2020 Oct 12

Investigation of transcriptome-wide STAU2 RNA-binding through CLIP analysis

Albert Martí I Sabarí, Dr Flora Lee, Dr Jernej Ule.
RNA Regulatory Networks Laboratory, Francis Crick Institute

CONTENTS

SUMMARY

Eukaryotic cells are characterised by their complex regulatory networks. RNA-binding proteins (RBPs) are a crucial part of these networks that bind RNA transcripts, forming ribonucleoproteins (RNPs), and localising them in specific parts of the cytoplasm to carry out different functions. In this project, I look at a specific RBP called Staufen2 (STAU2) that binds double-stranded RNA molecules and is known to be involved in the development of the nervous system and plasticity in adult brains.

A recent publication by Bajaj et al. identified STAU2 as a key regulator of the progression of aggressive myeloid leukaemias through an integrated analysis of CRISPR, RNA-Sequencing and eCLIP. In parallel, my host laboratory (the Ule Lab at the Francis Crick Institute) produced iCLIP data from the study of the role of STAU2 in neuronal development. These datasets contain a list of the genes found to bind STAU2 and their counts. By comparing them, I have identified a relatively small subset of genes that were strongly enriched in both datasets that serve as potential future candidates of study for both the study of leukaemia progression and the development of neurons.

INTRODUCTION

RNA-binding proteins (hereby referred to as RBPs) constitute an increasingly relevant type of proteins that bind cellular RNA, forming ribonucleoprotein (RNP) complexes with various subcellular localizations (Hentze et al., 2018). This diverse group of proteins interact in a complex regulatory network whereby different multimeric combinations in RNPs result in different outcomes, which overall affect virtually every aspect of post-transcriptional RNA metabolism and, ultimately, gene expression (Hogan et al., 2008).

A particularly important example of RBPs are Staufen proteins – the Stau1 and Stau2 orthologues in mammals –, which exhibit multiple double-stranded RNA-binding domains (dsRBDs). They have distinct evolutionarily conserved roles and mechanisms of mRNA segregation and formation of RNPs in bilaterian lineages (Heraud-Farlow & Kiebler, 2014). To date, Staufen proteins have been researched mostly in the nervous system, where they asymmetrically segregate cell fate determinants during development (Kusek et al., 2012; Li et al., 1997) and play important roles in synaptic function and plasticity in adult brains (Dubnau et al., 2003).

RBP misfunctions affecting RNP formation have been linked to a variety of neurological disorders, meaning a better understanding of how they bind RNA could be used in the development of treatments (Bassell & Kelic, 2004; Heraud-Farlow & Kiebler, 2014). In order to study this, crosslinking and immunoprecipitation (CLIP) methods have been developed to map the transcriptome-wide binding events of specific RBPs of interest (Lee & Ule, 2018; Ule et al., 2003). In short, these methods combine covalent UV crosslinking of RNA fragments to a protein of interest with immunoprecipitation that purifies the complexes formed, allowing for the isolation of the linked RNA fragments for their sequencing (Licatalosi et al., 2008). Many variants of CLIP methods have been developed in recent years, such as eCLIP (Van Nostrand et al., 2016) and iCLIP (König et al., 2010), which allow single-nucleotide resolution mapping of binding events by identifying crosslinking-induced truncations (Sugimoto et al., 2012).

In this project, I aim to characterise the conserved RNA binding and function of STAU2 through the computational analysis of two STAU2 CLIP datasets performed on distinct human cell lines. The first is derived from eCLIP experiments in K562 human bcCML cells recently published by Bajaj et al. In their paper, Bajaj et al. researched the molecular cues of aggressive myeloid leukaemias through a variety of molecular and computational approaches (Bajaj et al., 2020). Through genome-wide *in vivo* CRISPR screenings, they identified thousands of candidate genes involved in the development of leukaemia. This analysis revealed an overrepresentation of RBPs implicated in the progression of leukaemia, from which they identified STAU2 as a specific candidate for its enriched expression in leukaemia stem cells (LSCs). Furthermore, Stau2 siRNA knockdown experiments in mice LSCs showed robust effects on colony formation and reduced the onset of the disease, achieving similar results in human K562 cells. Following an enhanced crosslinking and immunoprecipitation (eCLIP) protocol in K562 human bcCML cells the authors identified thousands of Stau2 effectors, and they concluded STAU2 plays a major role in the regulation of chromatin-binding factors in myeloid leukaemia.

In parallel, the Ule Lab at the Francis Crick Institute generated individual-nucleotide resolution crosslinking and immunoprecipitation (iCLIP) data from human iPSC-derived motor neuron progenitor cells (NPCs) – generated following the protocol published by (Hall et al., 2017) – in the study of the role of Stau2 in neurogenesis.

By comparing the STAU2 eCLIP datasets from the study of leukaemia with the STAU2 iCLIP dataset from the study of neural development, I identified transcripts that are significantly bound by STAU2 in both contexts, revealing candidate genes and pathways which are conserved targets of STAU2 relevant for both areas of research.

METHODS

Datasets

This study compares two datasets of different origins and natures. On the one hand, the Stau2 eCLIP sequencing data published by Bajaj et al. (publicly available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE134971), which contains the genes that bind Stau2 in the human K562 cell line model for myeloid leukaemia (K562 dataset). On the other, the Stau2 iCLIP sequencing data from the host lab (RNA Regulatory Networks Laboratory – Francis Crick Institute), which contains the genes that bind Stau2 in human iPSC-derived motor neuron NPCs (NPC dataset). The details of the samples used are summarised in table 1.

eCLIP and iCLIP raw data analysis and processing

The processing of all the raw Stau2 sequencing data was done through the imaps (https://imaps.genialis.com/iclip) webserver, which is based on the iCount package developed by Dr. Tomaž Curk (https://icount.readthedocs.io/en/latest/index.html), for a standardised analysis. The FASTQ files were demultiplexed by experimental barcodes, the UMI positions were identified and the adapter sequences were trimmed. Reads were first pre-mapped to rRNAs and tRNAs with STAR (version 2.6.0) (Dobin et al., 2013), then subsequently aligned to the GRCh38/GENCODE v27 genome assembly human genome with options --alignEndsType Extend5pOfRead1 --outFilterMismatchNoverReadLmax 0.04. Crosslink sites were defined as the nucleotide position preceding the start of the cDNA insert (i.e. where the reverse transcription truncates) of uniquely mapping reads. cDNAs arising from PCR duplication were collapsed for sequencing reads which contain the same UMI sequence and the same position of crosslink site using the iCount xlsites function. This generated the crosslinks BED file with single-nucleotide coordinates of all crosslink sites and the counts at each site. The type and gene summary files were produced from each crosslink-site BED file to report distribution of cDNAs across primary types of RNA regions (CDS, 3'-UTR, 5'-UTR, ncRNA, intron and intergenic) and individual genes.

Building a counts matrix

The K562 dataset is formed by two replicate (REP) samples and an INPUT and the NPC dataset is formed by one REP sample and three INPUT samples. The gene and type summary files of all samples were imported formatted for analysis using RStudio v.4.0.2 (RStudio Team, 2019). From the gene summary files, the counts for each crosslink site were compiled using the dplyr package (Wickham et al., 2020) in a counts matrix, the rows and columns of which represented each gene and sample, respectively.

Studying differential expression

In order to identify enriched genes in the STAU2 dataset, which represents transcripts significantly bound by STAU2, I performed differential analysis between STAU2 ('REP') and INPUT samples in their respective datasets. This was performed by normalising and transforming the gene counts matrices using the DESEQ2 package (Love et al., 2014). The built-in rlog transformation was used to calculate the average logarithmic expression (avgLogExpr) for each gene across the samples, as well as the $\log_2$ fold change (rlogFC) between REPs and INPUTs. The gene lists in both the K562 and NPC datasets were then cropped based on their differential enrichment, considering only those whose rLogFC≥1 to be significantly bound by STAU2.

Analysing the Overlap between datasets

Using the dplyr package (Wickham et al., 2020), the DESeq2 outputs from both datasets were merged – only including the enriched genes. The list was ranked in descending order based on the average rLogFC obtained in both datasets to identify the genes that are highly bound by STAU2 in both datasets.

CLIP signal visualisation

In order to visualise the CLIP signal across all datasets, an unpublished tool was used, which uses bedgraph files for the different samples to align the CLIP signal onto an annotated human genome GTF file in a specific chromosomal region, determined using IGV (Robinson et al., 2011; Thorvaldsdóttir et al., 2013).

Gene ontology analysis

Significantly enriched genes and the overlapping data were input into GProfiler (Raudvere et al., 2019) and GORILLA (Eden et al., 2007, 2009) for functional profiling.

RESULTS

Overlapping genes

The DESeq2 tool calculated the avgLogExpr and rlogFC values for each gene in the datasets, which were plotted into an MA plot to visualise the differences between samples (Fig. 1). Initially, the K562 and NPC datasets contained 41737 and 36477 genes, respectively, from which 781 (1.87%) and 556 (1.52%) were differentially expressed in their respective cell types. From the enriched genes, 190 were found to be common to both K562 and NPC (Fig. 2). These 190 were ranked in descending order based on the sum of their rLogFC values in both datasets. The top 10 STAU2 targets (PGAM5, MRPS16, CRCP, MRPL30, METTL7A, DFFA, FCF1, NDUFC2, RRP36, RPL7L1) have their CLIP signals illustrated in figures 3-12. The crosslinking signals overlap similarly within sample types (REP and INPUT) in each dataset (K562 and NPC) and show clear differences between REP and INPUT samples. Moreover, the binding sites in the bottom section appear to overlap amongst themselves.

Binding regions

The type summary files of the K562 and NPC samples contain the counts for the nature of the RNA sequences bound by Stau2. These were used to produce figure 13, which illustrates the alignment of the binding regions of all samples. To focus on the binding of STAU2 on mRNAs, I also separately visualised the binding of STAU2 on mRNA regions, excluding ncRNAs and unannotated intergenic regions (Fig. 14). In both datasets, the REP samples show a higher incidence of 3'UTR binding sites than the INPUT.

Gene ontology

By inputting the enriched and overlapping genes in GProfiler and GORILLA, the gene ontologies were elucidated (Figs. 15 and 16). The K562 dataset appears to be dominated by sequences involved in metabolism, intracellular localization and symbiotic processes. The NPC dataset is dominated by genes regulating neural development and synaptic function. The overlapping STAU2 binding sequences were mainly involved in the regulation of elongation and termination of translation in the cytoplasm and mitochondria.

DISCUSSION

190 STAU2-binding transcripts identified

The identification of 190 common STAU2 binding sequences from the K562 and NPC datasets imply a conserved RNA-binding role by STAU2 in both cell types. Despite both cell types being 'stem cell-like', they are still very different in their expression profiles, as is reflected by the relatively few overlapping STAU2-binding transcripts (Fig. 2). However, methodological differences in the CLIP techniques used for the two datasets (K562: eCLIP, NPC: iCLIP) may have generated a slight bias. Additionally, the NPC dataset contains a single replicate sample, which does not allow for good estimation of variance in the data. To increase the robustness of this differential analysis, more experimental data should be gathered, perhaps adding an RNA-Seq dataset to resolve the issue of cell-specific gene expression. Furthermore, carrying out a similar study of STAU1 could help better understand the binding and functions of RBPs in both the leukaemia and neurogenesis contexts.

The visualisation of the CLIP signal (Figs. 3-12) showed, as expected, a similar pattern between sample types (REP vs. INPUT) and differences between them. The overlap of the binding sites suggests they are conserved in both contexts and perhaps reflects an intrinsic specificity for the binding of STAU2. Additionally, for the 10 visualised genes, the signal predominantly lies in the 3'UTR region, which is consistent with figures 13 and 14 showing an enrichment in 3'UTR binding regions in REP samples compared to INPUT for both datasets, suggesting a conserved subcellular localization in the cytoplasm and a preference for mature RNAs.

Gene ontology

The gene functions proposed by the GProfiler and GORILLA ontological analyses show significant differences based on cell type. Firstly, the enriched genes from the K562 dataset are mainly involved in metabolism, which is expected for a rapidly-growing cell population, intracellular localisation, remarking STAU2's role as an RBP, and symbiotic processes, which may be needed to interact with the immune system (Fig. 15). Secondly, the enriched genes from the NPC dataset are mainly involved in neural development and synaptic function, as expected in a neural progenitor (Fig. 16). However, the overlapping genes are mainly involved in the elongation and termination steps of translation, altogether suggesting STAU2 is involved in many different processes that are not in common between the cell types.

Unlike in Bajaj et al.'s paper, where STAU2 was identified as a regulator of chromatin-binding factors that influenced histone methylation, my analysis did not reveal enrichment of such STAU2 effectors. That could possible be attributed to the fact I focused on transcript-level enrichment, whereas a binding site-level enrichment may have increased sensitivity for STAU2 targets. Alternatively, the downstream RNA sequencing changes may not be directly caused by STAU2 binding and be an indirect consequence. This is supported by Bajaj et al.'s figure 5f, which shows a large subset of genes downregulated in RNA-Seq after STAU2 knockdown don't overlap with the CLIP data.

CONCLUSION

After comparing the K562 STAU2 eCLIP dataset from Bajaj et al.'s paper and the NPC STAU2 iCLIP dataset from the lab, 190 genes were identified as STAU2 targets based on a computational differential expression analysis. The binding sites of their transcripts have been shown to overlap between datasets, implying a conserved binding of STAU2 in both cell types, and are located on the 3'UTR of the transcripts in a small subset formed by the most enriched genes. The functional profiles of these sequences is not yet clear, but they appear to have conserved roles in the elongation and termination of translation in both the cytoplasm and the mitochondria. As more data is generated through similar experiments, possibly with the use of RNA-Seq techniques, screening for STAU2, as well as its relative STAU1, a better understanding of the role of these RBPs in the progression of leukaemia and the process of neurogenesis will be achieved, and will hopefully prove useful in the development of treatments.

ACKNOWLEDGEMENTS

I want to thank the Laidlaw Foundation for completely funding my research and for the opportunity to publish my research output in their website.

I also want to thank the Ule Lab for hosting my research project, especially Dr Jernej Ule and Dr Flora Lee for their supervision and support.

FIGURES

Sample	Dataset	Protein	Method
K562_REP1	K562	STAU2	eCLIP
K562_REP2	K562	STAU2	eCLIP
K562_INPUT	K562	INPUT	eCLIP
NPC_REP1	NPC	STAU2	iCLIP
NPC_INPUT1	NPC	INPUT	iCLIP
NPC_INPUT2	NPC	INPUT	iCLIP
NPC_INPUT3	NPC	INPUT	iCLIP

Table 1: Sample information.

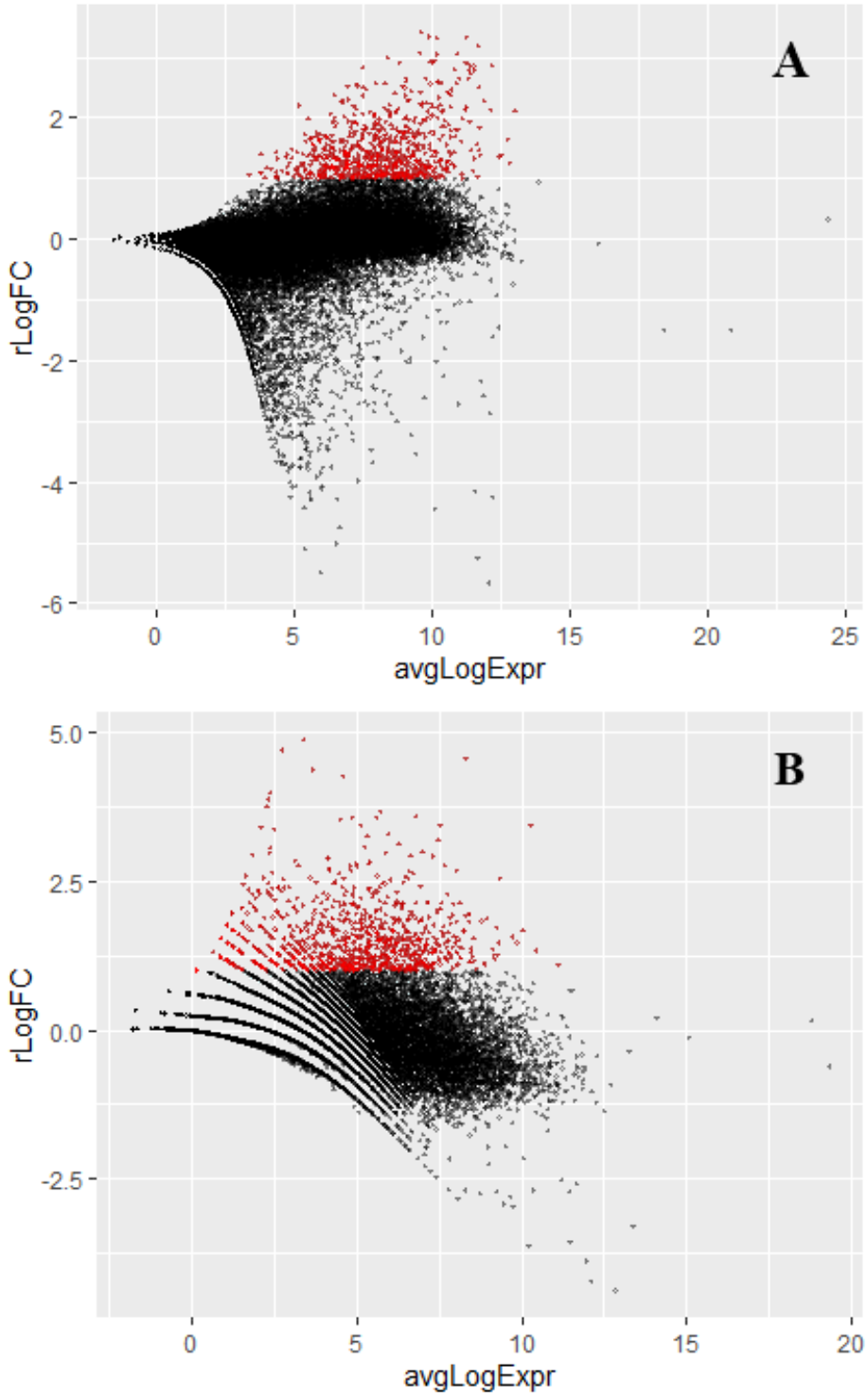


Figure 1: MA Plots illustrating the crosslinked genes in the K562 (A) and NPC (B) datasets. The red dots represent those genes that were significantly enriched (rLogFC≥1).

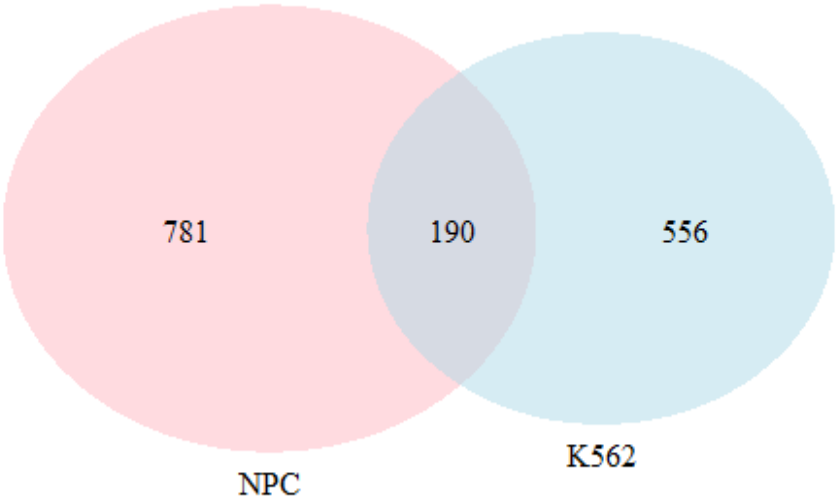
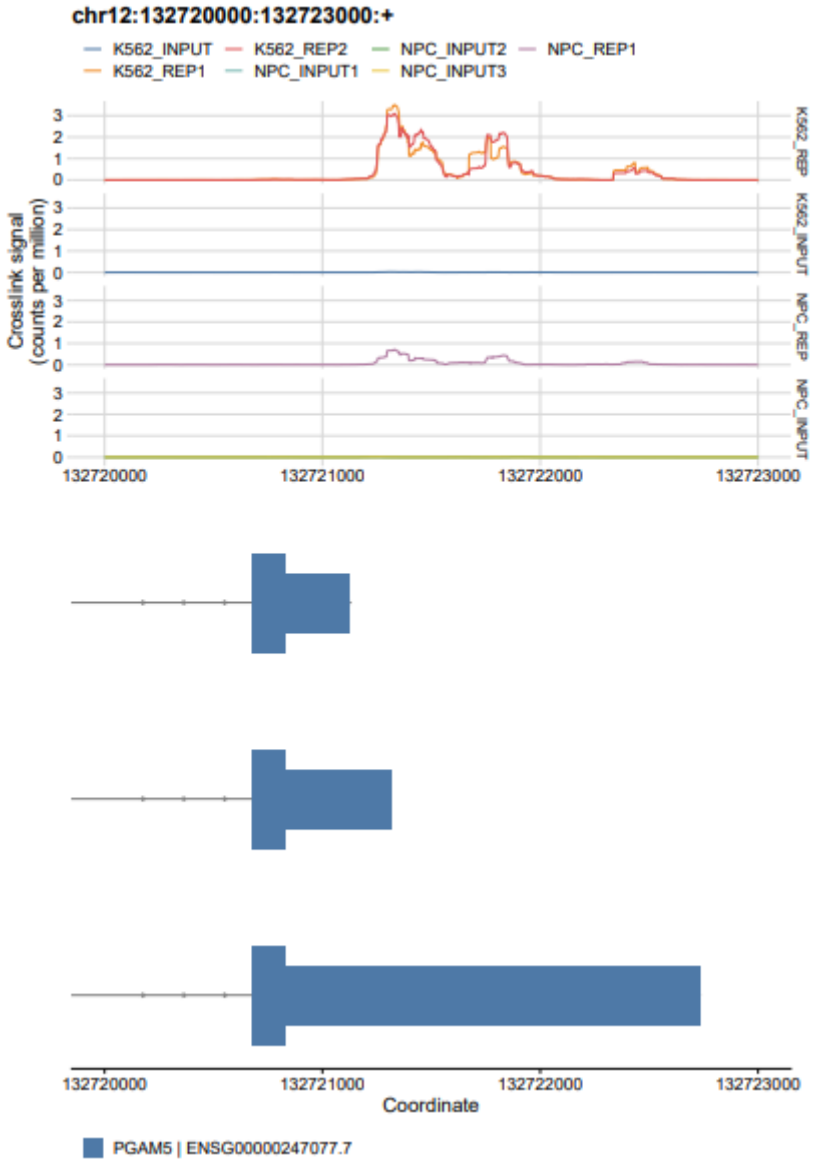
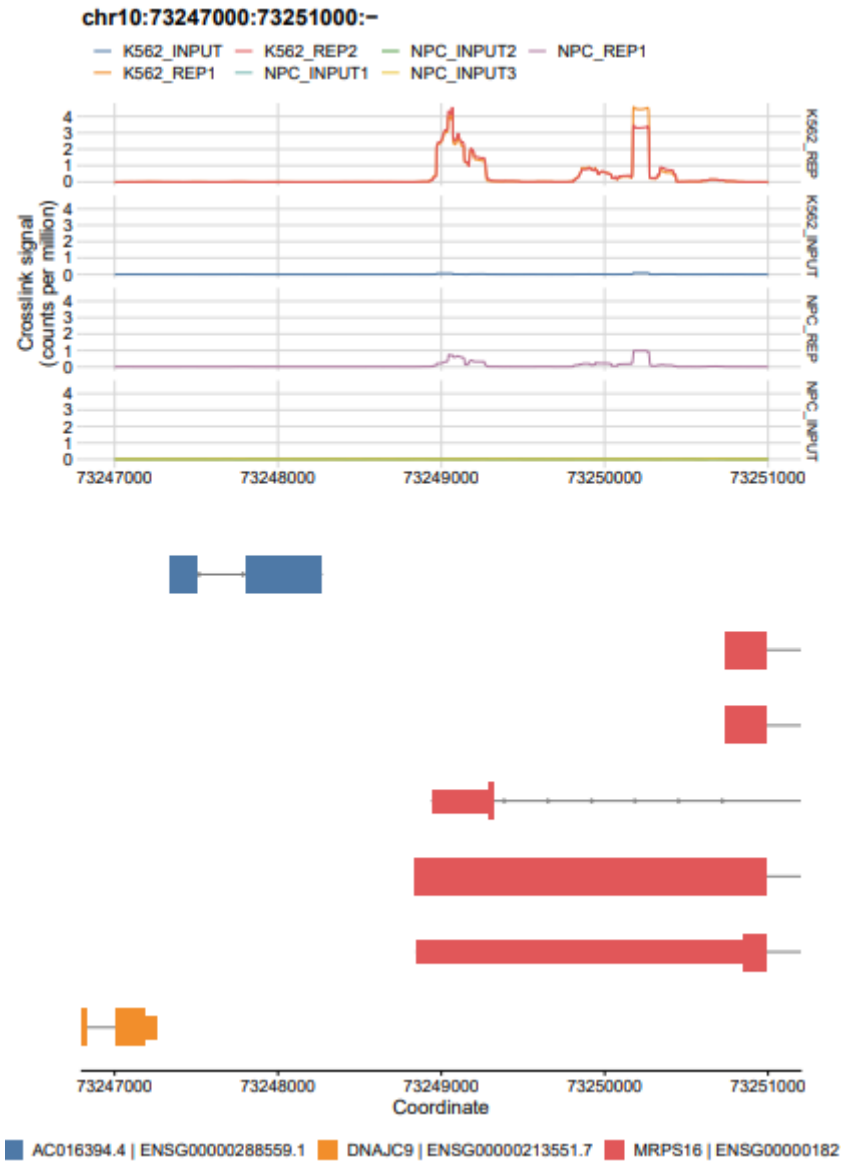
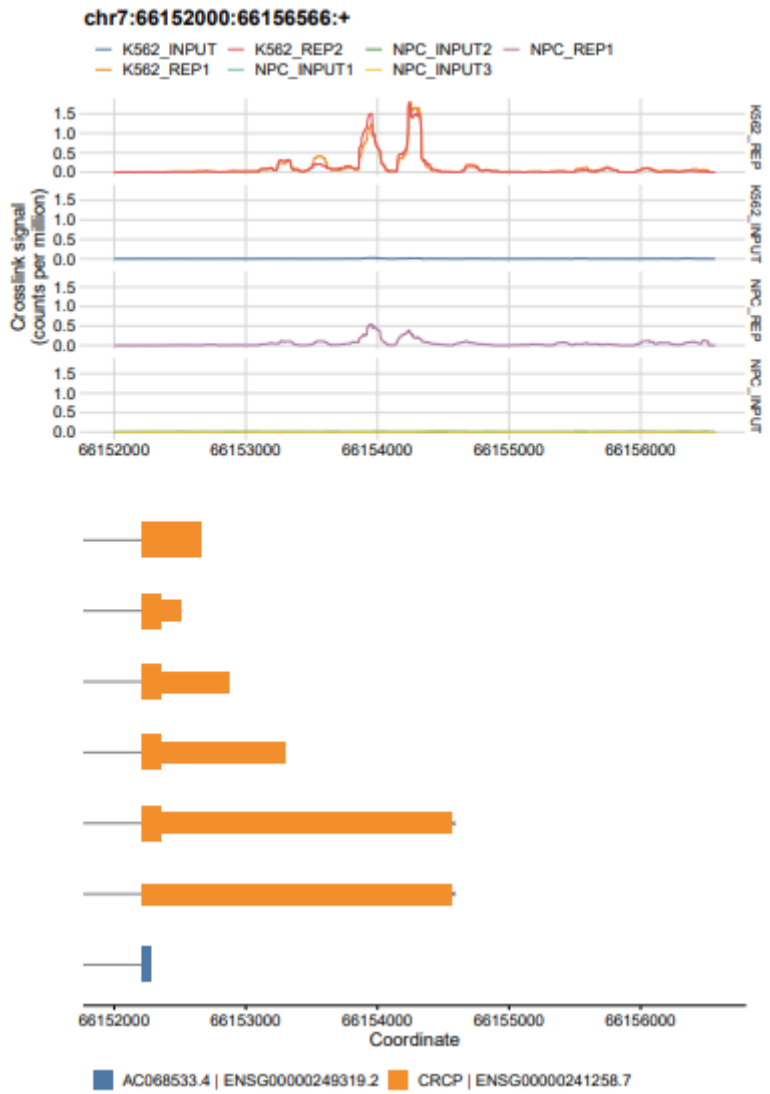
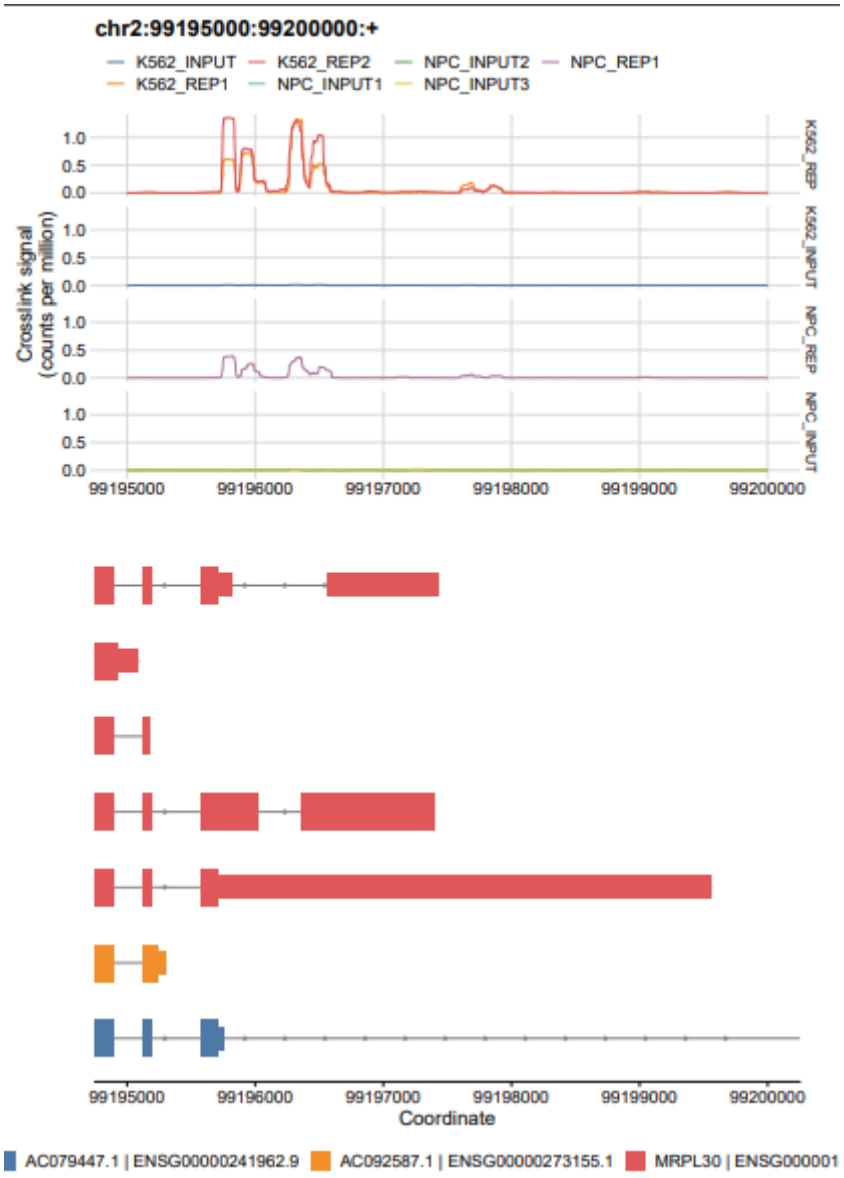


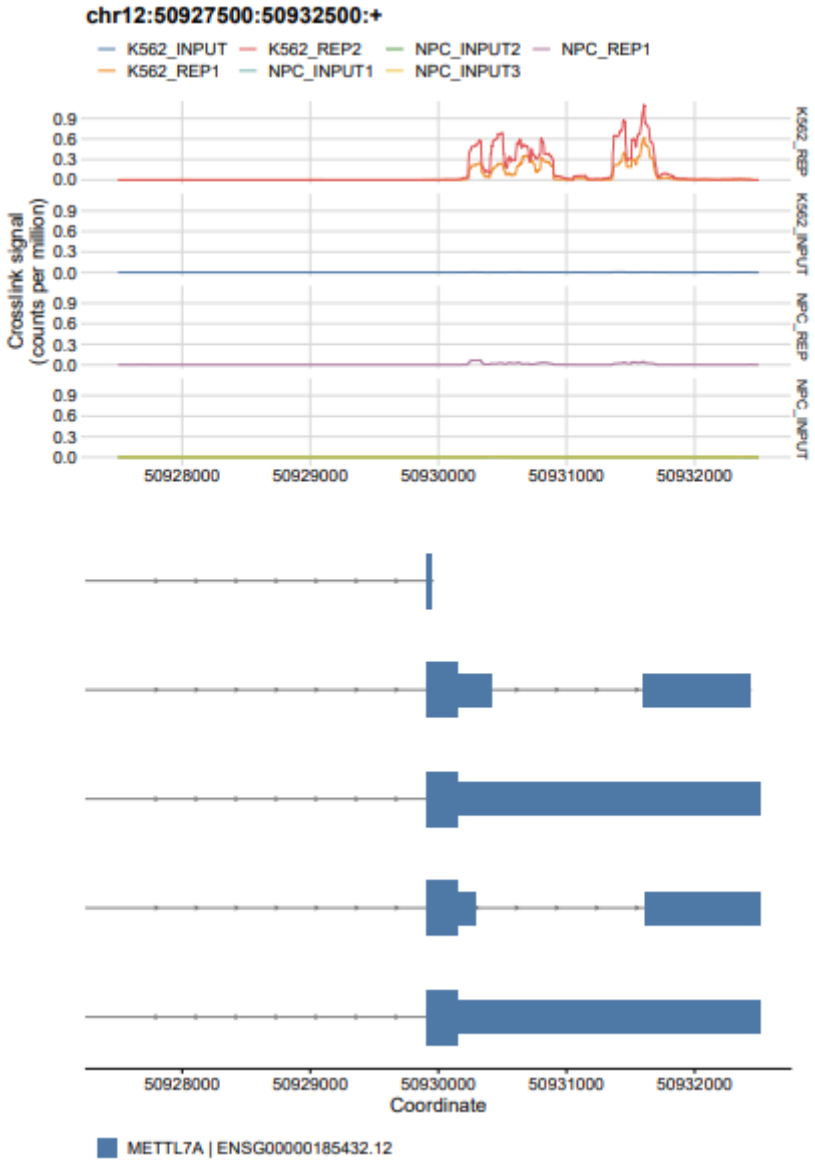
Figure 2: Venn Diagram of the K562 and NPC datasets. Each circle is sized in proportion to the number of enriched genes in the sample (noted inside the circle) and the overlapping area contains the number of genes that were significantly enriched in both samples.

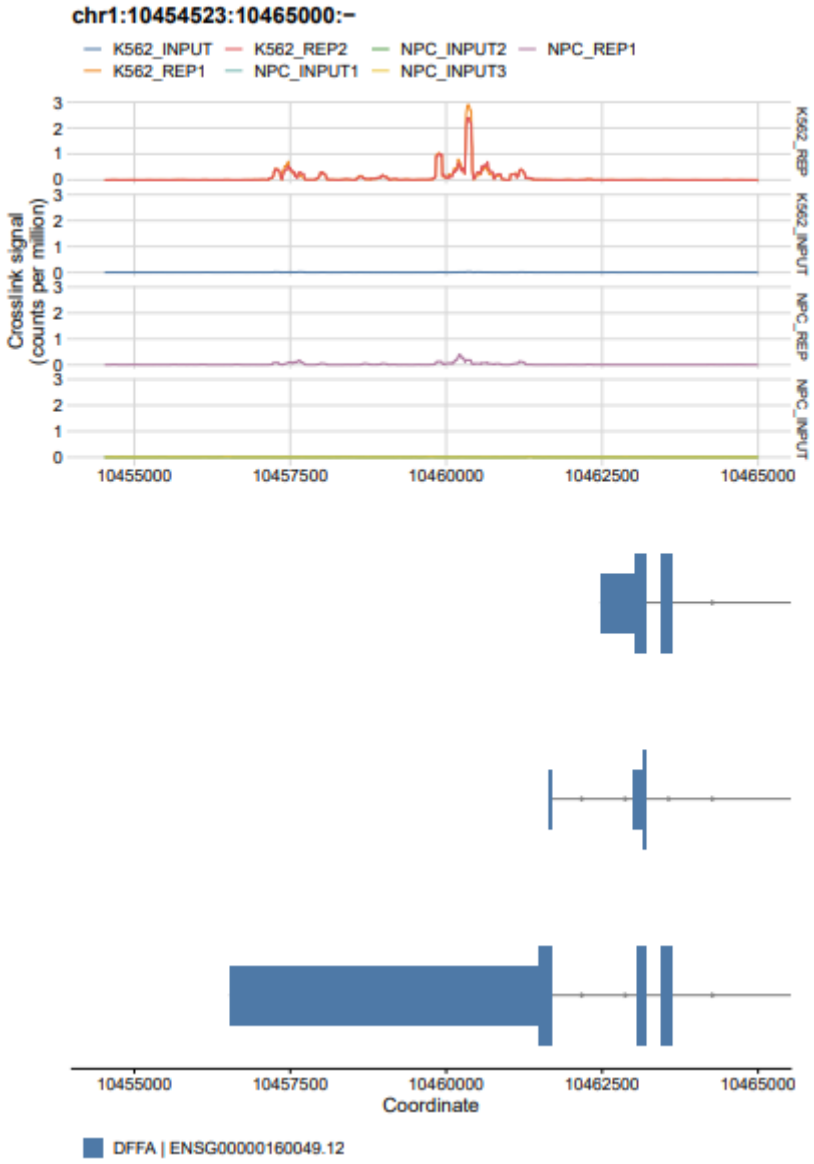

chr12:132720000:132723000:+
K562_INPUT
K562_REP2
NPC_INPUT2
NPC_REP1
K562_REP1
NPC_INPUT1
NPC_INPUT3
Crosslink signal (counts per million)
K562_REP
K562_INPUT
NPC_REP
NPC_INPUT
132720000
132721000
132722000
132723000
Coordinate
PGAM5 | ENSG00000247077.7

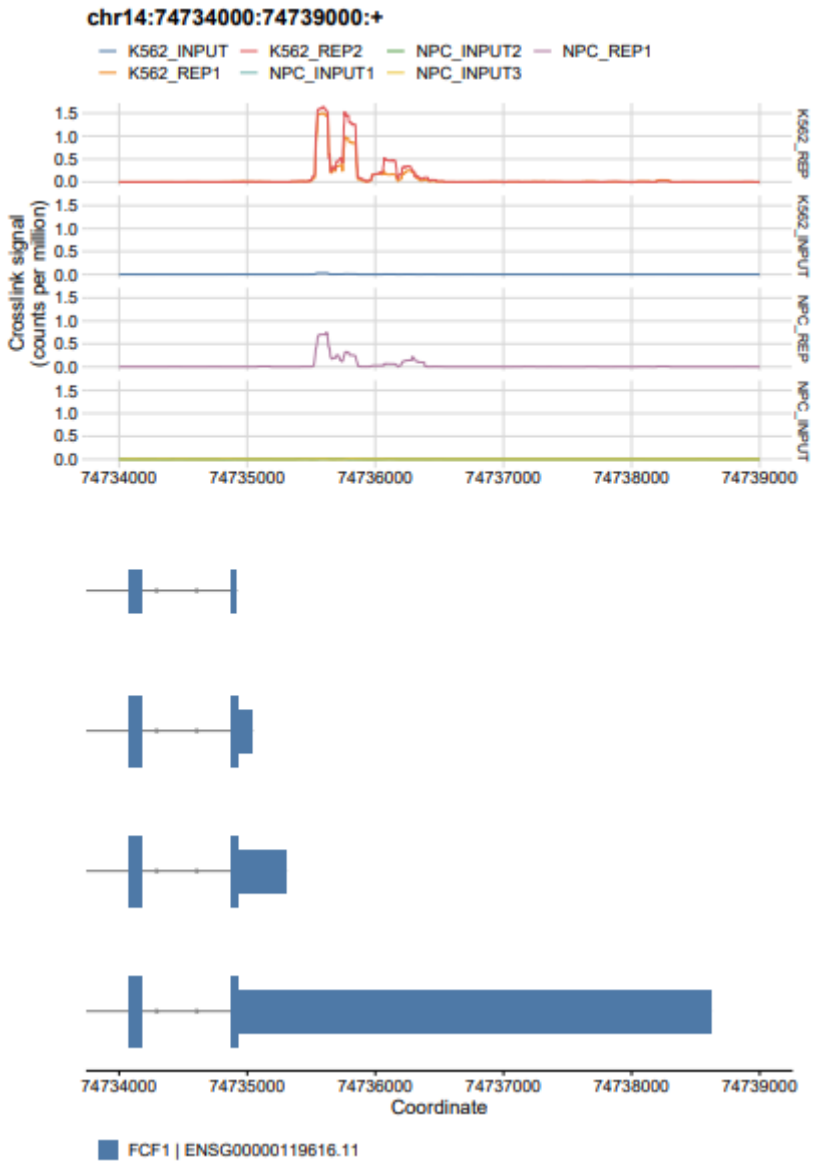

chr10:73247000:73251000:−
K562_INPUT
K562_REP2
NPC_INPUT2
NPC_REP1
K562_REP1
NPC_INPUT1
NPC_INPUT3
Crosslink signal (counts per million)
K562_REP
K562_INPUT
NPC_REP
NPC_INPUT
73247000
73248000
73249000
73250000
73251000
Coordinate
AC016394.4 | ENSG00000288559.1
DNAJC9 | ENSG00000213551.7
MRPS16 | ENSG00000182

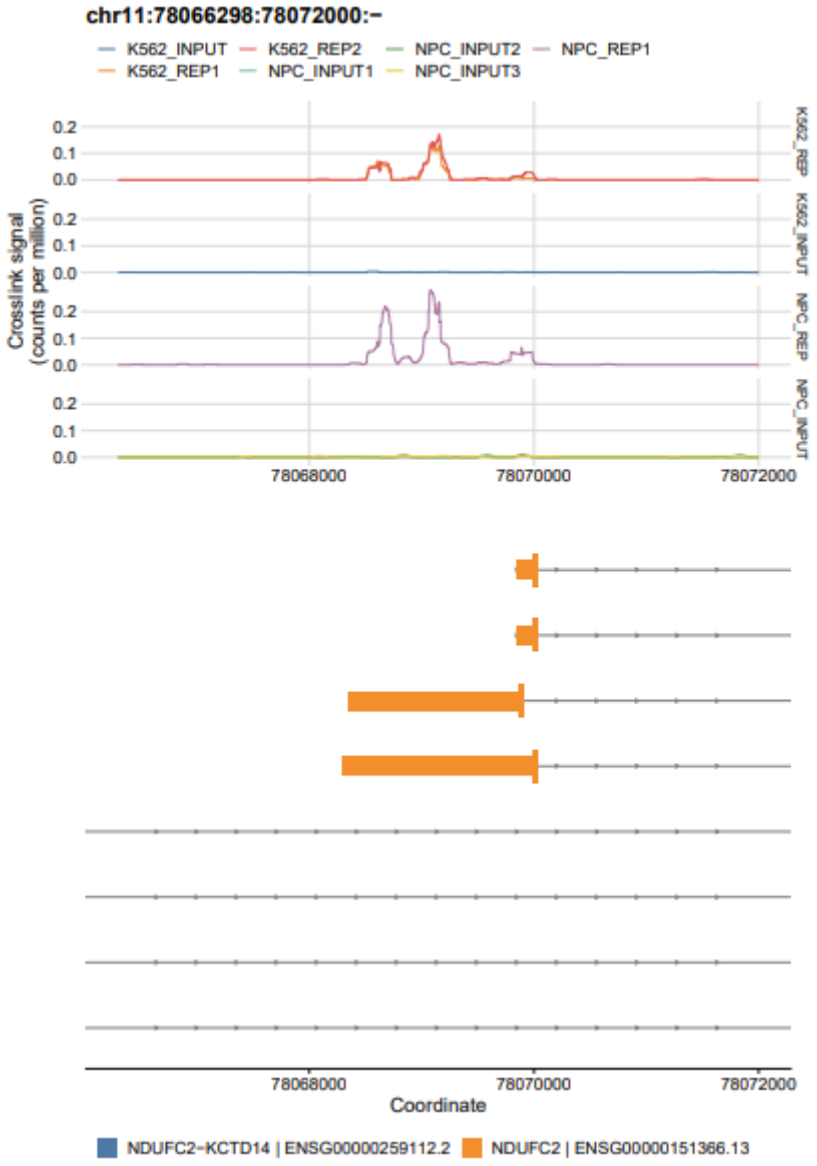

chr7:66152000:66156566:+
K562_INPUT
K562_REP2
NPC_INPUT2
NPC_REP1
K562_REP1
NPC_INPUT1
NPC_INPUT3
Crosslink signal (counts per million)
1.5
1.0
0.5
0.0
K562_REP
K562_INPUT
NPC_REP
NPC_INPUT
66152000
66153000
66154000
66155000
66156000
Coordinate
AC068533.4 | ENSG00000249319.2
CRCP | ENSG00000241258.7

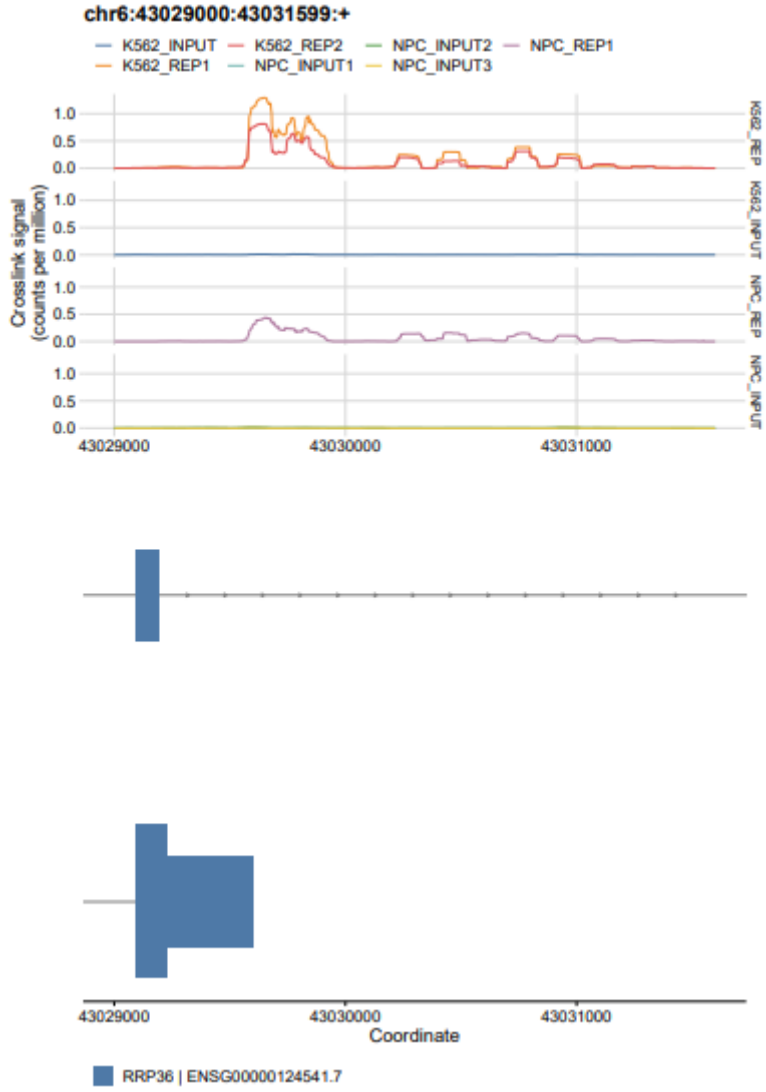

chr2:99195000:99200000:+
K562_INPUT
K562_REP2
NPC_INPUT2
NPC_REP1
K562_REP1
NPC_INPUT1
NPC_INPUT3
Crosslink signal (counts per million)
1.0
0.5
0.0
K562_REP
K562_INPUT
NPC_REP
NPC_INPUT
99195000
99196000
99197000
99198000
99199000
99200000
Coordinate
AC079447.1 | ENSG00000241962.9
AC092587.1 | ENSG00000273155.1
MRPL30 | ENSG000001

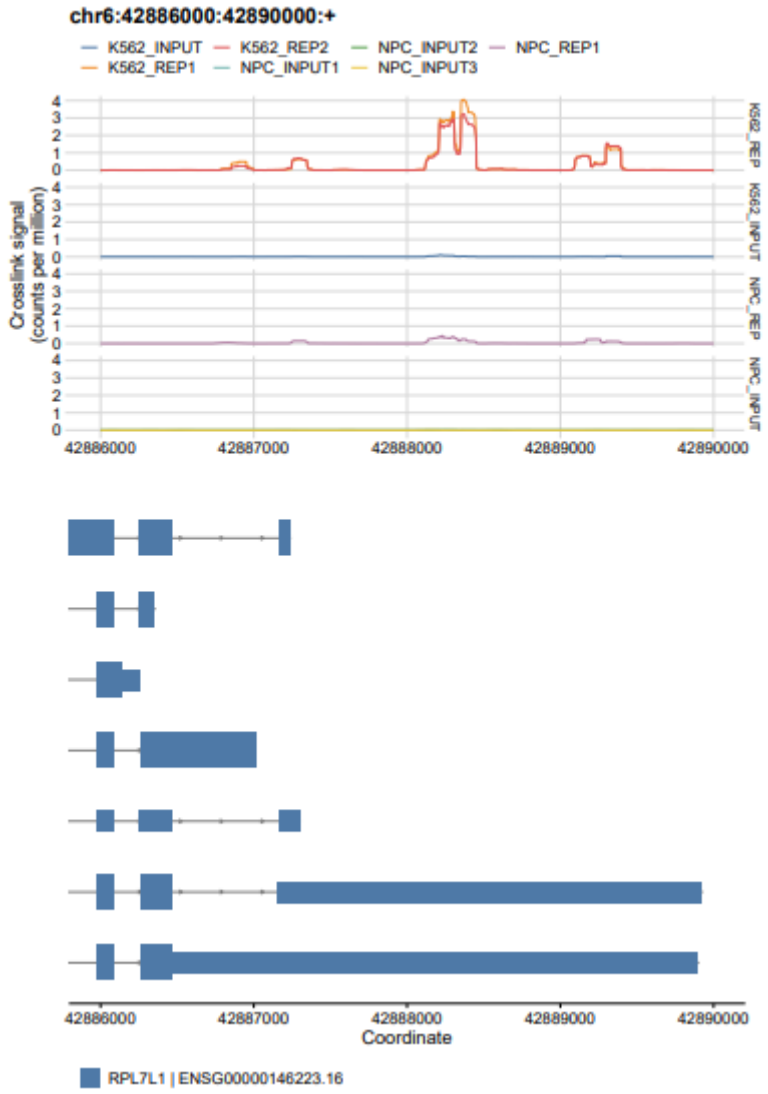

chr12:50927500:50932500:+
K562_INPUT
K562_REP2
NPC_INPUT2
NPC_REP1
K562_REP1
NPC_INPUT1
NPC_INPUT3
Crosslink signal (counts per million)
0.9
0.6
0.3
0.0
K562_REP
K562_INPUT
NPC_REP
NPC_INPUT
50928000
50929000
50930000
50931000
50932000
Coordinate
METTL7A | ENSG00000185432.12


chr1:10454523:10465000:−
K562_INPUT
K562_REP2
NPC_INPUT2
NPC_REP1
K562_REP1
NPC_INPUT1
NPC_INPUT3
Crosslink signal (counts per million)
K562_REP
K562_INPUT
NPC_REP
NPC_INPUT
10455000
10457500
10460000
10462500
10465000
Coordinate
DFFA | ENSG00000160049.12


chr14:74734000:74739000:+
K562_INPUT
K562_REP2
NPC_INPUT2
NPC_REP1
K562_REP1
NPC_INPUT1
NPC_INPUT3
Crosslink signal (counts per million)
1.5
1.0
0.5
0.0
K562_REP
K562_INPUT
NPC_REP
NPC_INPUT
74734000
74735000
74736000
74737000
74738000
74739000
Coordinate
FCF1 | ENSG00000119616.11


chr11:78066298:78072000:−
K562_INPUT
K562_REP2
NPC_INPUT2
NPC_REP1
K562_REP1
NPC_INPUT1
NPC_INPUT3
Crosslink signal (counts per million)
0.2
0.1
0.0
K562_REP
K562_INPUT
NPC_REP
NPC_INPUT
78068000
78070000
78072000
Coordinate
NDUFC2-KCTD14 | ENSG00000259112.2
NDUFC2 | ENSG00000151366.13


chr6:43029000:43031599:+
K562_INPUT
K562_REP2
NPC_INPUT2
NPC_REP1
K562_REP1
NPC_INPUT1
NPC_INPUT3
Crosslink signal (counts per million)
1.0
0.5
0.0
K562_REP
K562_INPUT
NPC_REP
NPC_INPUT
43029000
43030000
43031000
Coordinate
RRP36 | ENSG00000124541.7



Figures 3-12: Graphical representations of the CLIP signal and binding coordinates. Each figure visualizes a different genomic location containing a STAU2 target (PGAM5, MRPS16, CRCP, MRPL30, METTL7A, DFFA, FCF1, NDUFC2, RRP36, RPL7L1), which have the 10 highest rLogFC values – in descending order – combined in both datasets.

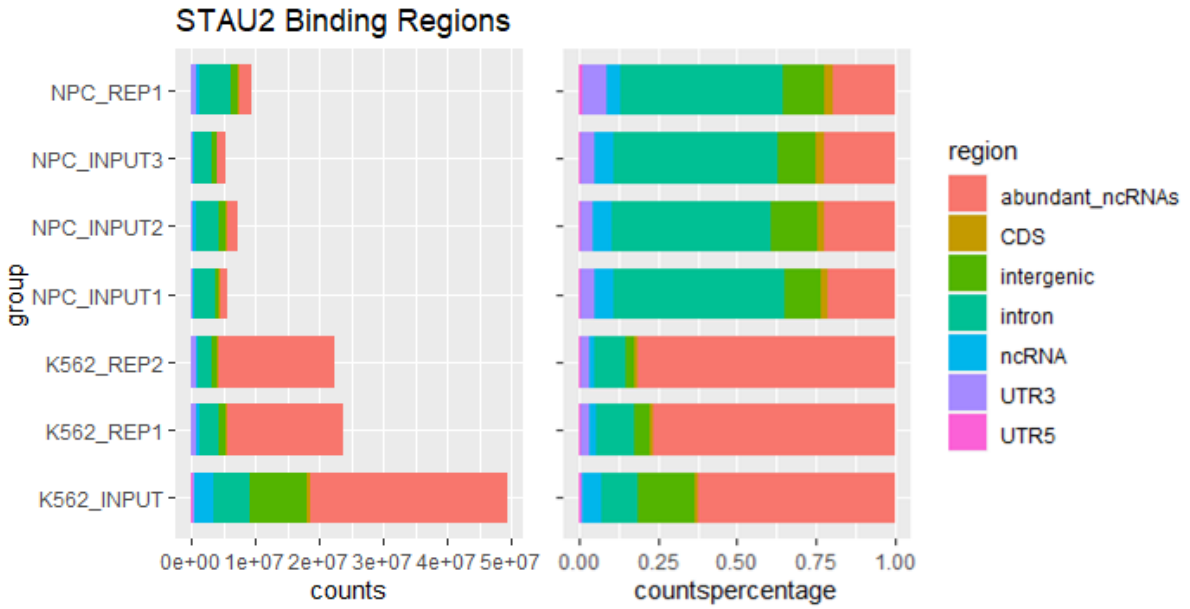


Figure 13: Aligned Stacked Bar Plots of the Stau2 Binding Regions in all samples. The left plot represents the number of counts cumulatively, and the right in percentage. Abundant non-coding RNAs (ncRNAs) are constituted by rRNA and tRNA. CDS: coding sequence.

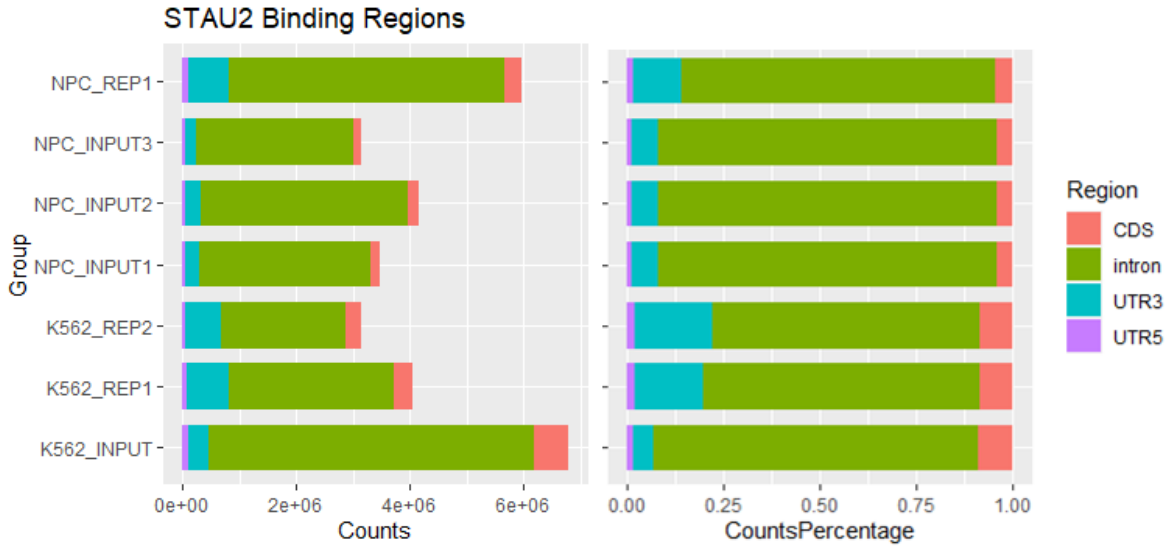
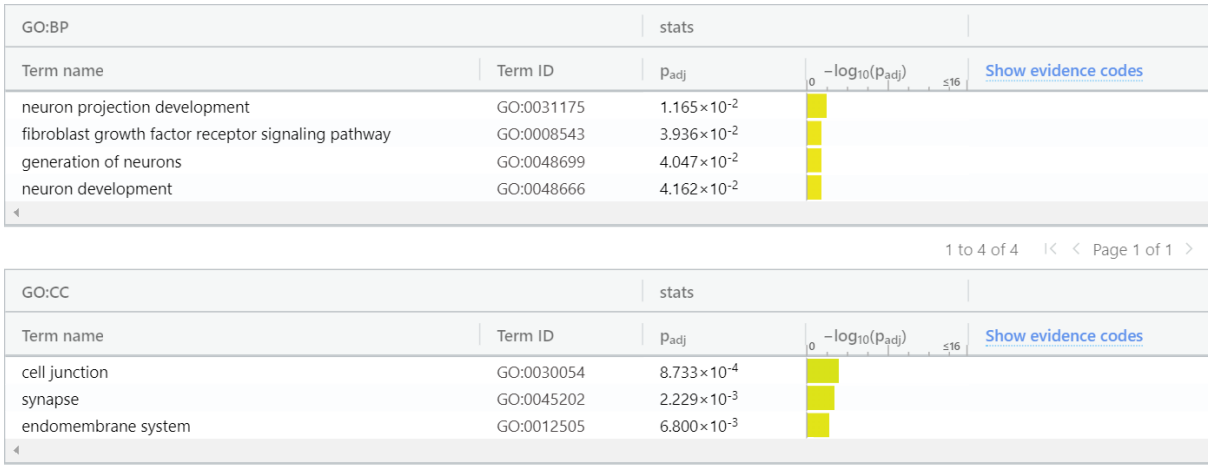


Figure 14: Refined Aligned Stacked Bar Plots of the Stau2 Binding Regions in all samples. The left plot represents the number of counts cumulatively, and the right in percentage. CDS: coding sequence.

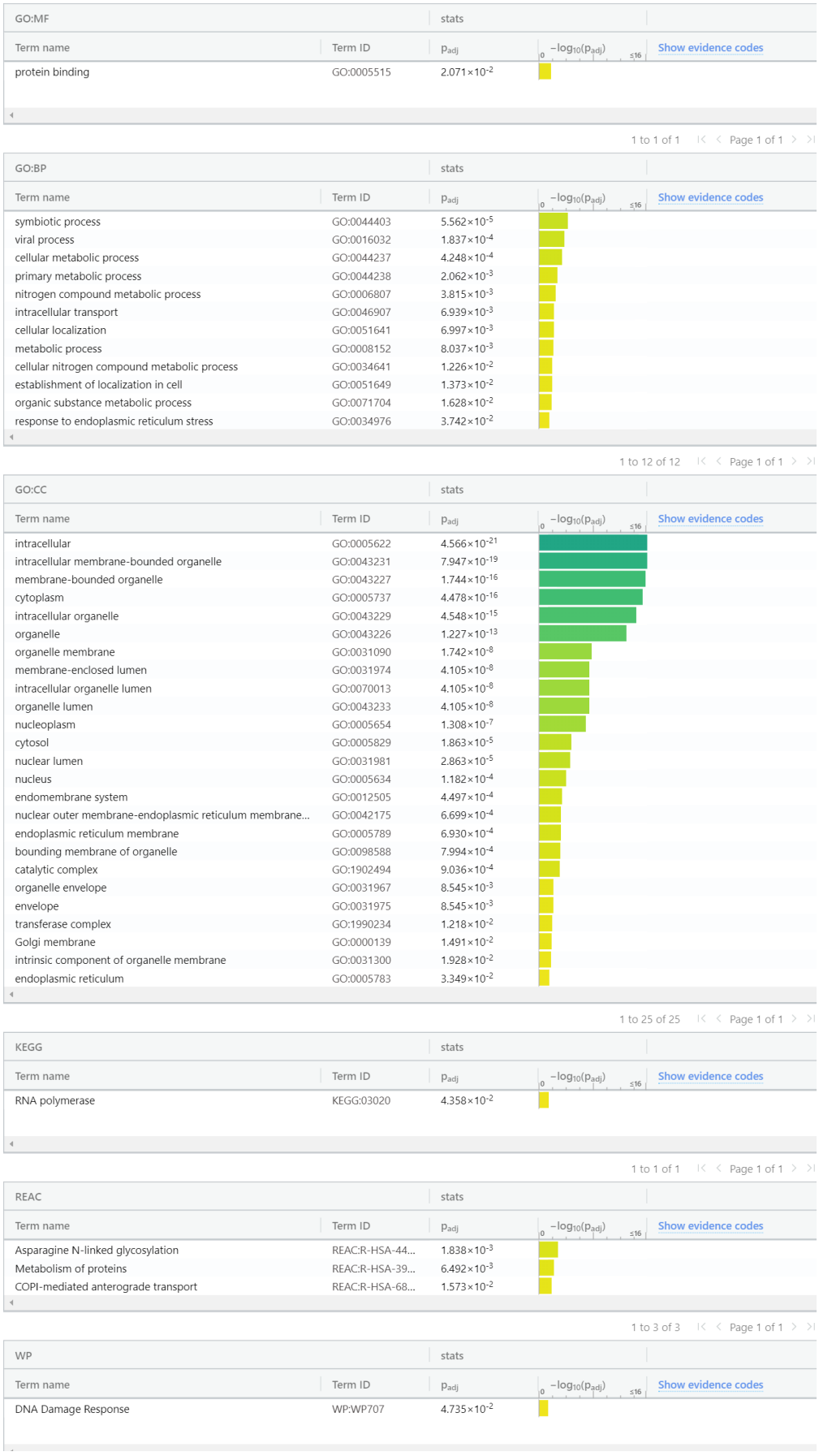


GO:BP		stats		
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16	Show evidence codes
neuron projection development	GO:0031175	1.165×10^{-2}		
fibroblast growth factor receptor signaling pathway	GO:0008543	3.936×10^{-2}		
generation of neurons	GO:0048699	4.047×10^{-2}		
neuron development	GO:0048666	4.162×10^{-2}		

1 to 4 of 4 Page 1 of 1

GO:CC		stats		
Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$ 0 ≤16	Show evidence codes
cell junction	GO:0030054	8.733×10^{-4}		
synapse	GO:0045202	2.229×10^{-3}		
endomembrane system	GO:0012505	6.800×10^{-3}		

Figure 15: GProfiler output for the NPC dataset



GO:MF

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
protein binding	GO:0005515	2.071×10^{-2}	

1 to 1 of 1 — Page 1 of 1

GO:BP

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
symbiotic process	GO:0044403	5.562×10^{-5}	
viral process	GO:0016032	1.837×10^{-4}	
cellular metabolic process	GO:0044237	4.248×10^{-4}	
primary metabolic process	GO:0044238	2.062×10^{-3}	
nitrogen compound metabolic process	GO:0006807	3.815×10^{-3}	
intracellular transport	GO:0046907	6.939×10^{-3}	
cellular localization	GO:0051641	6.997×10^{-3}	
metabolic process	GO:0008152	8.037×10^{-3}	
cellular nitrogen compound metabolic process	GO:0034641	1.226×10^{-2}	
establishment of localization in cell	GO:0051649	1.373×10^{-2}	
organic substance metabolic process	GO:0071704	1.628×10^{-2}	
response to endoplasmic reticulum stress	GO:0034976	3.742×10^{-2}	

1 to 12 of 12 — Page 1 of 1

GO:CC

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
intracellular	GO:0005622	4.566×10^{-21}	
intracellular membrane-bounded organelle	GO:0043231	7.947×10^{-19}	
membrane-bounded organelle	GO:0043227	1.744×10^{-16}	
cytoplasm	GO:0005737	4.478×10^{-16}	
intracellular organelle	GO:0043229	4.548×10^{-15}	
organelle	GO:0043226	1.227×10^{-13}	
organelle membrane	GO:0031090	1.742×10^{-8}	
membrane-enclosed lumen	GO:0031974	4.105×10^{-8}	
intracellular organelle lumen	GO:0070013	4.105×10^{-8}	
organelle lumen	GO:0043233	4.105×10^{-8}	
nucleoplasm	GO:0005654	1.308×10^{-7}	
cytosol	GO:0005829	1.863×10^{-5}	
nuclear lumen	GO:0031981	2.863×10^{-5}	
nucleus	GO:0005634	1.182×10^{-4}	
endomembrane system	GO:0012505	4.497×10^{-4}	
nuclear outer membrane-endoplasmic reticulum membrane...	GO:0042175	6.699×10^{-4}	
endoplasmic reticulum membrane	GO:0005789	6.930×10^{-4}	
bounding membrane of organelle	GO:0098588	7.994×10^{-4}	
catalytic complex	GO:1902494	9.036×10^{-4}	
organelle envelope	GO:0031967	8.545×10^{-3}	
envelope	GO:0031975	8.545×10^{-3}	
transferase complex	GO:1990234	1.218×10^{-2}	
Golgi membrane	GO:0000139	1.491×10^{-2}	
intrinsic component of organelle membrane	GO:0031300	1.928×10^{-2}	
endoplasmic reticulum	GO:0005783	3.349×10^{-2}	

1 to 25 of 25 — Page 1 of 1

KEGG

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
RNA polymerase	KEGG:03020	4.358×10^{-2}	

1 to 1 of 1 — Page 1 of 1

REAC

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
Asparagine N-linked glycosylation	REAC:R-HSA-44...	1.838×10^{-3}	
Metabolism of proteins	REAC:R-HSA-39...	6.492×10^{-3}	
COPI-mediated anterograde transport	REAC:R-HSA-68...	1.573×10^{-2}	

1 to 3 of 3 — Page 1 of 1

WP

Term name	Term ID	p_{adj}	$-\log_{10}(p_{adj})$
DNA Damage Response	WP:WP707	4.735×10^{-2}	

Figure 16: GProfiler output for the K562 dataset.

BIBLIOGRAPHY

Bajaj, J., Hamilton, M., Shima, Y., Chambers, K., Spinler, K., Van Nostrand, E. L., Yee, B. A., Blue, S. M., Chen, M., Rizzeri, D., Chuah, C., Oehler, V. G., Broome, H. E., Sasik, R., Scott-Browne, J., Rao, A., Yeo, G. W., & Reya, T. (2020). An in vivo genome-wide CRISPR screen identifies the RNA-binding protein Staufen2 as a key regulator of myeloid leukemia. *Nature Cancer*, *1*(4), 410–422. https://doi.org/10.1038/s43018-020-0054-2

Bassell, G. J., & Kelic, S. (2004). Binding proteins for mRNA localization and local translation, and their dysfunction in genetic neurological disease. *Current Opinion in Neurobiology*, *14*(5), 574–581. https://doi.org/https://doi.org/10.1016/j.conb.2004.08.010

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dubnau, J., Chiang, A.-S., Grady, L., Barditch, J., Gossweiler, S., McNeil, J., Smith, P., Buldoc, F., Scott, R., Certa, U., Broger, C., & Tully, T. (2003). The staufen/pumilio Pathway Is Involved in Drosophila Long-Term Memory. *Current Biology*, *13*(4), 286–296. https://doi.org/10.1016/S0960-9822(03)00064-2

Eden, E., Lipson, D., Yogev, S., & Yakhini, Z. (2007). Discovering Motifs in Ranked Lists of DNA Sequences. *PLOS Computational Biology*, *3*(3), e39. https://doi.org/10.1371/journal.pcbi.0030039

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, *10*(1), 48. https://doi.org/10.1186/1471-2105-10-48

Hall, C. E., Yao, Z., Choi, M., Tyzack, G. E., Serio, A., Luisier, R., Harley, J., Preza, E., Arber, C., Crisp, S. J., Watson, P. M. D., Kullmann, D. M., Abramov, A. Y., Wray, S., Burley, R., Loh, S. H. Y., Martins, L. M., Stevens, M. M., Luscombe, N. M., … Patani, R. (2017). Progressive Motor Neuron Pathology and the Role of Astrocytes in a Human Stem Cell Model of VCP-Related ALS. *Cell Reports*, *19*(9), 1739–1749. https://doi.org/10.1016/j.celrep.2017.05.024

Hentze, M. W., Castello, A., Schwarzl, T., & Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology*, *19*(5), 327–341. https://doi.org/10.1038/nrm.2017.130

Heraud-Farlow, J. E., & Kiebler, M. A. (2014). The multifunctional Staufen proteins: conserved roles from neurogenesis to synaptic plasticity. *Trends in Neurosciences*, *37*(9), 470–479. https://doi.org/10.1016/j.tins.2014.05.009

Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D., & Brown, P. O. (2008). Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biology*, *6*(10), e255–e255. https://doi.org/10.1371/journal.pbio.0060255

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., & Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, *17*(7), 909–915. https://doi.org/10.1038/nsmb.1838

Kusek, G., Campbell, M., Doyle, F., Tenenbaum, S. A., Kiebler, M., & Temple, S. (2012). Asymmetric Segregation of the Double-Stranded RNA Binding Protein Staufen2 during Mammalian Neural Stem Cell Divisions Promotes Lineage Progression. *Cell Stem Cell*, *11*(4), 505–516. https://doi.org/10.1016/j.stem.2012.06.006

Lee, F. C. Y., & Ule, J. (2018). Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Molecular Cell*, *69*(3), 354–369. https://doi.org/10.1016/j.molcel.2018.01.005

Li, P., Yang, X., Wasser, M., Cai, Y., & Chia, W. (1997). Inscuteable and Staufen Mediate Asymmetric Localization and Segregation of prosperoRNA during Drosophila Neuroblast Cell Divisions. *Cell*, *90*(3), 437–447. https://doi.org/10.1016/S0092-8674(00)80504-8

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., & Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, *456*(7221), 464–469. https://doi.org/10.1038/nature07488

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, *47*(W1), W191–W198. https://doi.org/10.1093/nar/gkz369

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M., & Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biology*, *13*(8), R67–R67. https://doi.org/10.1186/gb-2012-13-8-r67

Team, Rs. (2019). *RStudio: Integrated Development Environment for R*. RStudio, Inc. http://www.rstudio.com/

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics*, *14*(2), 178–192. https://doi.org/10.1093/bib/bbs017

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., & Darnell, R. B. (2003). CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science*, *302*(5648), 1212 LP – 1215. https://doi.org/10.1126/science.1090095

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M., & Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, *13*(6), 508–514. https://doi.org/10.1038/nmeth.3810

Wickham, H., François, R., Henry, L., & Müller, K. (2020). *dplyr: A Grammar of Data Manipulation* (R package version 1.0.2). https://cran.r-project.org/package=dplyr