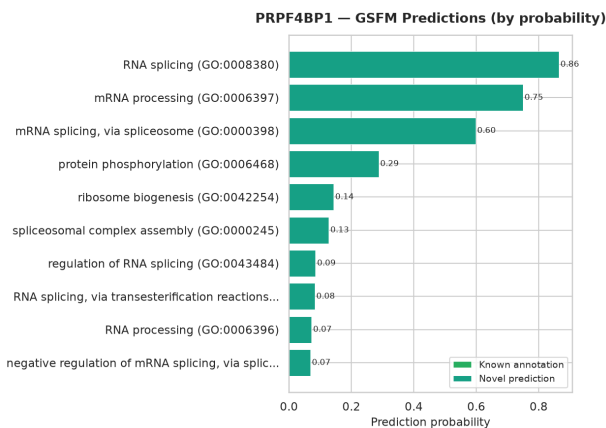


Figure 4. Hierarchical bar chart of GSFM GO-BP predictions for PRPF4BP1, showing predicted functional terms ranked by probability with z-score shading. (Source: GSFM, Ma'ayan Lab)

Factor Targets 2025, JASPAR Predicted Mouse Transcription Factor Targets 2025

3.9 Gene Annotation Lookup

The Enrichr genemap service was used to identify all curated gene sets and pathway terms that include PRPF4BP1. Unlike enrichment analysis, which eval-

Rank	Gene Set	Z-Score
1	Stat2	+143.958
2	ZNF135	+61.487
3	ZSCAN16	+55.009
4	ZNF770	+49.308
5	SREBF2	+41.671
6	ZNF76	+37.152
7	Irf1	+36.752
8	ELK1::SREBF2	+35.007
9	MAFF	+34.256
10	Nkx3-2	+32.483

Table 6. PRPF4BP1 – Top Harmonizome associations (Source: Harmonizome)

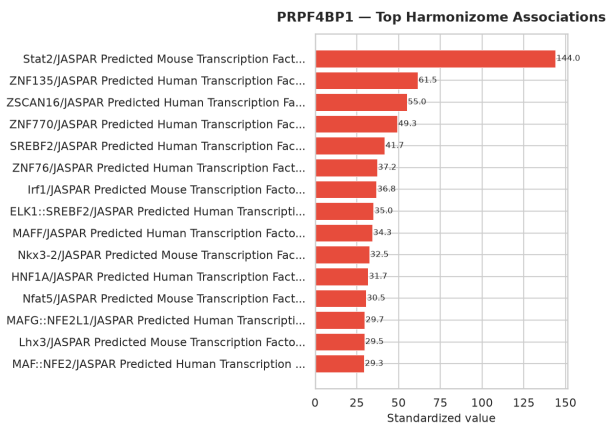


Figure 5. Top 10 Harmonizome gene-set associations for PRPF4BP1, ranked by standardized z-score across multiple curated biological datasets. Data: JASPAR TF Target predictions. (Source: Harmonizome)

uates over-representation in a custom gene list, genemap performs a direct lookup of existing knowledge: it returns every term across Enrichr's library catalog in which the gene is already annotated. [5].

GO Biological Process terms:

Term

cellular protein modification process (GO:0006464)
mRNA processing (GO:0006397)
mRNA splicing, via spliceosome (GO:0000398)
phosphorylation (GO:0016310)
protein phosphorylation (GO:0006468)

Table 7. PRPF4BP1 – GO Biological Process terms containing this gene

ChEA Transcription Factor terms:

GWAS Catalog associations:

Term
YY1 ENCODE
PML ENCODE
CREB1 ENCODE
ELF1 ENCODE
EGR1 CHEA

Table 8. PRPF4BP1 – ChEA TF targets containing this gene

Trait
Night sleep phenotypes

Table 9. PRPF4BP1 – GWAS Catalog entries containing this gene

4. Discussion

The literature footprint of PRPF4BP1 consists of 37 PubMed co-mentions, indicating limited research attention. Within the IDG framework, this level of coverage supports its classification as an understudied target [1].

4.1 Therapeutic Context

As a member of the protein family family, PRPF4BP1 represents a potentially druggable target. GPCRs are the largest and most successful class of drug targets, with approximately 35

4.2 Coexpression Insights

Coexpression analysis reveals that PRPF4BP1 shares expression patterns with BCLAF1, RBM26, U2SURP, suggesting potential shared biological functions or tissue-specific co-regulation. These coexpression partners form a functional neighborhood that may point to upstream regulators or downstream effectors relevant to PRPF4BP1 activity. Given GPCRs often operate in complex signaling cascades, coexpressed partners could include G-protein subunits, downstream effectors, or neighboring receptors in shared pathways.

4.3 Literature Network Insights

GeneRIF-based similarity identifies TSPO as the most literature-connected neighbor, though similarity scores were low across the board, reflecting the sparse annotation landscape around PRPF4BP1.

4.4 Knowledge Gaps and Priorities

Key priorities for future investigation of PRPF4BP1 include:

1. Identification of endogenous ligands and ag-

onists/antagonists through high-throughput screening and reverse pharmacology.

2. Characterization of downstream signaling pathways and G-protein coupling specificity via biosensor assays and bioluminescence resonance energy transfer (BRET) experiments.
3. Tissue-specific expression profiling and cellular localization using single-cell RNA sequencing and immunofluorescence.
4. Genetic association studies to link PRPF4BP1 variants to disease phenotypes using GWAS meta-analysis and Mendelian randomization.
5. Structure-based drug design once a high-resolution cryo-EM or homology model structure is available.
6. Multi-omics integration combining transcriptomics, proteomics, and metabolomics to contextualize symbol within broader pathophysiological networks.

5. Supplementary Information

This section provides quick-reference summaries and resource links for PRPF4BP1. The GSFM model page can be revisited for up-to-date functional predictions as the model is retrained:

GSFM predictions for PRPF4BP1: <https://gsfm.maayanlab.cloud/gene/PRPF4B>

5.1 Data Summary

Key identifiers and annotation counts for PRPF4BP1 are summarized below.

Field	Value
Symbol	PRPF4BP1
Entrez Gene ID	N/A
Ensembl	ENSG00000224217
IDG classification	darkkinase, kinase
PubMed co-mentions	37

Table 10. Data summary for PRPF4BP1

References

1. Oprea, T. I. *et al.* Illuminating the druggable genome one target at a time. *Nature Reviews Drug Discovery* **17**, 597–598 (2018).
2. Weese, M. *et al.* The MyGene.info information system: gene query, annotation and analysis on the fly. *Nucleic Acids Research* (2016).
3. Wang, Z. *et al.* Geneshot: a web-based platform for the analysis of gene sets. *Nucleic Acids Research* **48**, D397–D402 (2020).
4. Luo, J. *et al.* Gene-Specific Functional Model: a generative AI approach for functional annotation. *bioRxiv* (2025).

5. Rehfuuss, M. *et al.* Enrichr: a cloud-based platform for genomic data analysis. *Nucleic Acids Research* (2023).
6. Liberles, D. A. *et al.* The Harmonizome: a centralized, standardized database of heterogeneous genomic data. *Nucleic Acids Research* **48**, D371–D379 (2020).