Proteomics analysis on proximity labeling of AGO2 interacting proteins

August 8, 2017

Motivation

Proximity labeling is a powerful tool for revealing potential protein interaction partners. Its application has helped to expand our understanding of the molecular functions of cellular proteins, a central quest in molecular cell biology. The goal of this study is to show that biotin-labeling proteins, an engineered ascorbate peroxidase (APEX) and biotin ligase (BirA), can be employed to track both protein and RNA interactors of an RNA-binding protein. Argonaute 2 (AGO2) is a well-studied example of such protein that plays a critical role in miRNA-mediated gene silencing. By harnessing the high-throughput capability of mass-spectrometry-based proteomics for protein characterization, we hope to both confirm known and uncover novel AGO2 proximity partners.

Workflow

1. Construct HEK cells that constitutively express APEX (control), fused APEX-AGO2, BirA (control), or fused BirA-AGO2.
2. Lyse cells and pull down biotinylated proteins with streptavidin-bound magnetic beads.
    - **400 ug protein input** from BirA and BirA-AGO2 samples (bio-replicate 1)
    - **1200 ug protein input** from BirA and BirA-AGO2 samples (bio-replicate 2) and APEX and APEX-AGO2 samples (bio-replicate 1)
3. Wash beads.
    - 2X 2% SDS
    - 1X nadeoxycholate/triton/high salt/HEPES
    - 1X nadeoxycholate/np40/lithiumchlorise/tris buffer
    - 1X 50mM Tris pH 8
    - **Final: beads in 500 uL 50mM Tris**
4. Switch out Tris with lysis buffer (1M guanidium chloride, 0.1M Tris pH 8.5, 10mM TCEP, 40mM 2-CAA, 1mM $CaCl_2$, 1X HALT inhibitor).
5. Overnight digest with 2 or 3 ug trypsin, depending on protein input.
6. Desalt peptides with SOLA C18 columns.
7. Data acquisition with LC-MS/MS on a 2-hour gradient.
8. Data analysis by MaxQuant software.
    - Label-free quantification (LFQ) normalizes protein abundance measurements across samples.
    - Match-between-runs setting boosts the total number of proteins detected and quantified.
9. Data organization and visualization in R.

Results

Data Tables

The *proteinGroups* MaxQuant output containing LFQ intensity, an indirect measurement of abundance, for each protein was used for downstream statistical analyses. Please see the processed Excel file for the data tables. The contents of each table are outlined below.

- **raw_data**: Unmodified *proteinGroups* output from MaxQuant. Please see the key at the end of this report for the column name descriptions.
- **cleaned_data**: **raw_data** table with contaminants and reverse proteins removed. Added columns on $\log_2$-transformed LFQ intensities.
- **APEX_data**: Contains data associated with APEX. Removed proteins **WITHOUT ANY** quantification in either of the two groups – AGO2-APEX (*AGO2.APEX* columns) and APEX control (*ctrl.APEX* columns). Computed the mean of the logarithmized LFQ intensities for each protein in each group. This operation is trivial for the current APEX data set since there are no replicates. Last, the differences in the means, which represent fold changes in protein enrichment over control, were calculated (*AGO2.APEX - ctrl.APEX*).
  - Green background = Greater than 2-fold enrichment in the fusion protein sample over control.
  - Red background = Greater than 2-fold depletion in the fusion protein sample over control.
  - Red letters = **Known AGO2 interactors**.
- **BirA_data**: Contains data associated with BirA. Removed proteins **WITHOUT ANY** quantification in either of the two groups – AGO2-BirA (*AGO2.BirA* columns) and BirA control (*ctrl.BirA* columns). This leaves some missing values in the data table, which were imputed by assuming that the abundances are low for unquantified proteins and that their distributions are normal (see Figure 1). The logarithmized LFQ intensities were then averaged across the two biological replicates. Finally, the differences in the means, which represent fold changes in protein enrichment over control, were calculated (*AGO2.BirA - ctrl.BirA*).
  - Green background = Greater than 2-fold enrichment in the fusion protein sample over control.
  - Red background = Greater than 2-fold depletion in the fusion protein sample over control.
  - Red letters = **Known AGO2 interactors**.
  - Blue letters = **Candidate AGO2 interactors** (see Figure 6).
- **merged_data**: Joining of **APEX_data** and **BirA_data** by common *Gene.names*.
  - Dark green background = Greater than 2-fold enrichment in the fusion protein sample over control from both APEX and BirA labeling.
  - Light green background = Enrichment ($diff.mean.LOG2 > 0$) in the fusion protein sample over control from both APEX and BirA labeling.
  - Red background = Depletion ($diff.mean.LOG2 < 0$) in the fusion protein sample over control from both APEX and BirA labeling.
  - White background = Disagreements on proximity labeling between APEX and BirA approaches.
  - Red letters = **Known AGO2 interactors**.
- **AGO2_interactors**: A list of 45 AGO2 interactors annotated in the UniProt *Interaction* tab. Corresponding *UniProt ID* and *Protein names* from Swiss-Prot were included when available.

Table 1: Summary statistics on the processed samples

Samples	# protein groups identified	# proteins after valid values filtering	# missing values imputed	% imputed out of total post-filter
ctrl.APEX.bR01	1285	459	0	0
AGO2.APEX.bR01	1285	459	0	0
ctrl.BirA.bR01	1285	424	82	19.3
AGO2.BirA.bR01	1285	424	52	12.3
ctrl.BirA.bR02	1285	424	155	36.6
AGO2.BirA.bR02	1285	424	94	22.2

Data Exploration

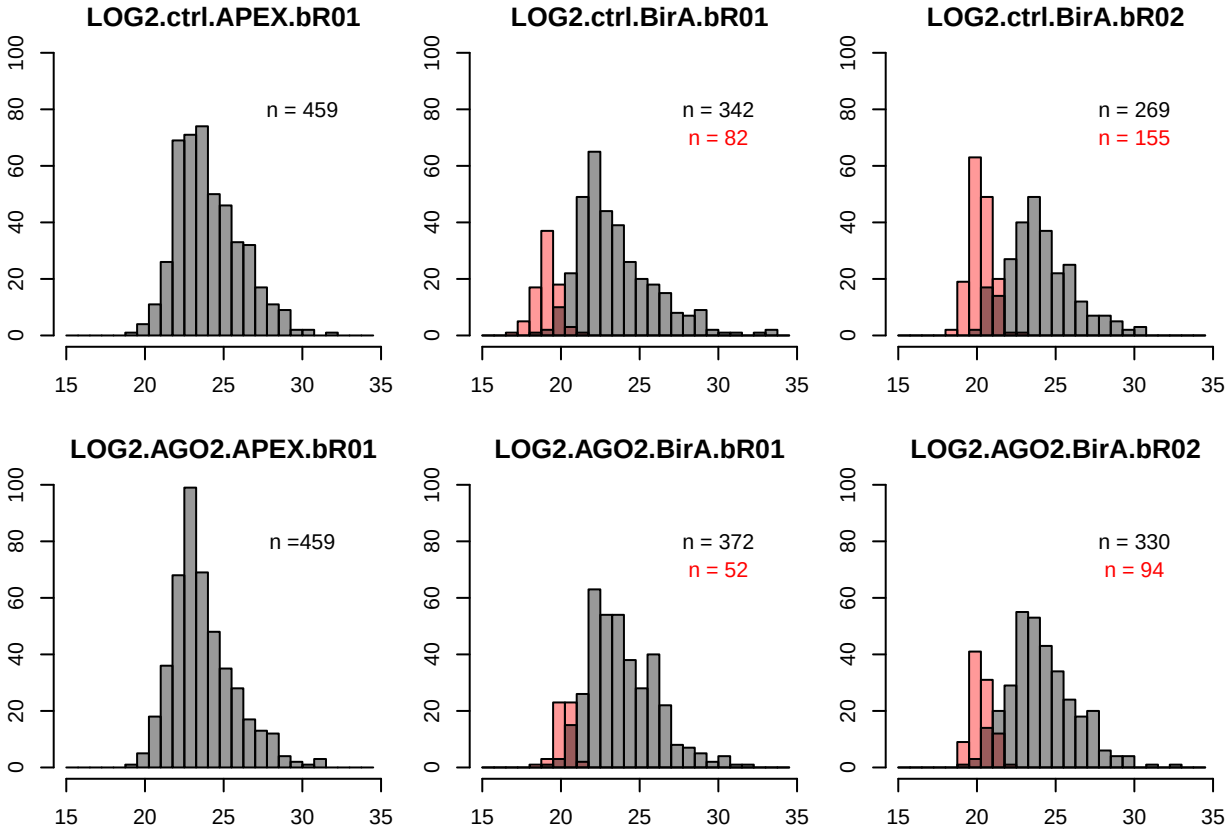

Figure 1: Histogram of logarithmized LFQ intensities by sample.

The first step in analyzing the MaxQuant *proteinGroups* output is to filter out proteins lacking quantification. In order to properly compare the protein abundances in the AGO2-tethered vs control groups, we require **at least one** quantified value (LFQ intensity > 0) in each group for each protein. APEX and BirA samples were evaluated separately in this process. For APEX, this means that the protein must be quantified across the board since only one experiment was performed. Admittedly, this is a conservative approach as proteins enriched in the AGO2 sample but lacking measurements in the control, which should qualify as potential candidates, are ignored. With two replicates of the BirA experiment, we encounter a different set of challenges after data filtering. The rules above capture proteins lacking quantification in some cases, which hinders downstream analyses. A standard imputation method was implemented to fill the gaps by randomly drawing from a normal distribution with a spread and center defined on a per-sample basis over the $\log_2$-transformed LFQ intensities: 1) the standard deviation (SD) of the distribution is 0.3 times the sample SD and 2) its mean is the difference between the sample mean and 1.8 times the sample SD. Low-value, sample-based assignments to missing elements reflect the scanty nature of proteins detected yet unquantified and accounts for the variability in sample preparation and mass spectrometery analysis from one run to the next.

The distributions of $\log_2$-transformed LFQ after filtering and imputation are summarized in Figure 1 and Table 1. Black represents proteins quantified by MaxQuant, and red represents proteins whose values were imputed. As expected, a larger fraction of values were imputed in the control group compared to the AGO2 group. Notably, a greater proportion of proteins were imputed in the second BirA biological replicate compared to the first. This considerably changes the overall shapes of distributions away from normal. Alternatively, a more stringent filter requiring quantification in all channels could displace the need for imputation (but at a cost of coverage as only 169 proteins would remain vs 424).

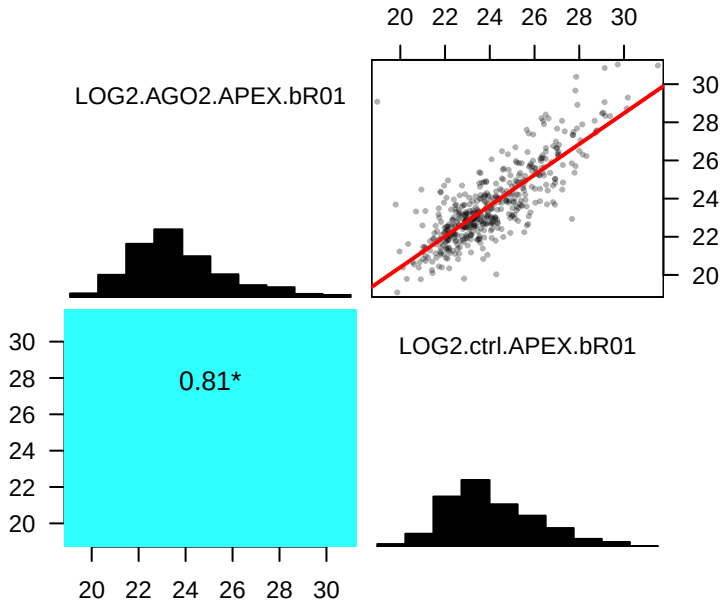

Figure 2: Correlation between APEX sample runs after data filtering.

Following data cleaning and organization, comparisons between samples are now possible. Figure 2 is a pairs plot displaying a linear fit (red line) and Pearson correlation between the protein abundances in the AGO2 (y-axis) and control (x-axis) group for the APEX experiment. Given that the APEX control provides a measure of background labeling, proteins falling closer to the upper-left corner in the scatter plot are of interest as they show greater enrichment in the AGO2-APEX sample than in the control. This is explored further in Figure 4.

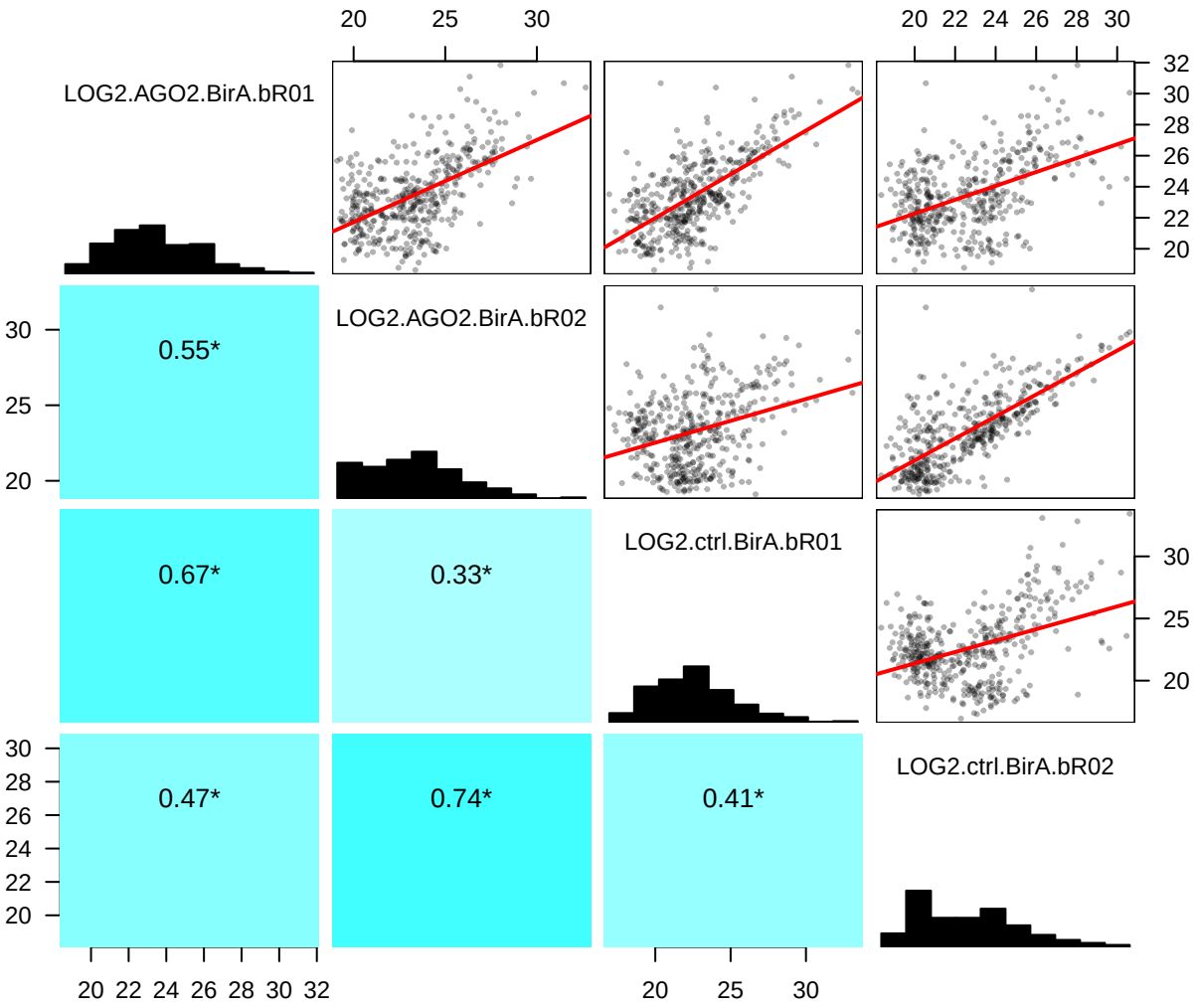

Figure 3: Correlations between BirA sample runs after data filtering and imputation.

Figure 3 shows an analogous pairs plot for the BirA experiment. A difference in the strengths of the correlations between the APEX and BirA data set is obvious – the linear trends are less defined between BirA samples. This is likely an artifact of randomly assigning low-value measurements to missing data. Interestingly, the strongest correlation is observed between samples from the same replicate rather than from the same group. It is unclear whether this is due to variabilities from one sample preparation/analysis to the next or the effects of imputation. What is evident is that any interpretations of the BirA data should be made with caution.

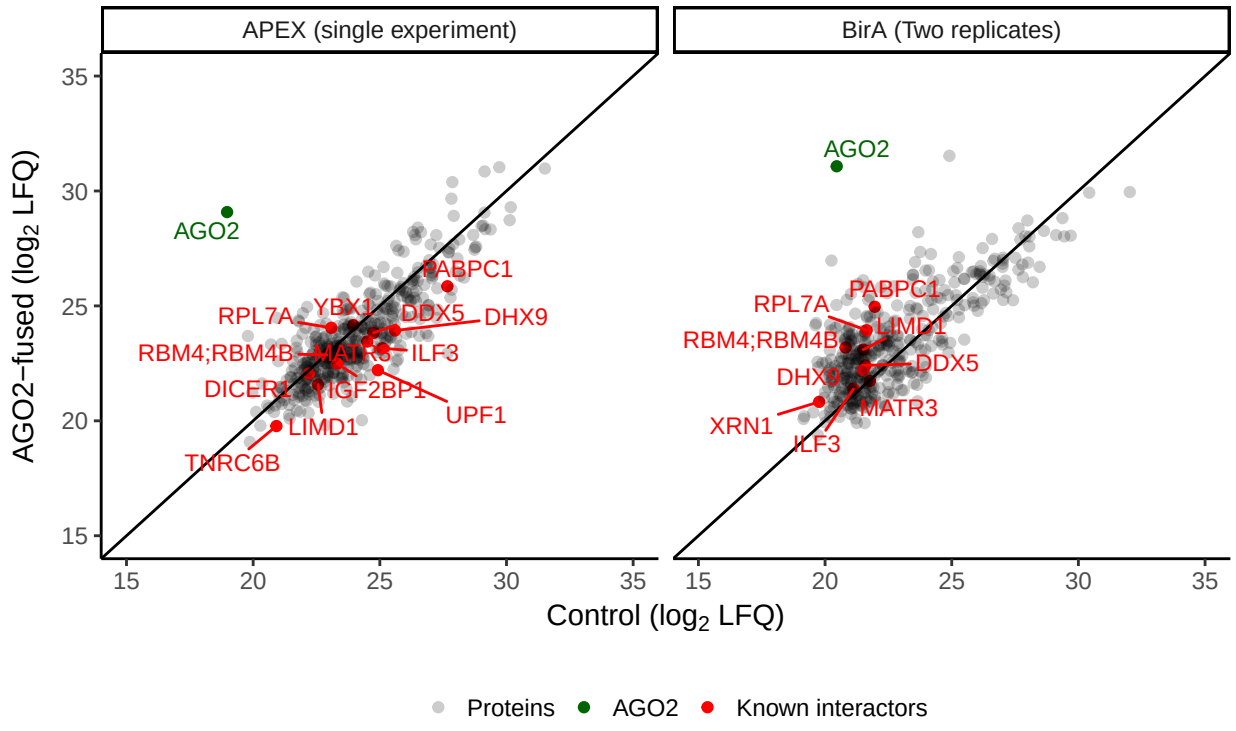

Figure 4: Scatter plot of the mean $\log_2$ LFQ for proteins labeled in AGO2 vs control group.

A more informative way of examining the BirA data is to combine the results for the two groups across replicates. To this end, the mean of the $\log_2$-transformed LFQs was computed for each protein in each group. The results are shown in Figure 4 alongside the APEX scatter. A line of unity was placed for reference. Known interactors are marked in red. The list of interactors used in the search was derived from UniProt annotations (see **merge_data** table).

The protein AGO2 is highly enriched in both APEX and BirA, nicely validating the experimental design. In BirA labeling, all known interactors except for one (MATR1) are more highly enriched in the AGO2 group than the control. The same case cannot be made for APEX as the majority in red are found below the reference line. Additionally, the weight of the points appears to be flipped for BirA and APEX with a majority of the proteins located above and below the diagonal, respectively. Perhaps more replicates of APEX could refine the signal and shift the known interactors above the line.

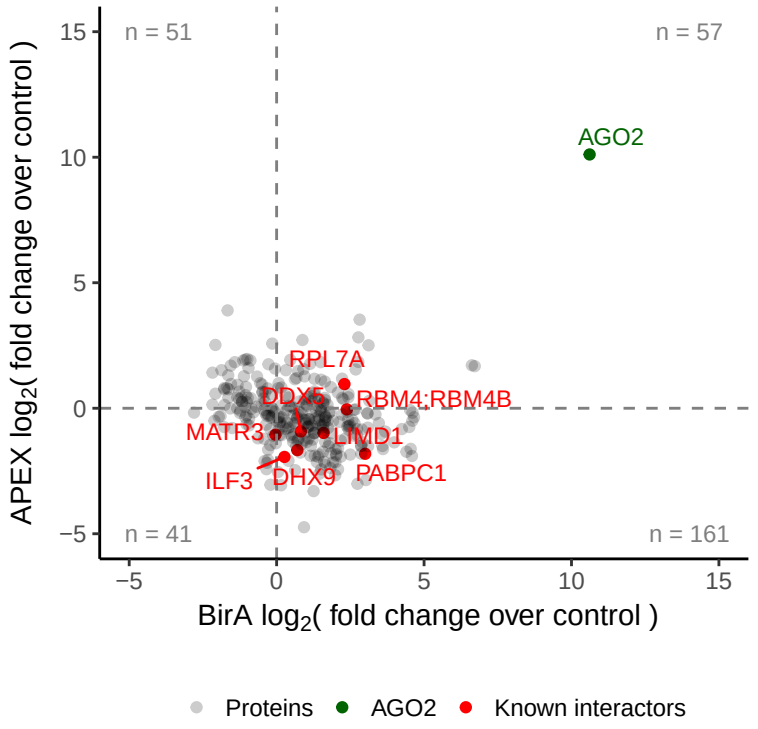

Figure 5: Comparison of protein enrichment in APEX vs BirA labeling.

APEX and BirA are biotin-labeling enzymes that mark neighboring proteins via distinct mechanisms. In other words, they are designed to achieve the same goal in their own ways. Therefore, we hypothesize that proteins enriched under one experimental set-up, APEX or BirA, should also be enriched under the other; the same should apply for depleted proteins. Figure 5 depicts the effect of the labeling strategy on protein enrichment. Each point represents a protein present in both APEX and BirA samples. If our hypothesis were true, all the points would lie in either the upper-right or lower-left quadrants. However, over half of the proteins show a lack of consensus in enrichment between the two labeling techniques with a majority residing in the lower-right quadrant, recapitulating the findings from Figure 4 that a bulk of the proteins are enriched in the BirA experiment but depleted when labeled with APEX. Nonetheless, these results should be taken with a grain of salt as we are comparing experiments with different numbers of replicates followed data manipulation in unique ways.

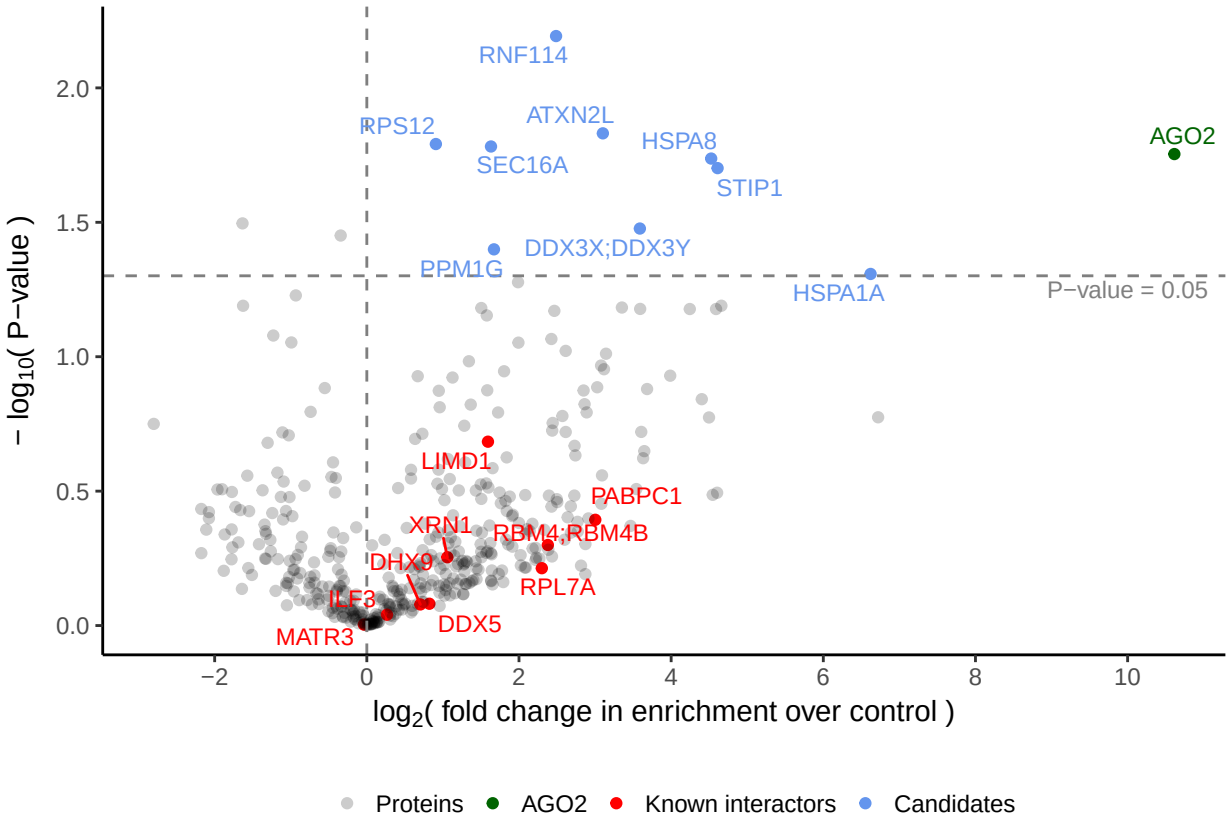

Figure 6: Volcano plot of $\log_2$ LFQ differences between AGO2-BirA and BirA control labeling.

Finally, the goal of this study is to identify novel candidates associated with AGO2 *in vivo*. A volcano plot is useful for this purpose as it graphically displays biological and statistical significance on the x-axis and y-axis, respectively. More specifically, proteins are arranged along the horizontal axis based on fold change in protein abundance over non-specific, background labeling and the vertical axis based on statistical significance levels from sample t-tests on biological replicates. Figure 6 shows such representation on the BirA data set. The asymmetric form of the scatter is reassuring and supports the thinking that AGO2-BirA labeling has greater specificity for select proteins over indiscriminate biotinylation by BirA alone. In addition, the threshold for candidacy (in blue) are set at a P-value less than 0.05 and a fold change greater than 1 (or $\log_2$ fold change $>$ 0). It is reasonable to think that these putative interactors appear significant only by chance, possibly raised to the top by exceptionally low imputed values in the controls. However, these proteins, surprisingly, were fully quantified, indicating that the basis of their significance is biological. One final note – the fact that known interactors are excluded by the prescribed cutoff suggests that many potential candidates may be overlooked. Please refer to the **BirA_data** table for a complete list of positively enriched proteins.

Bottom Line

- Nine AGO2-interacting candidates were identified by statistical tests on the BirA data.
- AGO2 was detected and highly enriched in both the APEX and BirA experiments, serving as a successful positive control.
- More replicates of the APEX would increase statistical power and enable comparisons with BirA labeling.
- The current requirement for data filtering eliminates proteins without any quantification in the control. While it narrows the data set to proteins with more confident signals, many potential candidates are missed as a result.
- The **AGO2_interactors** table used to search for known interactors is not exhaustive as one of the identified candidates (DDX3) has been shown to interact with AGO2 in literature but was not described in the UniProt annotations.

Protein groups

The Protein Groups table contains information on the identified proteins in the processed raw-files. Each single row contains the group of proteins that could be reconstructed from a set of peptides.

Name	Separator	Description
Protein IDs		Identifier(s) of protein(s) contained in the protein group. They are sorted by number of identified peptides in descending order.
Majority protein IDs		These are the IDs of those proteins that have at least half of the peptides that the leading protein has.
Peptide counts (all)		Number of peptides associated with each protein in protein group, occuring in the order as the protein IDs occur in the 'Protein IDs' column. Here distinct peptide sequences are counted. Modified forms or different charges are counted as one peptide.
Peptide counts (razor+unique)		Number of peptides associated with each protein in protein group, occuring in the order as the protein IDs occur in the 'Protein IDs' column. Here distinct peptide sequences are counted. Modified forms or different charges are counted as one peptide.
Peptide counts (unique)		Number of peptides associated with each protein in protein group, occuring in the order as the protein IDs occur in the 'Protein IDs' column. Here distinct peptide sequences are counted. Modified forms or different charges are counted as one peptide.
Protein names		Name(s) of protein(s) contained within the group.
Gene names		Name(s) of the gene(s) associated to the protein(s) contained within the group.
Fasta headers		Fasta headers(s) of protein(s) contained within the group.
Number of proteins		Number of proteins contained within the group. This corresponds to the number of entries in the colum 'Protein IDs'.
Peptides		The total number of peptide sequences associated with the protein group (i.e. for all the proteins in the group).
Razor + unique peptides		The total number of razor + unique peptides associated with the protein group (i.e. these peptides are shared with another protein group).
Unique peptides		The total number of unique peptides associated with the protein group (i.e. these peptides are not shared with another protein group).
Peptides AGO2-APEX		Number of peptides (distinct peptide sequences) in experiment AGO2-APEX
Peptides AGO2-BirA		Number of peptides (distinct peptide sequences) in experiment AGO2-BirA
Peptides ctrl APEX		Number of peptides (distinct peptide sequences) in experiment ctrl APEX
Peptides ctrl BirA		Number of peptides (distinct peptide sequences) in experiment ctrl BirA
Razor + unique peptides AGO2-APEX		Number of razor + unique peptides (distinct peptide sequences) in experiment AGO2-APEX
Razor + unique peptides AGO2-BirA		Number of razor + unique peptides (distinct peptide sequences) in experiment AGO2-BirA
Razor + unique peptides ctrl APEX		Number of razor + unique peptides (distinct peptide sequences) in experiment ctrl APEX
Razor + unique peptides ctrl BirA		Number of razor + unique peptides (distinct peptide sequences) in experiment ctrl BirA
Unique peptides AGO2-APEX		Number of unique peptides (distinct peptide sequences) in experiment AGO2-APEX
Unique peptides AGO2-BirA		Number of unique peptides (distinct peptide sequences) in experiment AGO2-BirA
Unique peptides ctrl APEX		Number of unique peptides (distinct peptide sequences) in experiment ctrl APEX
Unique peptides ctrl BirA		Number of unique peptides (distinct peptide sequences) in experiment ctrl BirA
Sequence coverage [%]		Percentage of the sequence that is covered by the identified peptides of the best protein sequence contained in the group.
Unique + razor sequence coverage [%]		Percentage of the sequence that is covered by the identified unique and razor peptides of the best protein sequence contained in the group.
Unique sequence coverage [%]		Percentage of the sequence that is covered by the identified unique peptides of the best protein sequence contained in the group.
Mol. weight [kDa]		Molecular weight of the leading protein sequence contained in the protein group.

Sequence length		The length of the leading protein sequence contained in the group.
Sequence lengths		The length of all sequences of the proteins contained in the group.
Q-value		This is the ratio of reverse to forward protein groups.
Identification type AGO2-APEX		Indicates whether this experiment was identified by MS/MS or only by matching between runs.
Identification type AGO2-BirA		Indicates whether this experiment was identified by MS/MS or only by matching between runs.
Identification type ctrl APEX		Indicates whether this experiment was identified by MS/MS or only by matching between runs.
Identification type ctrl BirA		Indicates whether this experiment was identified by MS/MS or only by matching between runs.
Sequence coverage AGO2-APEX [%]		Percentage of the sequence that is covered by the identified peptides in this sample of the longest protein sequence contained within the group.
Sequence coverage AGO2-BirA [%]		Percentage of the sequence that is covered by the identified peptides in this sample of the longest protein sequence contained within the group.
Sequence coverage ctrl APEX [%]		Percentage of the sequence that is covered by the identified peptides in this sample of the longest protein sequence contained within the group.
Sequence coverage ctrl BirA [%]		Percentage of the sequence that is covered by the identified peptides in this sample of the longest protein sequence contained within the group.
Intensity		Summed up eXtracted Ion Current (XIC) of all isotopic clusters associated with the identified AA sequence. In case of a labeled experiment this is the total intensity of all the isotopic patterns in the label cluster.
Intensity AGO2-APEX		Summed up eXtracted Ion Current (XIC) of all isotopic clusters associated with the identified AA sequence. In case of a labeled experiment this is the total intensity of all the isotopic patterns in the label cluster.
Intensity AGO2-BirA		Summed up eXtracted Ion Current (XIC) of all isotopic clusters associated with the identified AA sequence. In case of a labeled experiment this is the total intensity of all the isotopic patterns in the label cluster.
Intensity ctrl APEX		Summed up eXtracted Ion Current (XIC) of all isotopic clusters associated with the identified AA sequence. In case of a labeled experiment this is the total intensity of all the isotopic patterns in the label cluster.
Intensity ctrl BirA		Summed up eXtracted Ion Current (XIC) of all isotopic clusters associated with the identified AA sequence. In case of a labeled experiment this is the total intensity of all the isotopic patterns in the label cluster.
LFQ intensity AGO2-APEX		
LFQ intensity AGO2-BirA		
LFQ intensity ctrl APEX		
LFQ intensity ctrl BirA		
MS/MS Count AGO2-APEX		
MS/MS Count AGO2-BirA		
MS/MS Count ctrl APEX		
MS/MS Count ctrl BirA		
MS/MS Count		
Only identified by site		When marked with '+', this particular protein group was identified only by a modification site.
Reverse		When marked with '+', this particular protein group contains no protein, made up of at least 50% of the peptides of the leading protein, with a peptide derived from the reversed part of the decoy database. These should be removed for further data analysis. The 50% rule is in place to prevent spurious protein hits to erroneously flag the protein group as reverse.
Potential contaminant		When marked with '+', this particular protein group was found to be a commonly occurring contaminant. These should be removed for further data analysis.
id		A unique (consecutive) identifier for each row in the proteinGroups table, which is used to cross-link the information in this file with the information stored in the other files.
Peptide IDs		Identifier(s) of the associated peptide sequence(s) summary, which can be found in the file 'peptides.txt'.
Peptide is razor		Indicates for each peptide ID if it is a razor or group unique peptide (true) or a non unique non razor peptide (false).
Mod. peptide IDs		
Evidence IDs		
MS/MS IDs		
Best MS/MS		The identifier of the best (in terms of quality) MS/MS scans identifying the peptides of this protein, referenced against the msms table.

Oxidation (M) site IDs		Identifier(s) for site(s) associated with the protein group, which show(s) evidence of the modification, referenced against the appropriate modification site file.
Oxidation (M) site positions		Positions of the sites in the leading protein of this group.