Supplementary Document for "*Analysis of Protein Complexes via Model-based Biclustering of Label-free Quantitative AP-MS Data*"

Table of Contents

Search Parameters for PP2A Dataset

In short, mzXML files were generated from ThermoFinnigan *.RAW files using the ReAdW tool (http://tools.proteomecenter.org/software.php). The protein sequence database contained Human IPI fasta sequence file (version 3.38), common contaminants, and appended with a reversed version of the same IPI database. mzXML files were searched with X!Tandem/ k-score (Craig and Beavis, 2004; MacLean *et al*, 2006) using the following parameters. For Goudreault *et al* (2009): b- and y-ion series; partial trypsin digestion, allowing one missed cleavage site, methionine oxidation and N-terminal acetylation specified as variable modifications. The fragment mass tolerance was 0.8 Da (monoisotopic mass), and the mass window for the precursor was from -1 to 4 Da (average mass). For Glatter *et al* (2009): b- and y-ion series; trypsin digestion, allowing for two missed cleavage sites, methionine oxidation specified as a variable modification, and cysteine alkylation as a fixed modification. The fragment mass tolerance was 0.8 Da (monoisotopic mass), and the mass window for the precursor was from -3 to 3 Da (monoisotopic mass). In the refinement mode, the analysis was extended to semi-tryptic peptides, and allowing phosphorylation of S, T, and Y residues.

Prior Elicitation for DPM Model

Specifying the base distribution of $\{\theta_k\}_{k=1}^{K}$ as a hierarchical Dirichlet process is equivalent to projecting the initial multivariate probability distribution into a collection of *p* correlated univariate probability distributions. The advantage of this simplification is that exactly the same model parameters $\{\theta_k\}_{k=1}^{K}$ are shared across different preys, and subsequently so are the clustering labels indicating abundance levels.

For the base distribution in the TIP49a/b dataset, we specified $H_0 \sim N(0,10^2) \times IG(1,1)$, the product of a Gaussian prior for the mean and an inverse Gamma prior for the variance. Since all scaled data vary between 0 and 40 in the two datasets, one standard deviation of the base distribution covers the range of the entire data. Variations on the hyperprior for the mean parameter had nearly no impact on the posterior distributions, while those for the variance parameter heavily influenced the mixture model estimation in this dataset. Shape 1 and scale 1 for the Gamma distribution was chosen to allow every mixture component to cover sufficient range of values after observing the scaled NSAF values. We varied the scale parameter from 1 to 10, but this did not affect the analysis outcome. The DPM model concentration parameters were set at $\alpha=\rho=\gamma=1$ for these data.

In the PP2A dataset, we set the hyperpriors at $H_0 \sim N(5,10^2) \times IG(1,10)$, which targets the high abundance range more than the low abundance range and results in larger variance than in the previous dataset. In addition, we set α=25, ρ=1000, *and* γ=15 to reduce the parsimony imposed by the model. This specification was intended to circumvent the possibility that the homogeneity within STRIPAK complex can drive the nonparametric mixture model extremely concentrated in a small number of mean and variance values.

Sampling-based Inference with Markov chain Monte Carlo (MCMC)

The proposed method performs statistical estimation and inference by repeatedly drawing model parameters from the appropriate posterior distribution (50,000 iterations and adjustable depending on convergence diagnostic). A Gibbs sampling algorithm for the stick-breaking construction of DPM has already been used widely (Ishwaran and James, 2001; Sethuraman, 1994). The reservoir of sample bait clusters and nested prey clusters provides a rich source for statistical inference. One can reference the final clustering outcome by comparing it to the distribution across all samples. Here, the bait cluster configuration is selected as the one that maximizes the posterior probability, also known as *maximum a posteriori* (*MAP*) estimate. However, a given clustering outcome may not be the only solution with a dominating posterior probability score and there can be many other competing models with similarly high probabilities. To examine this, inter-bait distances were computed from the posterior sample models as follows. For bait *i* and *j*, let c_i and c_j denote their respective cluster labels. Then the inter-bait distance is calculated by

$$P(c_i \neq c_j) = \frac{\sum_{t=1}^{T} 1(\hat{c}_i^{(t)} \neq \hat{c}_j^{(t)})}{T}$$

where $\hat{c}_i^{(t)}$ and $\hat{c}_j^{(t)}$ are the posterior samples for baits *i* and *j* at iteration *t* of the sampler, and *T* is the number of iterations in MCMC. This distance measure gives the assessment of tightness between members of a protein complex (bait cluster), and is partially related to the robustness of the *MAP* estimate. In turn, this measure is conceptually equivalent to the opposite of clustering coefficient.

Comparison to Other Clustering Methods

The nested clustering algorithm was compared with a representative hierarchical clustering method in PP2A dataset (application of hierarchical clustering to TIP49a/b dataset was extensively discussed in (Sardiu *et al*, 2009)). When agglomerative hierarchical clustering was

applied to the prey data with Jaccard and Euclidean distances on binary and quantitative data respectively, the STRIPAK complex was clearly separated from other proteins in both cases, and the regulatory subunits PPP2R2 (A,B,C,D) were widely dispersed in the dendrogram (prey direction; see Supplementary Figures 3A and 3B). However, when using Jaccard distance, hierarchical clustering placed the CCT complex components under the same branch as the catalytic subunit PPP2CA and the scaffolding subunit PPP2R1A. This is likely an incorrect clustering since the result heavily relies on the interaction between the CCT components and a fraction of STRIPAK complex (see above). The clustering outcome became more congruent with the known biology when using quantitative data with the Euclidean distance. The chaperones formed a cluster independent from other proteins and only a single core member (CTTNBP2) of STRIPAK complex was separated from the rest of the complex, which happened because CTTNBP2 exhibited most interactions when used as bait but not as prey (likely due to a lower endogenous abundance as compared to the CTTNBP2NL paralog). Nevertheless, it was difficult to interpret the result for the catalytic, scaffolding, and regulatory subunits of PP2A based on this hierarchical clustering dendrogram (using either binary or Euclidean distances) because they were placed next to completely unrelated preys with no common patterns in spectral counts. This once again illustrates the limitation of clustering in one dimensional space (partitions and tree-structures) for accurate protein complex assembly. The catalytic, scaffolding, and regulatory subunits of PP2A showed a small number of interactions as preys, whereas they had many interactions when used as baits. Since hierarchical clustering computes the distance between prey proteins, the more complete bait-wise interaction data for these proteins were not utilized in full. In other words, when a prey has interactions only with a relatively small number of the bait proteins in the dataset, hierarchical clustering for that prey is driven more by the absence of interactions with other prey proteins than by the presence of interactions with the baits.

The comparison with the existing biclustering methods was performed using biclust R package (http://cran.r-project.org). These methods include BiMax (Prelic *et al*, 2006), Cheng and Church (CC) (Cheng and Church, 2000), and PLAID (Lazzeroni and Owen, 2002). BiMax performs biclustering by searching for submatrices of ones in a binary data matrix. Here binary data are equivalent to the one in which any positive spectral count is considered as one and all zeros are considered as zeros. Cheng and Church method (CC) report submatrices with a type of mean squared error lower than a specified threshold (delta) in standardized data. PLAID model explains the data as a sum of multiple layers of mean values and identifies submatrices by

minimizing a standard loss function, allowing identification of overlapping submatrices. Note that these methods may identify a huge number of submatrices since the number of possible submatrices increases in a combinatorial way as the number of baits and preys grows. Hence the analysis reported in the main text was performed based on up to 16 highest scoring biclusters in each of the three methods.

In TIP49a/b dataset, the biclusters reported by Bimax and CC methods did not match any of the four proteins complexes clearly. On the other hand, PLAID was the only method that recovered hINO80 complex and the subcluster shared by SRCAP and TRRAP/TIP60 complexes, which also coincides with the complexes identified by nested clustering. However, PLAID reported three biclusters at most and failed to identify the Prefoldin and TRRAP complexes (see Supplementary Figure 4). The three biclustering methods used in TIP49a/b dataset were applied to this dataset for comparison as well. Both BiMax and CC algorithms largely failed to recover meaningful biclusters in top scoring solutions. Of the three methods, the PLAID model produced results most congruent with the nested clustering. For instance, the biclusters reported from the PLAID model included the set of interactions between two core subunits of PP2A and other regulatory subunits (Supplementary Figure 5). However, the model reported only two biclusters as statistically significant and failed to identify the large STRIPAK complex, which renders the overall interpretation of the data unclear.

References

Cheng Y, Church GM (2000) Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* **8:** 93-103.

Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **20:** 1466-1467.

Glatter T, Wepf A, Aebersold R, Gstaiger M (2009) An integrated workflow for charting the human interaction proteome: insights into the PP2A system. *Mol Syst Biol* **5:** 237.

Goudreault M, D'Ambrosio LM, Kean MJ, Mullin MJ, Larsen BG, Sanchez A, Chaudhry S, Chen GI, Sicheri F, Nesvizhskii AI, Aebersold R, Raught B, Gingras AC (2009) A PP2A phosphatase high density interaction network identifies a novel striatin-interacting phosphatase and kinase complex linked to the cerebral cavernous malformation 3 (CCM3) protein. *Mol Cell Proteomics* **8:** 157-171.

Ishwaran H, James LF (2001) Gibbs sampling methods for stick-breaking priors. *J Am Statist Assoc* **96:** 161-173.

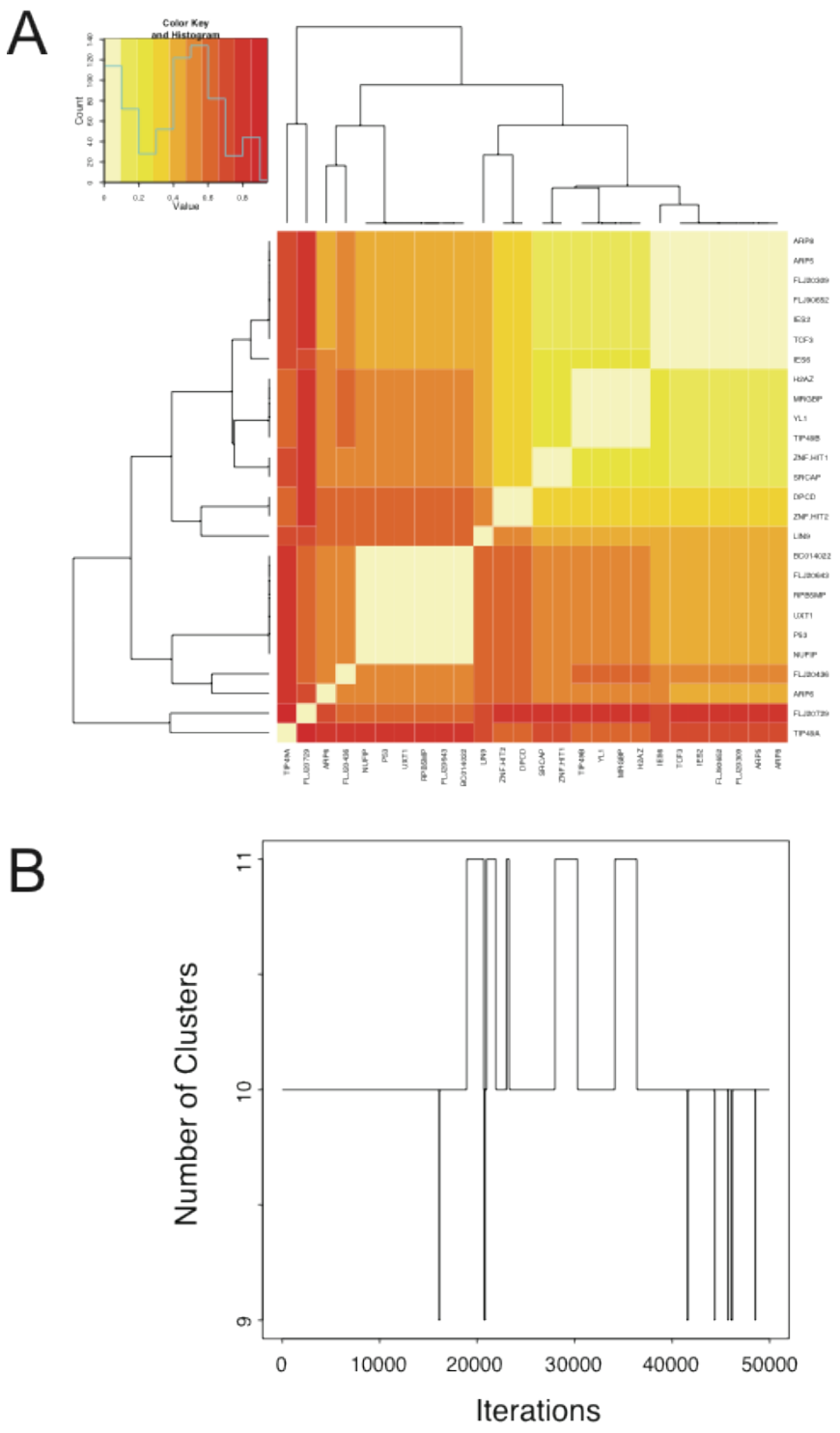
Lazzeroni L, Owen A (2002) Plaid Models for Gene Expression Data. *Stat Sinica* **12:** 61-86.

MacLean B, Eng JK, Beavis RC, McIntosh M (2006) General framework for developing and evaluating database scoring algorithms using the TANDEM search engine. *Bioinformatics* **22:** 2830-2832.

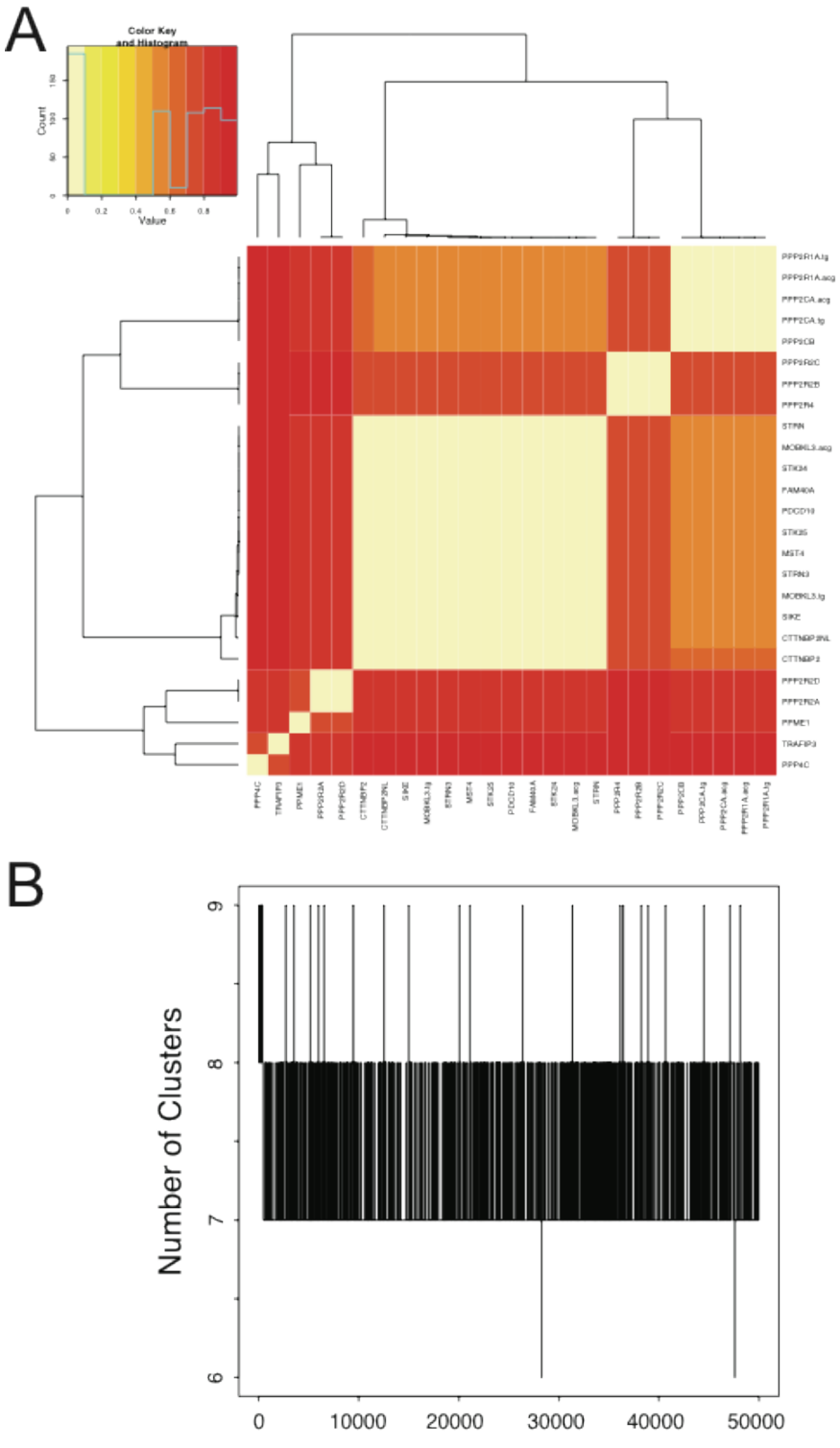
Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* **22:** 1122-1129.

Sardiu ME, Florens L, Washburn MP (2009) Evaluation of clustering algorithms for protein complex and protein interaction network assembly. *J Proteome Res* **8:** 2944-2952.

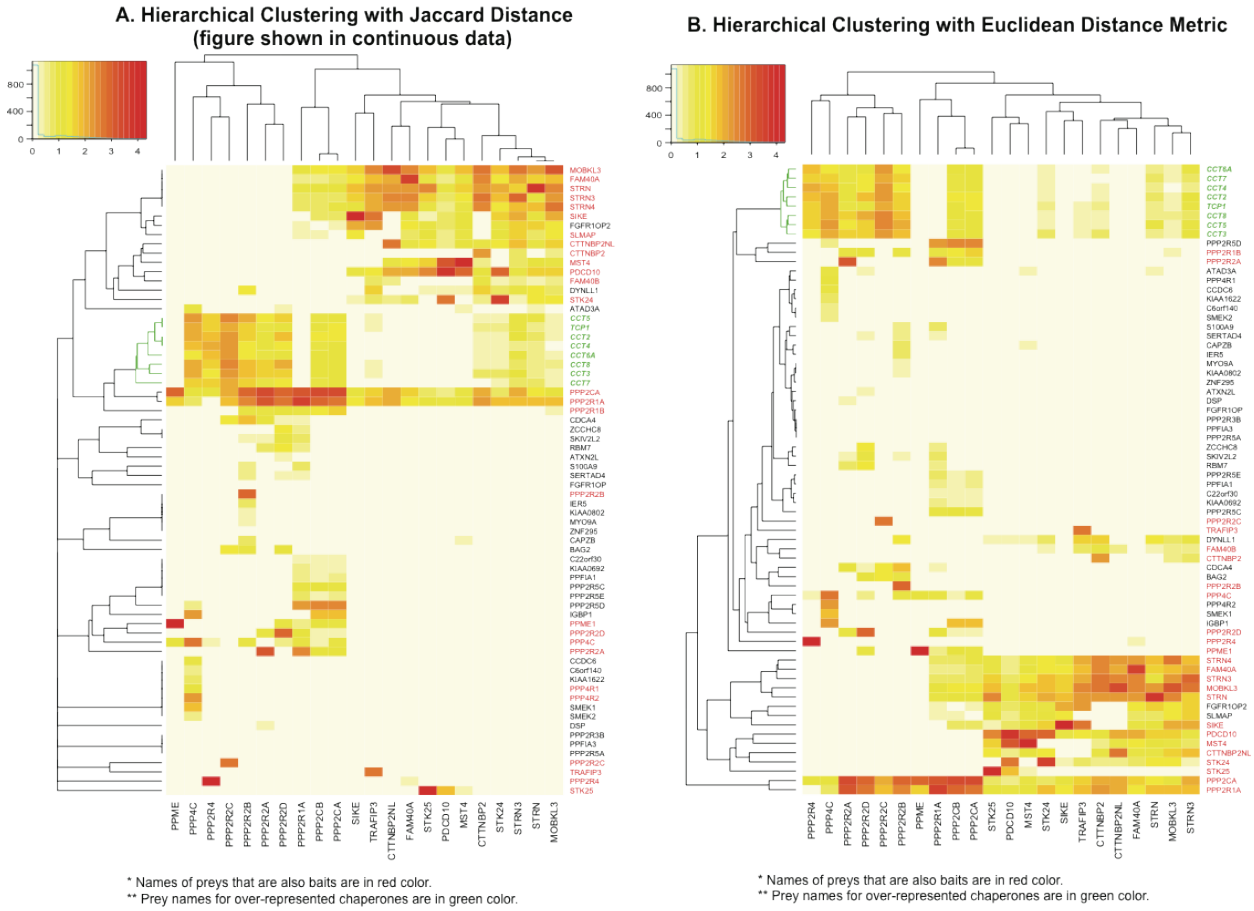
Sethuraman (1994) A constructive definition of Dirichlet priors. *Stat Sinica* **4:** 639-650.



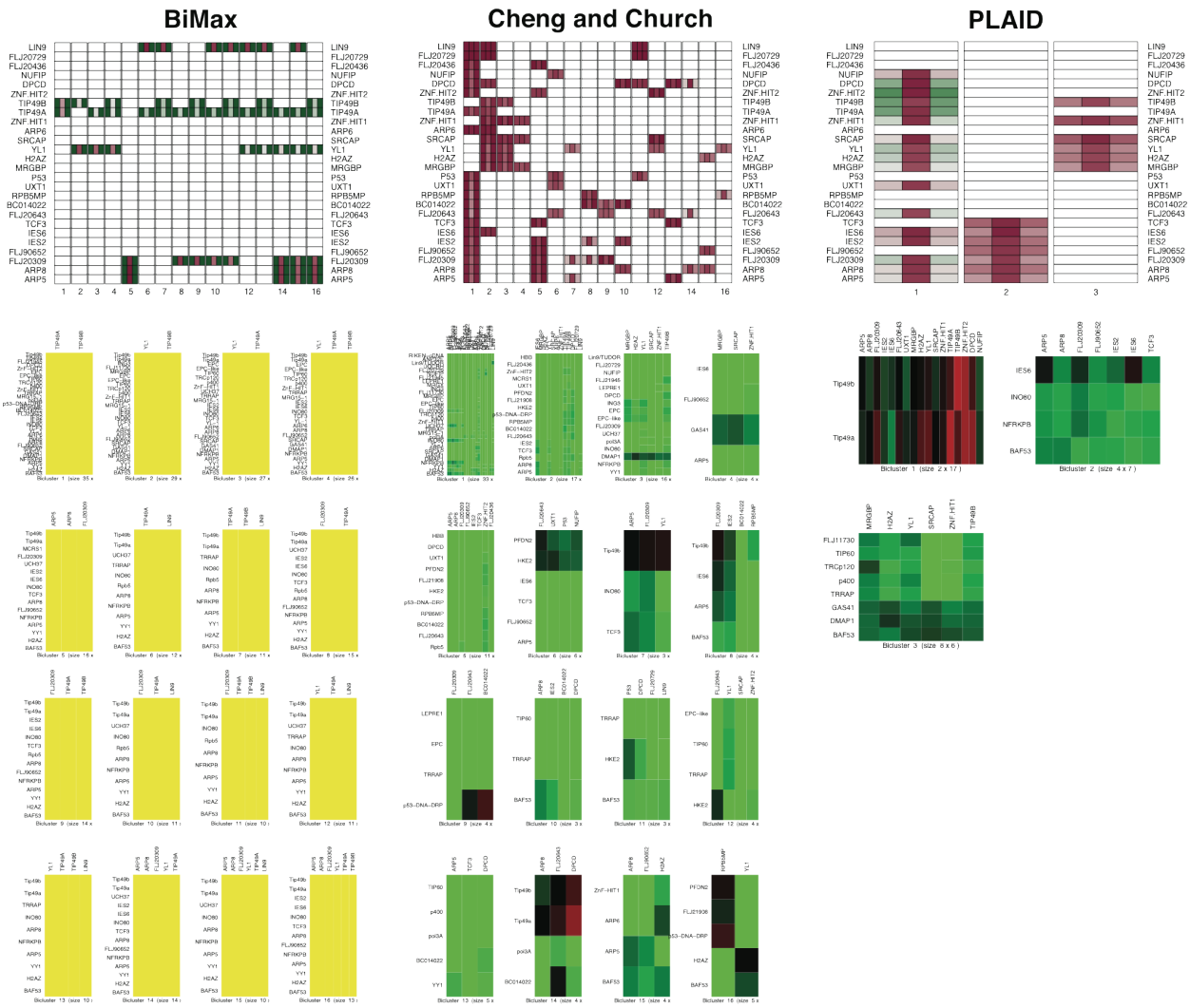
Supplementary Figure 1
A. Inter-bait probability distance matrix in the TIP49a/b dataset, organized by hierarchical clustering with Euclidean distance metric. B. Sampling trajectory of number of bait clusters in the algorithm.



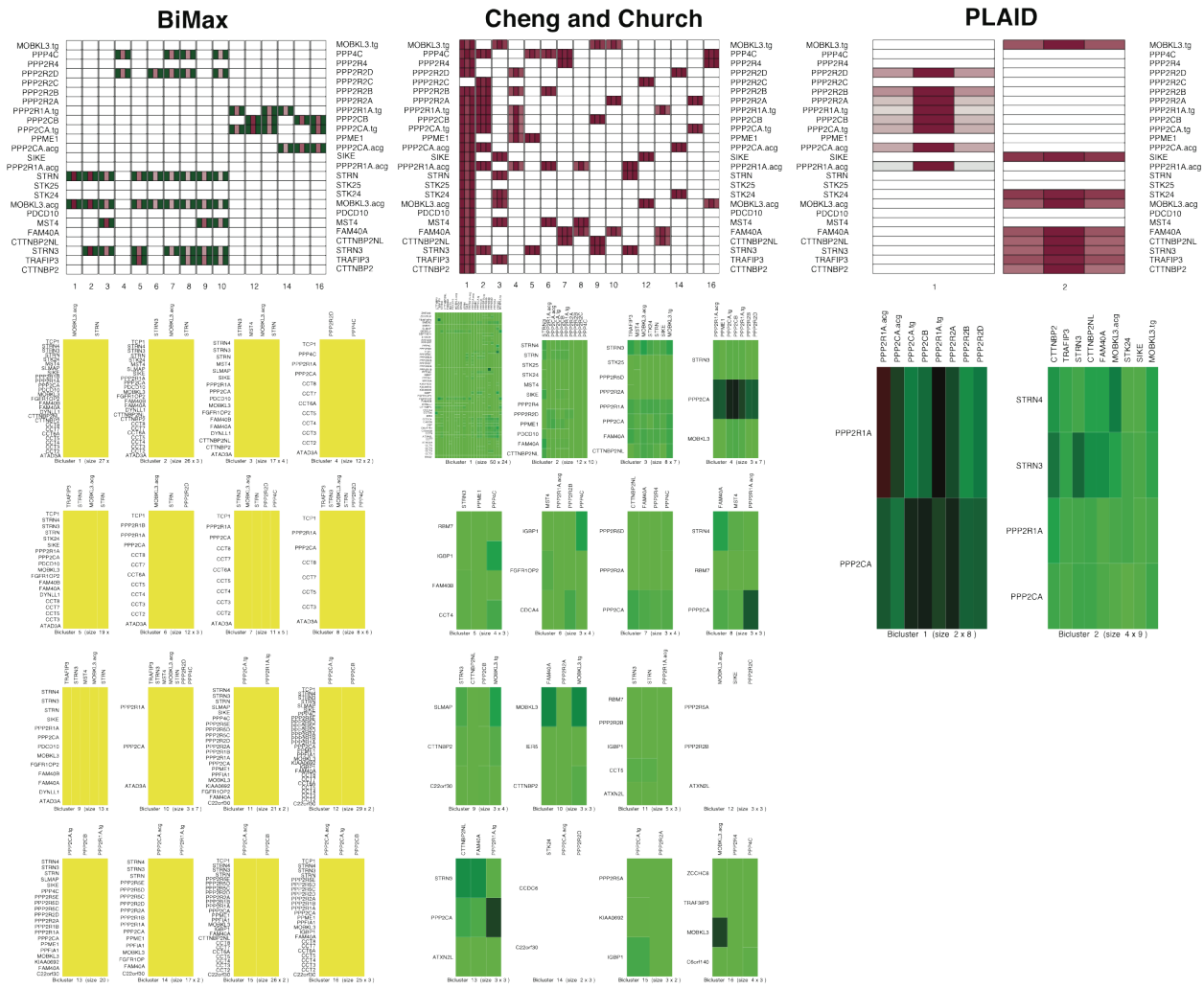
Supplementary Figure 2
Inter-bait probability distance matrix in the PP2A dataset, organized by hierarchical clustering with Euclidean distance metric. B. Sampling trajectory of number of bait clusters in the algorithm.



Supplementary Figure 3
Hierarchical clustering applied to binary and quantitative data of PP2A with the Jaccard distance and Euclidean distance metrics respectively.



Supplementary Figure 4
Application of BiMax, CC, and PLAID model to the TIP49a/b dataset. In each method, the top panel shows the bait clustering configurations for the top 16 submatrices. The bottom panel shows the corresponding submatrices.



Supplementary Figure 5

Application of BiMax, CC, and PLAID model to the PP2A dataset. In each method, the top panel shows the bait clustering configurations for the top 16 submatrices. The bottom panel shows the corresponding submatrices.