Scoring Large-Scale Affinity Purification Mass Spectrometry Datasets with MiST

Erik Verschueren,[1,6] John Von Dollen,[1,6] Peter Cimermancic,[1,3,6] Natali Gulbahce,[1] Andrej Sali,[4,5,6] and Nevan J. Krogan[1,2,6]

[1]Department of Cellular & Molecular Pharmacology, University of California, San Francisco, San Francisco, California
[2]Gladstone Institutes, University of California, San Francisco, San Francisco, California
[3]Graduate Group in Biological and Medical Informatics, University of California, San Francisco, San Francisco, California
[4]Department of Bioengineering and Therapeutic Science, University of California, San Francisco, San Francisco, California
[5]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, California
[6]California Institute for Quantitative Biomedical Sciences, San Francisco, California

High-throughput Affinity Purification Mass Spectrometry (AP-MS) experiments can identify a large number of protein interactions, but only a fraction of these interactions are biologically relevant. Here, we describe a comprehensive computational strategy to process raw AP-MS data, perform quality controls, and prioritize biologically relevant bait-prey pairs in a set of replicated AP-MS experiments with Mass spectrometry interaction STatistics (MiST). The MiST score is a linear combination of prey quantity (abundance), abundance invariability across repeated experiments (reproducibility), and prey uniqueness relative to other baits (specificity). We describe how to run the full MiST analysis pipeline in an R environment and discuss a number of configurable options that allow the lay user to convert any large-scale AP-MS data into an interpretable, biologically relevant protein-protein interaction network. © 2015 by John Wiley & Sons, Inc.

Keywords: affinity purification mass spectrometry • protein interactions • scoring algorithms • interaction networks • proteomics

How to cite this article:
Verschueren, E., Von Dollen, J., Cimermancic, P., Gulbahce, N., Sali, A., and Krogan, N. 2015. Scoring Large-Scale Affinity Purification Mass Spectrometry Datasets with MiST. *Curr. Protoc. Bioinform.* 49:8.19.1-8.19.16.
doi: 10.1002/0471250953.bi0819s49

INTRODUCTION

Affinity Purification Mass Spectrometry (AP-MS) is one of the primary methods to discover protein interactions in an unbiased manner. In recent years, due to advances in bottom-up mass spectrometry and affinity tagging methods, this method has been applied in high throughput to chart the protein-protein interaction networks or 'interactomes' of entire pathways from viral, bacterial, and eukaryotic organisms (Arifuzzaman et al., 2006; Sowa et al., 2009; Jäger et al., 2011). A high-throughput dataset containing hundreds of AP-MS samples, each with several replicates, poses a clear challenge for human processing, but also presents an opportunity to mine the data collectively with computational algorithms. To this end, a number of recent studies, which used AP-MS in a high-throughput fashion, developed computational algorithms that transform such a dataset into a list of bait-prey pairs ranked according their predicted biological

Current Protocols in Bioinformatics 8.19.1-8.19.16, March 2015
Published online March 2015 in Wiley Online Library (wileyonlinelibrary.com).
doi: 10.1002/0471250953.bi0819s49

significance (Sowa et al., 2009; Choi et al., 2011; Jäger et al., 2011). Knowing that a cellular protein is predicted to have on average five to eight biologically relevant interactions (Grigoriev, 2003), prioritizing these from over hundreds of proteins and thousands of spectra identified by MS is far from trivial.

To understand the high number of ‘false positives’ identified by AP-MS, it helps to categorize interactions into four broad classes: (1) biologically relevant interactions; (2) specific, non-biologically-relevant interactions between proteins from different cellular compartments in lysed cells; (3) unspecific interactions with contaminants or highly abundant proteins; and (4) non-existing interactions caused by residual peptides from previous runs or MS identification errors. Conversely, ‘false negatives’ occur because not every biologically relevant interaction is reproducibly detectable, especially if the protein is not very abundant, has peptides difficult to detect by MS, or interacts only transiently (Yu et al., 2009; MacLean et al., 2010). To account for both false-positive and false-negative errors, high-throughput experimental setups need to be designed with proper controls and a sufficient number of biological replicates (at least triplicates), and processed with a computation AP-MS scoring algorithm to separate signal from noise (Jäger et al., 2011).

The protocols we present here outline our ‘best-practices’ approach to convert a large data set of replicated AP-MS experiments into a list of bait-prey pairs ranked according to their predicted biological relevance. After installing MiST as described in the Support Protocol, the data is pre-processed in Basic Protocol 1 to generate a bait-prey matrix that can be subjected to the quality-control protocol in Basic Protocol 2. Basic Protocol 3 is then used to calculate a MiST score.

BASIC PROTOCOL 1

DATA PRE-PROCESSING

Prior to computing MiST scores, it is required that the search results be converted into a format compatible with the MiST algorithm. The MiST pre-processing pipeline was initially designed to work with a Prospector (Clauser et al., 1999; *http://prospector.ucsf.edu*) protein report file but virtually any report file that lists uniquely identified proteins and their observed peptide frequencies in tabular format is currently supported. Additionally, we built a number of filtering steps, such as contaminant removal and carryover removal, into the pre-processing and formatting steps. The result of this protocol is a bait-prey matrix that can be subjected to the quality-control protocol (Basic Protocol 2) and scored by the MiST protocol (Basic Protocol 3) (see Fig. 8.19.1).

Necessary Resources

Hardware

Workstation running any current OS; Unix environment recommended

Software

MiST pipeline (installed as described in the Support Protocol)

Files

`data` file and `keys` file (see below)
`remove` file and `collapse` file (optional; see below)

Data preparation

Prior to running the MiST pre-processing protocol, the user needs to have at least the `data` file and the `keys` files available (See Fig. 8.19.2A and Table 8.19.1).

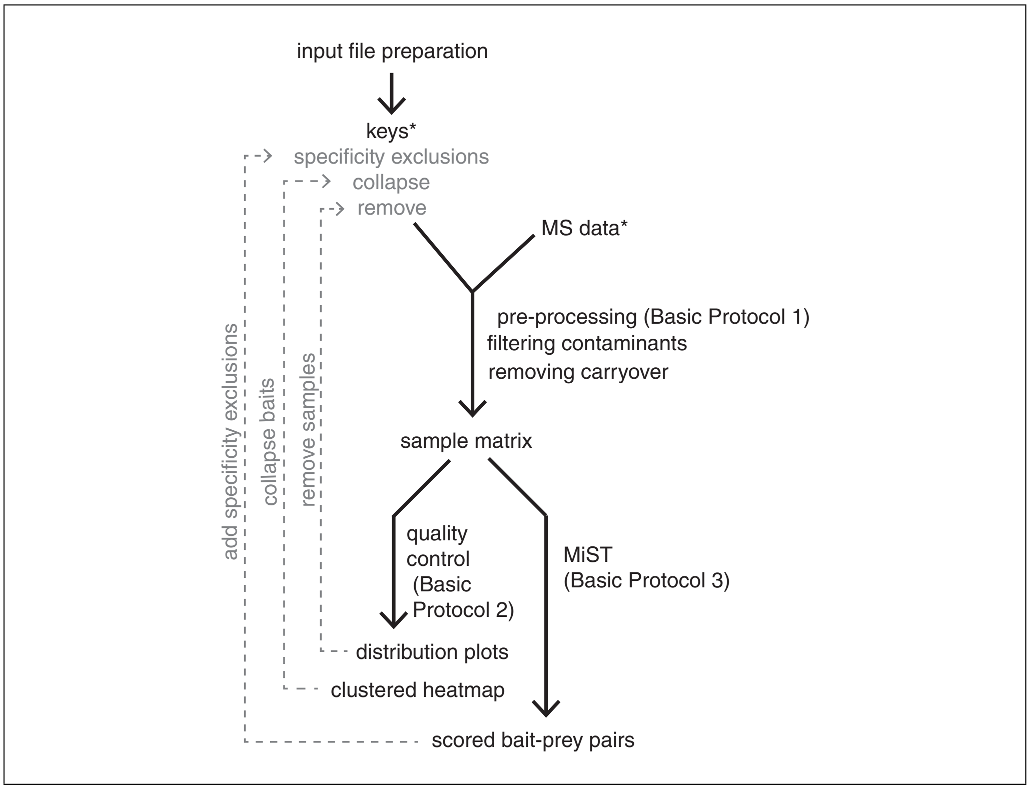


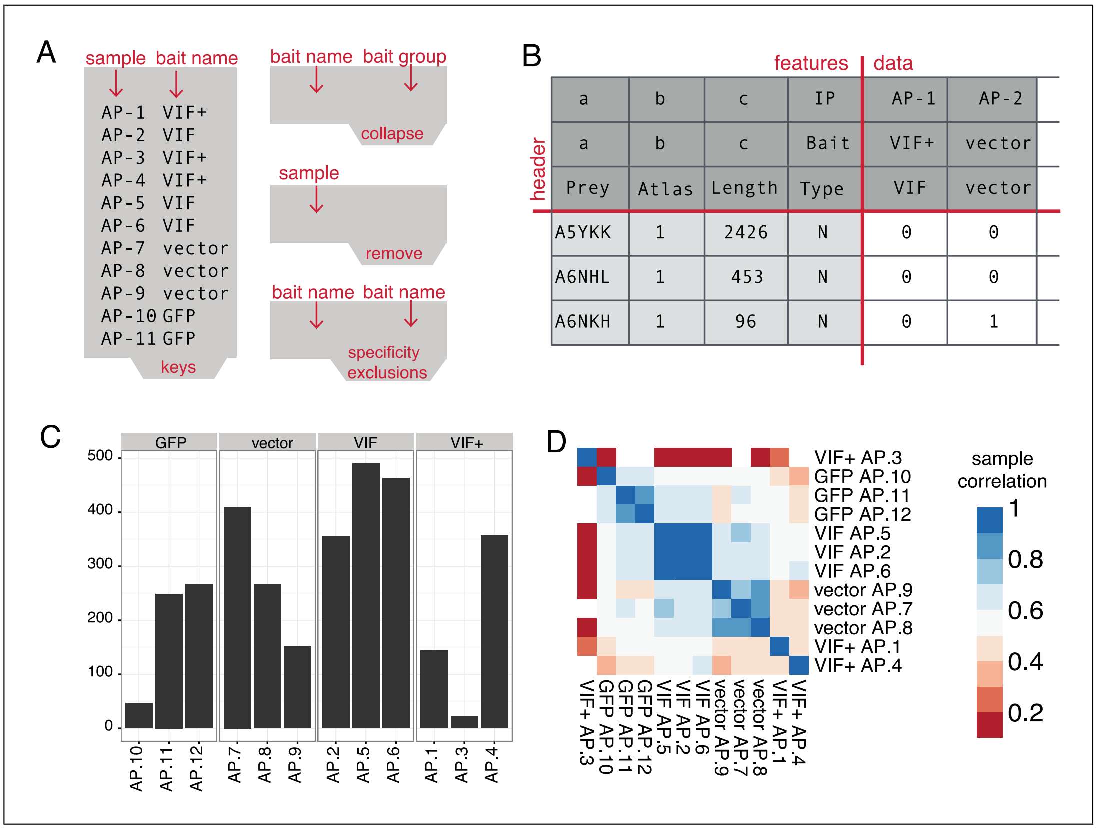
Figure 8.19.1 Flow chart of the Mass Spectrometry Interaction Statistics (MiST) analysis pipeline. Required input files are marked with (*). Optional steps in the protocol and their corresponding input files are indicated in light gray.

Table 8.19.1 Configurable Parameters for file Input/Output

Name	Type	Description
`data`	string	Path to the data file with the identified and quantified proteins (see data preparation)
`keys`	string	Path to the keys file matching samples to bait names (see data preparation)
`remove`	string	Path to the file listing samples that should be excluded from analysis
`collapse`	string	Path to the file listing groups of baits that can be treated as biological replicates
`specificity _exclusions`	string	Path to the file listing baits to mutually exclude when computing specificity (see Basic Protocol 2)
`output_dir`	string	Directory to write output files

1. `data` *file (required):* The `data` file should be in tab-delimited format with a descriptive header for each column. The number of features in this file is in principal unconstrained but the following are required:

   a. *Sample identifier*: A unique identifier for each AP-MS run.
   b. *Protein identifier*: A unique identifier for each protein (i.e., Uniprot accession code).
   c. *Observed peptide frequency*: A quantitative value for each protein, for example (from less to most quantitative):
      i. Number of unique peptides per protein or;
      ii. Summed spectral counts per protein or;



```
AP-1  VIF+
AP-2  VIF
AP-3  VIF+
AP-4  VIF+
AP-5  VIF
AP-6  VIF
AP-7  vector
AP-8  vector
AP-9  vector
AP-10 GFP
AP-11 GFP
```

a	b	c	IP	AP-1	AP-2
a	b	c	Bait	VIF+	vector
Prey	Atlas	Length	Type	VIF	vector
A5YKK	1	2426	N	0	0
A6NHL	1	453	N	0	0
A6NKH	1	96	N	0	1

Figure 8.19.2 (**A**) Tab-separated input files required to run a first pass of MiST. (**B**) Output bait-prey matrix after the pre-processing protocol (1) that serves as input for the MiST protocol (3). (**C**) Output of the quality control protocol depicting the number of detected proteins across replicates for a single bait. Samples AP-10 and AP-3 are candidates for removal from the data set. (**D**) Output of the quality control protocol depicting a hierarchically clustered heatmap of the pairwise sample correlation matrix. The signal from sample AP-3 is clearly uncorrelated with the VIF+ replicates while AP.10 correlation with the GFP group is acceptable.

iii. Summed MS1-intensities per protein, preferably log2-transformed.

d. *Protein molecular weight*: The molecular weight will be used as a scaling factor to normalize the quantitative value for each protein to its size.

2. `keys` *file (required):* The goal of the `keys` file is two-fold: (1) describing which bait and experimental conditions are in the sample with a human-readable description; and (2) providing a unique name to group biological replicates. This file also needs to be tab-separated with two columns created by the user:

    a. *Sample identifier*: a unique identifier for each AP-MS experiment that matches a sample identifier in the `data` file

    b. *Bait name*: A unique identifier for each bait, grouping all replicates. This can be the bait's Uniprot accession number or a more easily readable name (i.e., gene name).

3. `remove` *file (optional):* The `remove` option is a feature we put in place to dynamically exclude entire samples from the scoring process while keeping the original `data` files intact. Single samples might need to be removed for quality reasons, which we will address later, while an entire range of samples might be excluded to score subsets of the complete data set. This file is formatted as a single column, consisting of all sample identifiers that need to be excluded (one sample per line).

4. `collapse` *file (optional):* The `collapse` option serves to merge samples belonging to different baits into a single group while keeping the original `data` files intact.

Table 8.19.2 Configurable Parameters for Pre-Processing

Name	Type	Description
`remove _carryover`	boolean	Enables the attempted removal of carryover proteins from a previous run
`flter _contaminants`	boolean	Enables the removal of known contaminants from the prey list
`contaminants _file`	string	Path to file listing all known contaminants in FASTA format
`Id_colname`	string	Column identifier for sample identifier
`prey_colname`	string	Column identifier for identified proteins in `data` file
`pepcount _colname`	string	Column identifier for observed peptides per protein in `data` file
`mw_colname`	string	Column identifier for protein molecular weight in `data` file

This is useful when a particular experimental condition (i.e., compound addition, bait mutation, different affinity tags, differently tagged termini, organelle extraction) shows no perceivable difference from the wild-type experiment. In this case, it might be desirable to treat these samples as additional replicates of the wild-type bait to improve reproducibility estimates. In the following section we will discuss how you can use a clustered heat map to reveal such patterns. The `collapse` file is formatted as tab-separated entries: one for the original bait name, corresponding to an entry in the `keys` file, followed by the new name or composite group name.

To guide the reader through the various protocols, we prepared an example project that can be downloaded together with the MiST source code from the github repository at https://github.com/everschueren/mist. For instructions on downloading and installing see Support Protocol After installation, the example configuration file `mist_small_test.yml` *and the corresponding input files can be found in* `$INSTALL_DIR/tests/small/`*. The configuration file follows the Yaml Ain't Markup Language (YAML) (http://www.yaml.org) format that allows defining conceptual blocks of parameters. The first parameter block deals with file input/output (See Table 8.19.1) and the consecutive blocks correspond to configuration options for the three basic protocols outlined in this unit. The* `enabled [0/1]` *parameter in each parameter block turns each basic protocol off or on respectively.*

Running the pre-processing protocol (see Table 8.19.2)

5. The first pre-processing step is to remove common contaminants such as all the keratins and known cross-reacting proteins with beads or affinity tags during purification. To enable this option, turn on `filter_contaminants` in the configuration file and define the path to a text file listing all contaminants using the same identifiers as in the `data` file.

   Even though it is up to the user to define a custom set of contaminants, we provided a minimal list, based on the Maxquant (Cox and Mann, 2008) contaminant file augmented with most keratin proteins we regularly come across and a 'decoy' entry for Prospector false hits. Since contaminants might be condition- and even machine-specific, we encourage users to search the recently published Crapome (Mellacheruvu et al., 2013) database for cell-, tag,- or bead- specific contaminants matching their experimental setup.

6. The second pre-processing step is to computationally remove peptide counts of proteins that are due to sample 'carryover' from a previous run on the MS. To enable this feature turn `remove_carryover` on.

Accurately preventing and detecting sample carryover between consecutive MS runs is not a trivial problem, and is an active topic of discussion in the MS community. Carryover is caused by residual peptides from a previous sample and is generally serial in nature, often affecting several samples in a sequence (Hughes et al., 2007). We have empirically observed that hydrophobic proteins in combination with over-expression of the bait protein can lead to a higher number of carryover peptides. We recommend addressing this issue by (1) testing various experimental conditions that minimize sample carryover, and (2) shuffling the order in which biological replicates of samples are run. A third option is to remove any false hits that are detectable after the fact by a computational procedure similar to the one we implemented. The current procedure checks for carryover in up to four samples following the current sample. Carryover is determined if the following conditions are met in the samples tailing a sample: (i) there is a positive number of unique peptides for a specific protein, (ii) the number of unique peptides is less than half of the current sample, (iii) the number of occurrences of this prey in the data set is less than one third of the number of total experiments.

7. Map the values of the column names in the tab-delimited `data` file to the column description parameters (See Data preparation section for details):

   a. `id_colname` : sample identifier
   b. `prey_colname` : protein identifier
   c. `pepcount_colname` : observed peptide frequency
   d. `mw_colname` : protein molecular weight

8. run the MiST pipeline as follows: `$INSTALL DIR/main.R --config [path to YML file]` (see Support Protocol for installation notes).

9. Inspect the bait-prey matrix called `preprocessed_MAT.txt` in the output folder defined by `output_dir`, or the `processed` directory in the executable path if no output directory was defined. The bait-prey matrix is formatted according the data preparation guidelines in the next section (see Fig. 8.19.2B).

BASIC PROTOCOL 2

QUALITY CONTROL

We built a number of quality control and data summary plots such as MS yield statistics and a hierarchically clustered heat map of the experiment matrix into the MiST package. The output of these quality control scripts can help to decide critical parameters that influence the MiST scores, such as samples to remove or conditions to group together as replicates.

Necessary Resources

Hardware

Workstation running any current OS; Unix environment recommended

Software

MiST pipeline (installed as described in Support Protocol)

Files

Data pre-processed as described in Basic Protocol 1

Data preparation

1. After the preprocessing protocol of the MiST pipeline has been run, the input data should be properly reformatted for the quality control algorithm. For convenience, the quality control and scoring input follows the same formatting guidelines as the SAINT algorithm bait-prey input matrix. If the pre-processing step was skipped, the user should make sure to format the input for all subsequent steps as follows (See Fig. 8.19.2B):

Table 8.19.3 Configurable Parameters for Quality Control

Name	Type	Description
`matrix_file`	string	Path to preprocessed matrix file. The matrix created in the preprocess step will be used if this is left blank.
`cluster`	boolean	Whether to perform a hierarchical cluster analysis on the data
`cluster_font _scale`	integer	Row/column font adjustment factor for the clustered heat map
`ip_distributions`	boolean	Whether to perform a protein count and peptide distribution analysis per group of biological replicates

a. Columns:

i. *Preys:* All distinct protein identifiers found as preys in the complete set of samples. These correspond to the unique identifiers in the `prey_colname` of the `data` file

ii. *PepAtlas:* If Peptide Atlas counts are not known, set this to 1. PepAtlas counts are for SAINT compatibility. (Choi et al., 2012).

iii. *Length:* Scaling factor derived from the indicated molecular weight per protein in the `mw_colname` column.

iv. *PreyType:* Prey type is set to `N` to maintain matrix structure compatible with SAINT.

b. Header rows:

i. *row 1:* All unique sample identifiers. Sample identifiers correspond to the unique identifiers in the `id_colname` of the `data` file and the first column in the `keys` file.

ii. *row 2:* Bait names for samples on row 1. Bait names correspond to the mapping of baits to samples described in the `keys` file.

iii. (row 3) specificity exclusions for baits on row 2. Specificity exclusions are described in the `specificity_exclusions` file and will be discussed in the section on MiST scoring (Basic Protocol 3).

c. *Data in Row 4-[unique preys] × Column 5-[samples]:* Peptide counts per prey/sample from the `pepcount_colname` column in the `data` file.

Running the quality control protocol (see Table 8.19.3)

2. If `matrix_file` is left blank, then the output matrix produced by the pre-processing step (Basic Protocol 1) is used; otherwise, this parameter should point to the path of a correctly formatted bait-prey matrix as described above.

3. If `ip_distributions` is enabled, then a number of plots summarizing sample features per group of biological replicates (see Fig. 8.19.2C) are saved into the `output_dir`:

   a. `_proteincounts.pdf` shows the total number of identified proteins with unique peptides. (see Fig. 8.19.2C)

   b. `_NumUniqPep.pdf` shows the distribution via boxplot of the peptide counts by the replicates grouped by bait. Replicate distributions that are very different may imply something went wrong with that sample, leading to it being removed in future scoring.

4. If `cluster` is enabled, then a hierarchically clustered heatmap showing the pair-wise signal correlation between all samples is saved into the `output_dir` (See Fig. 8.19.2D).

The correlation between the observed peptide counts for each pair of samples is measured by the Pearson correlation coefficient. Proteins that were not identified in a sample are given a zero peptide count. The resulting symmetric correlation matrix is then clustered with R's `hclust` *algorithm using the default Euclidian distance metric and visualized with R's* `pheatmap` *library. Except the* `cluster_font_scale` *parameter, which can be decreased or increased for larger or smaller datasets, respectively; the remaining described parameters are currently not configurable.*

BASIC PROTOCOL 3

CALCULATING THE MiST SCORE

The MiST score is a weighted sum of three features: (1) normalized protein abundance measured by peak intensities, spectral counts, or unique number of peptide per protein (abundance); (2) invariability of abundance over replicated experiments (reproducibility); and (3) a measure of how unique a bait-prey pair is compared to all other baits (specificity). The weights of the three features are configurable in three different ways: first, pre-configured fixed weights can be used; second, they can be trained de novo on a custom list of trusted bait-prey pairs identified in the data set; lastly, a principal component analysis (PCA) can be run to assign the feature weights according their contribution to the variance in the data set.

Necessary Resources

Hardware

Workstation running any current OS, Unix environment recommended

Software

MiST pipeline (installed as described in Support Protocol)

Data preparation

1. After the preprocessing protocol of the MiST pipeline has been run, the input data should be properly reformatted for the main scoring algorithm. For convenience, the scoring input follows the same formatting guidelines as the SAINT algorithm bait-prey input matrix. If the pre-processing step was skipped, the user should make sure to format the input for all subsequent steps as described in the data preparation section of the Quality Control step (See also Fig. 8.19.2B).

   In addition to the bait-prey matrix input file, the user also has the option of setting bait exclusion rules when calculating the MiST scores. These rules only apply when computing MiST's specificity feature value. In brief, every exclusion rule defines which baits should be excluded from the specificity denominator. Even though the definition of bait exclusion rules is an optional parameter, these rules can highly influence the results. Therefore, we further discuss their proper use in the Critical Parameters section of this unit.

   To apply specificity exclusion rules, create a tab-delimited file (See Fig. 8.19.2A) where every row consists of:

   a. *column 1:* The name of the bait whose specificity exclusion rules you would like to define. Make sure that the bait name corresponds to the name that was used in the `keys` or `collapse` file.
   b. *column 2:* The names of the baits that you would like to exclude when specificity is being computed for the bait listed in column 1. Multiple baits can be excluded by separating them with a pipe (|) symbol. Again, make sure that all bait names correspond to names that were used in the `keys` or `collapse` file.

Table 8.19.4 Configurable Parameters for the MiST Scoring Algorithm

Name	Type	Description
`matrix_file`	string	Path to preprocessed matrix file. The matrix created in the preprocess step will be used if this is left blank.
`weights`	string	One of three possible values: fixed/training/PCA (See MiST section)
`training_file`	string	Path to the file containing bait-prey pairs for training (only if weights : training)
`reproducibility`	double	MiST weight [0-1] for the reproducibility feature
`abundance`	double	MiST weight [0-1] for the abundance feature
`specificity`	double	MiST weight [0-1] for the specificity feature

Running the MiST scoring protocol (see Table 8.19.4)

2. *Optional:* Create a specificity exclusion file (see step 1) and define the path to this file through the `specificity_exclusions` entry in the `files` block of the configuration file.

3. If `matrix_file` is left blank, then the output matrix produced by the pre-processing step (Basic Protocol 1) is used; otherwise, this parameter should point to the path of a correctly formatted bait-prey matrix.

4. Decide on a strategy to combine the abundance, reproducibility, and specificity features into a single MiST score. To do so, set the weights parameter to either:

   a. *fixed*: Choose a decimal value between 0 and 1 for reproducibility, abundance and specificity.

   When choosing fixed values, make sure that the sum of these values is 1. If no weight values are chosen, the weights default to 0.309 for reproducibility, 0.685 for specificity, and 0.006 for abundance. These weights were established in the first MiST publication (Jäger et al., 2011) and are a good choice to select reproducible, specific bait-prey pairs.

   b. *training*: MiST will use a `training_file` to exhaustively test the performance of different parameter combinations and select the optimal configuration.

   We recommend using this option only when a sufficiently large benchmark set is available. See the Advanced Parameters section, below, for details.

   c. *PCA*: MiST will perform a Principal Component Analysis (PCA) on the three-dimensional feature matrix and select weights that project the feature values on the first principal component.

   We recommend using this option only when the suggested fixed weights do not perform well and insufficient training data is available.

Result file

5. The result file is a tab-delimited file with a unique entry for each observed bait-prey pair organized in the following columns:

   Bait: Bait name as described in the `keys` or `collapse` file
   Prey: Protein identifier as listed in the `prey_colname` column of the `data` file
   Abundance: MiST abundance feature value of the bait-prey pair
   Reproducibility: MiST reproducibility feature value of the bait-prey pair
   Specificity: MiST specificity feature value of the bait-prey pair
   MiST: The total MiST score value of the bait-prey pair
   Ip: All samples in which the bait-prey was observed.

SUPPORT PROTOCOL 1

INSTALLATION OF MiST

This MiST pipeline is implemented in R, an open-source programming language for statistical computing and graphics. Here, we describe how the current version (1.0) of the MiST pipeline can be downloaded from GitHub and installed by any user with access to an online computer.

Necessary Resources

Hardware

Workstation running any current OS; Unix environment recommended

Software

R package (*http://www.r-project.org*)
R packages: `getopt`, `optparse`, `reshape2`, `pheatmap`, `RcolorBrewer`, `ggplot2`, `MESS`, `yaml`
MiST source code (*https://github.com/everschueren/MiST*)
Git (optional) (*http://git-scm.com*)

Setting up MiST

1. Download the MiST source code to your workstation.

   a. Download the MiST package as a `.zip` archive from the public GitHub repository by clicking on the "Download ZIP" button on the bottom right, unzip the files, and move the directory to a permanent location.
   b. Alternatively, you may check out the MiST package through Git as follows:

   ```
   git clone https://github.com/everschueren/MiST.git
   MiST
   ```

2. The MiST pipeline is designed to run from a terminal using R. This requires the user to have executable permissions. To set these permissions in a Unix environment, navigate in the terminal to the MiST directory, hereafter referred to as the `$INSTALL_DIR`, then type: `sudo chmod -R 775 *`

GUIDELINES FOR UNDERSTANDING RESULTS

Quality control plots

Different protein baits can have drastically different amounts of interacting proteins. Nevertheless, these levels should be consistent across biological replicates. A simple bar plot grouping all replicates for a single bait can help in spotting samples of lower quality, which can consequently be removed from the dataset (See Fig. 8.19.2C and 8.19.3A).

The primary use of the hierarchically clustered heatmap is to validate that samples are indeed more correlated within their group of replicates compared to negative controls and different baits. If this is not the case and a sufficient number of replicates are available, the sample can be removed from the data by adding its identifier to the `remove` file. In addition, carefully inspecting clusters can reveal accidentally mislabeled samples. For example, if one sample clusters more tightly with an unrelated group of baits compared to the other samples in its group, this could be an indication that the sample was accidentally mislabeled. If this is the case, it is helpful to inspect the number of bait peptides that should be detected at high levels in the sample. Lastly, if a bait is purified under multiple experimental conditions (mutations, beads, tags, drugs, etc.), the correlation between these conditions can be an indication of the effect of the condition. Instead of discarding these samples, the user can make a conscious choice to treat the conditional purifications as replicates of the wild type (See Figs. 8.19.2D and 8.19.3B).

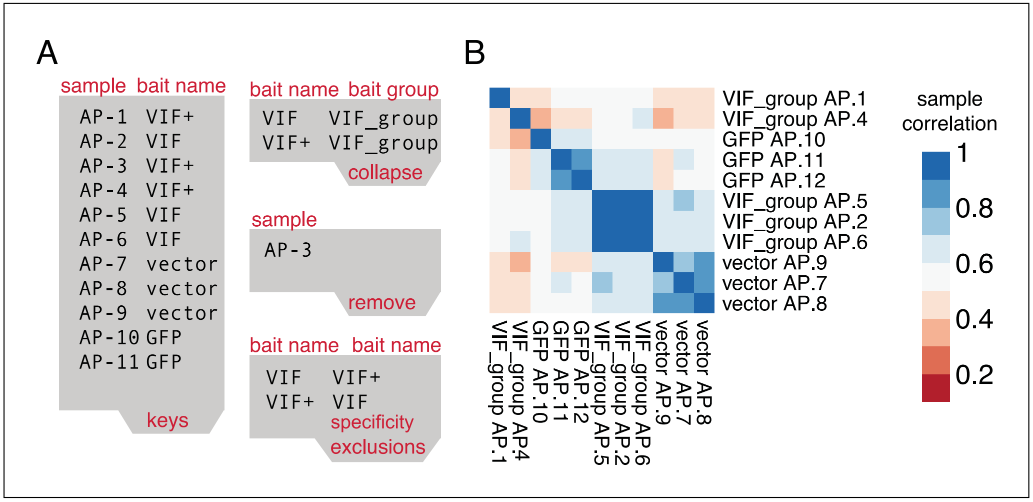


Figure 8.19.3 (**A**) Tab-separated input files adjusted based on the Quality Control observations: (1) sample AP-3 is 'removed' from the data. (2) VIF and VIF+ are 'collapsed' into a VIF_group and optionally VIF and VIF+ could have been mutually excluded for specificity calculations. (**B**) The clustered heatmap after adjustments shows no low-quality data and clearly distinct replicate clusters aligned with the three different bait groups.

MiST score

MiST aims to predict whether a protein-protein interaction detected in a set of repeated AP-MS experiments is biologically relevant using three features: reproducibility, abundance, and specificity. To get as close as possible to the optimal MiST score of 1 for an interaction, it is important that this interaction scores well across all features (See Fig. 8.19.4A). Since values of 1 or close to 1 are rare, a minimum threshold can be applied to separate the predicted true interactions from the rest. However, choosing the appropriate threshold that makes a good trade-off between prediction sensitivity (detecting all true interactions) and specificity (minimizing the number of false positives) can be challenging and depends on the choice of feature weights.

The easiest way to pick a threshold is therefore to stick to the recommended value for specific weights. For example, for the previously published HIV–host interaction network, a MiST lower-bound threshold of 0.75 is recommended. When your dataset is comparable to a reference set this is an easy and sound solution.

A slightly harder but more preferred way of picking a threshold is by making prediction plots such as a Receiver Operating Curve (ROC), or a precision-recall curve, and computing prediction accuracy statistics like the f1 score. The main drawback of this approach is that these metrics depend on the presence of 'true' positive and negative bait-prey interactions in the MS data set. While the so-called 'negative' interactions are often picked randomly from the data, the absence of known 'positive' interactions for a protein is often the very reason to perform an AP-MS experiment. When compiling a benchmark set, we recommend using a positive set of at least 20 interactions and picking a negative set roughly 100 times larger than the positive set.

Finally, if neither of the aforementioned strategies for choosing a threshold is feasible, we advise users to respect two rules of thumb. First, never pick a threshold lower than the highest feature weight value. For example, if specificity has the highest weight of 0.68, the MiST threshold should be strictly greater than 0.68. As explained before, biologically relevant interactions are expected to be reproducible and specific (See Fig. 8.19.4A). Interactions with a perfect reproducibility score but zero specificity score are most likely 'background' interactions with highly abundant proteins. Conversely, interactions with a perfect specificity score and zero reproducibility score are likely to be 'one-hit-wonders'.

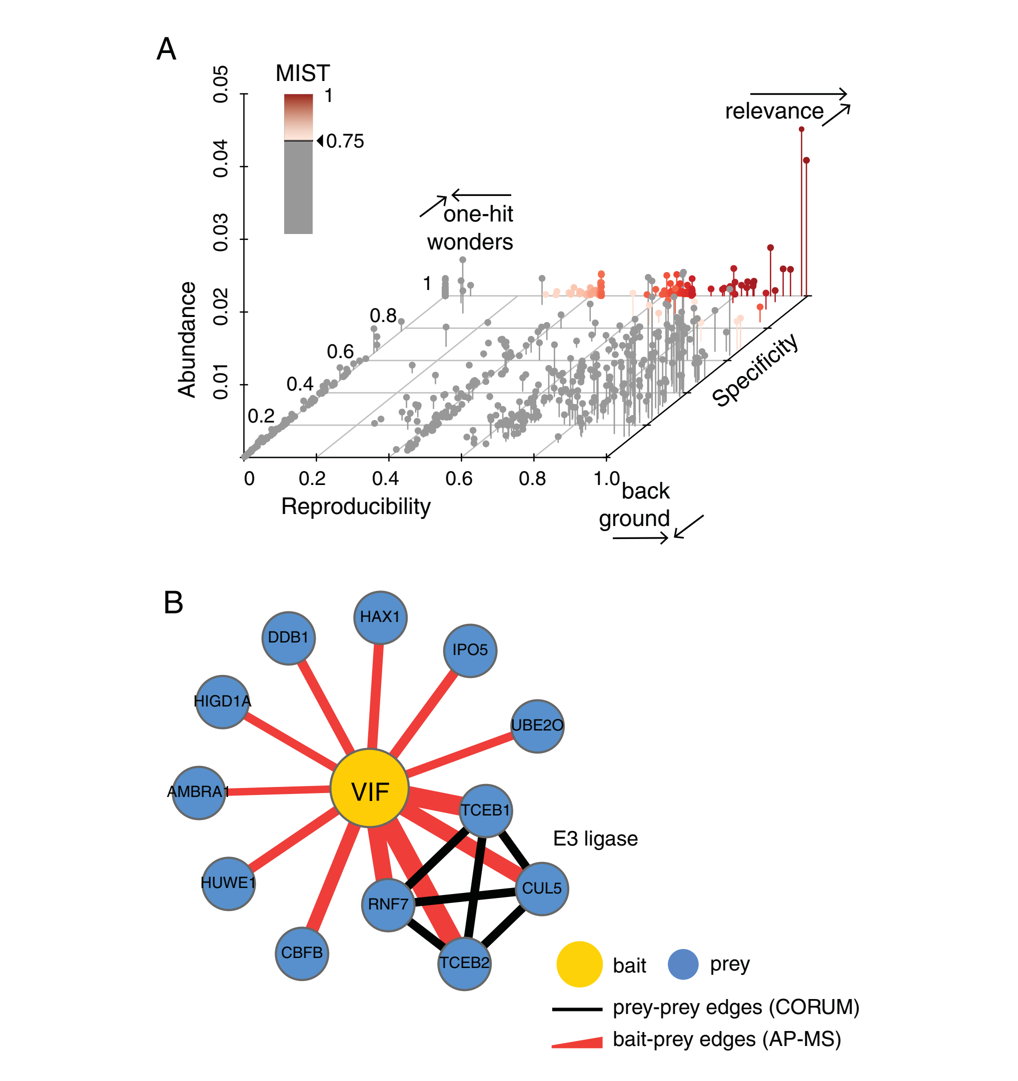


Figure 8.19.4 (**A**) A 3-D scatterplot illustrating three different areas in the three-dimensional MiST feature space: (1) biologically relevant interactions (red gradient, MiST scores $>$ 0.75) are specific and reproducible with variable abundance. (2) Nonspecific, reproducible interactions are often background proteins, and (3) specific, irreproducible interactions are often one-hit-wonders including contaminants and MS artifacts. (**B**) The scored bait-prey list at a high MiST score threshold visualized as a protein interaction network with Cytoscape. Red edges depict interactions identified by AP-MS while black edges are mined from the CORUM database. Edge width corresponds to the MiST score. High MiST scores for multiple subunits of a described complex add confidence to observed AP-MS interactions.

Second, keep in mind that the goal of scoring an AP-MS data set should be to approach the true biologically relevant interaction network as closely as possible. Even though it is known that some proteins act as hubs and others have just one single interaction partner, current studies estimate the average number of interactions to be around five to eight per protein. This expected bait-prey ratio could therefore be used to determine a reasonable cutoff for the MiST scores. Since the primary purpose of an AP-MS experiment is to discover new interactions, and these studies are often followed up with a more targeted experiment, it is still acceptable to consciously allow a higher number of potential false positives.

COMMENTARY

Background Information

The MiST score was originally developed to rank bait-prey pairs in an ex vivo HIV-human data set (Jäger et al., 2011). When this data set was being produced, SAINT [*UNIT 8.15* (Choi et al., 2011), Choi et al., 2012 and the CompPASS-D (Sowa et al., 2009) score were the two main computational algorithms that were suited to analyze large-scale AP-MS data sets. However, their prediction performance was only reported on a data set of human-human protein interactions. The uniqueness of the HIV-human data sets became the primary reason to develop MiST, a custom AP-MS scoring algorithm. Although MiST and CompPASS both use exclusively abundance, reproducibility, and specificity as predictive features, making them therefore somewhat comparable, MiST scores are easier to interpret because their feature value and total score varies between 0 and 1. Even though a ranked bait-prey list based on CompPASS scores is quite accurate, the actual scores are by definition harder to interpret because they vary between 0 and extremely high numbers.

SAINT scores, on the other hand, vary between 0 and 1, but are conceptually very different because they describe the probability that a bait-prey pair is true based on a distribution model of its abundance values. For optimal SAINT performance, it is therefore recommended to have a well-defined set of negative-control affinity purifications to compare against. Neither MIST nor CompPASS require explicit definition of negative controls; in fact, negative controls are treated just the same as any another bait purification in the dataset.

To assess the accuracy of MiST, we compared it to the SAINT and CompPASS scores. The accuracy of each score was evaluated by its recall rate for the set of 39 well-characterized biologically relevant HIV-human bait-prey pairs. The MiST score was the most accurate among all the tested scores (Jäger et al., 2011). For example, at the threshold of 0.75, the recall number of known bait-prey pairs for the SAINT, CompPASS and MiST scores was 19, 29, and 32, respectively. Furthermore, 97 out of 127 (76% recall) top-ranked interactions predicted by MiST were validated using co-immunoprecipitation followed by Western blotting as an orthogonal assay.

For an additional test, we counted bait-prey pairs involving ribosomal proteins, which are a good indicator of biologically irrelevant bait-prey pair predictions (Ewing et al., 2007). Again, MiST was the most accurate score, resulting in only three HIV-ribosomal protein bait-prey pairs, compared to 32 and 75 for SAINT and CompPASS, respectively.

Critical Parameters

Specificity exclusions

We repeatedly observed in gold-standard data sets that good prediction performance goes hand in hand with a high weight for specificity. Preys exclusively identified with a single bait protein will receive a high specificity value; therefore, these preys will get high MiST scores overall. Conversely, prey proteins identified with several baits are not bait specific and will receive overall lower scores. However, often large-scale interrogations of biological systems are designed in a way that inherently introduces a bias into the number of times preys are identified across multiple baits. For 'specificity' to optimally work the way it is intended to, it is therefore important to remove design bias from the calculations. Here, we illustrate how you can achieve this by describing the three most common examples:

Conditional interactions: When an experiment is designed to determine whether a drug or mutation influences one specific interaction between two proteins, most of the unaffected interactions will still be detected in the samples with the drug or mutation. To ensure that the scores of both the wild-type and conditional sample are not incorrectly penalized by their shared interactions, we add exclusion rules between the (bait, bait + condition) pair and the (bait + condition, bait) pair (see Fig. 8.19.3B).

Highly homologous baits: When two baits in a data set share a more than expected degree of sequence identity, they likely share a significant number of interactions too. The most common examples are intra- or inter-species homologs, different isoforms of the same protein, and cleaved protein products from poly-proteins.

Complex subunits: When subunits of a stable multi-subunit complex are all used as baits, the remaining subunits of the full complex will be identified recurrently in all samples.

Beyond these obvious cases, we recommend that readers not get carried away with specificity exclusions. Only data sets that have a high percentage of baits that match the conditions above should be scored with a specificity exclusion list. If there *might* be a set of common interactions between baits that could throw off the specificity score, then score the data without excluding the baits from each other and check whether the MiST scores are affected by a low-specificity component.

Advanced Parameters

MiST training with gold-standard interactions

The training file contains known true interactions that were identified in the data. This file should be tab-delimited, listing baits with the bait name described in the `keys` or `collapse` file and preys with their name in the `prey_colname` column of the `data` file. MiST will label these bait-prey pairs as a positive interactions set and compile a 100 times larger negative interaction set randomly selected from the data.

MiST will then run a simulation that cycles through all possible assignments of the three weights with 0.01 increments that together sum to 1, and 0.01 increments of the combined score threshold between 0 and 1. In every simulation cycle, MiST scores using these weights are computed and the subset greater than the threshold is compared to the positive and negative set. We compare the performance of the simulated weights at a given threshold by computing the precision and recall rates along with the 'f1 score' to measure overall accuracy.

Finally, the combination of weights with the best f1 score will be selected to compute the MiST scores, and an additional file with the summary of all simulations is written to the `output_dir`.

Suggestions for Further Analysis

Here we describe a number of suggestions for further analysis, starting from the MiST scores output file.

Protein interaction networks

After MiST outputs the scored list of bait-prey pairs and the appropriate threshold to select the high-confidence pairs is determined, the next step is usually to convert this filtered list into a more visual representation as a protein interaction network. Cytoscape [Shannon, 2003; *UNIT 8.13* (Su et al., 2014)] is by far the most popular piece of software used to create such networks. An in-depth tutorial on how to create interaction networks using AP-MS data is recently reviewed by Morris et al. (2014).

Connecting protein complexes

Unlike yeast-two-hybrid, which is binary in nature, interactions identified by AP-MS can be either in direct contact with the bait or indirectly mediated through members of the same protein complex. By overlaying published interaction data as edges between two 'prey' proteins, it is easier to see which protein complexes a particular bait protein interacts with. There are many online resources that collect and curate large-scale interaction data from different experimental sources, such as Biogrid (Stark et al., 2006), STRING (Mering, 2003), Corum (Ruepp et al., 2009), and Compleat (Vinayagam et al., 2013), that allow queries with a list of proteins to return all observed interactions between them (see Fig. 8.19.4B). If the bait-prey list is imported into Cytoscape, this process can be done automatically by using the BisoGenet plugin (Martin et al., 2010).

Biological annotation enrichment

To make sense out of larger protein-interaction data sets, it is useful to annotate all identified proteins with meaningful terms describing their biological function, domain composition, pathway involvement, disease involvement, and cellular localization. Again, there are many well-established online resources that collect and curate protein annotation terms such as GO [Gene and Consortium, 2000; *UNIT 7.2* (Blake and Harris, 2008)], KEGG [Ogata et al., 1999; *UNIT 1.12* (Tanabe and Kanehisa, 2012)] or PFAM [*UNIT 2.5* (Coggill et al., 2008); Finn et al., 2014]. Next, these annotated protein lists can then be analyzed to test whether ontology terms are overrepresented in the full set or specifically with respect to a single bait protein. The Cytoscape software can be a valuable tool for this purpose too, because plugins such as BINGO (Maere et al., 2005) take care of network annotation and enrichment tests in a few easy steps. Alternatively, complete lists of scored protein interactions or shorter lists of interactions above a certain threshold can be uploaded to online tools such as DAVID (Dennis et al., 2003) or GORILLA (Eden et al., 2009) respectively.

Acknowledgement

E.V., J.V.D., and N.K. are supported by NIH grants P50 GM082250, P50 GM081879, P01 AI090935, P01 AI091575, P01 AI106754, R01 GM084279, U19 AI106754, and the DARPA grant HR0011-11-C-0094. A.S. is supported by NIH R01

GM083960 and P41 GM109824. P.C. is supported by a Howard Hughes Medical Institute Predoctoral Fellowship.

Literature Cited

Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R., Ara, T., Nakahigashi, K., Huang, H.-C., Hirai, A., Tsuzuki, K., Nakamura, S., Altaf-Ul-Amin, M., Oshima, T., Baba, T., Yamamoto, N., Kawamura, T., Ioka-Nakamichi, T., Kitagawa, M., Tomita, M., Kanaya, S., Wada, C., and Mori, H. 2006. Large-scale identification of protein-protein interaction of *Escherichia coli* K-12. *Genome. Res.* 16:686-691.

Blake, J.A. and Harris, M.A. 2008. The Gene ontology (GO) project: Structured vocabularies for molecular biology and their application to genome and expression analysis. *Curr. Protoc. Bioinformatics* 23:7.2:7.2.1-7.2.9.

Choi, H., Liu, G., Mellacheruvu, D., Tyers, M., Gingras, A.C., and Nesvizhskii, A.I. 2012. Analyzing protein-protein interactions from affinity purification-mass spectrometry data with SAINT. *Curr. Protoc. Bioinformatics* 39:8.15.1-8.15.23.

Choi, H., Larsen, B., Lin, Z.-Y., Breitkreutz, A., Mellacheruvu, D., Fermin, D., Qin, Z.S., Tyers, M., Gingras, A.-C., and Nesvizhskii, A.I. 2011. SAINT: Probabilistic scoring of affinity purification-mass spectrometry data. *Nat. Methods* 8:70-73.

Clauser, K.R., Baker, P., and Burlingame, A.L. 1999. Role of accurate mass measurement (±10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* 71:2871-2882.

Coggill, P., Finn, R.D., and Bateman, A. 2008. Identifying protein domains with the Pfam database. *Curr. Protoc. Bioinformatics* 23:2.5:2.5.1-2.5.17.

Cox, J. and Mann, M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 26:1367-1372.

Dennis, G., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. 2003. DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol* 4:P3.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. 2009. GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48.

Ewing, R.M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M.D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y.V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.P., Duewel, H.S., Stewart, I.I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.L., Moran, M.F., Morin, G.B., Topaloglou, T., and Figeys, D. 2007. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* 3:89.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L.L., Tate, J., and Punta, M. 2014. Pfam: The protein families database. *Nucleic Acids Res.* 42:D222-D230.

Gene, T. and Consortium, O. 2000. Gene ontology: Tool for the. *Nat. Gen.* 25:25-29.

Grigoriev, A. 2003. On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res.* 31:4157-4161. Available at: *http://nar.oxfordjournals.org/content/31/14/4157.full*.

Hughes, N.C., Wong, E.Y. K., Fan, J., and Bajaj, N. 2007. Determination of carryover and contamination for mass spectrometry-based chromatographic assays. *AAPS J.* 9:E353-E360.

Jäger, S., Cimermancic, P., Gulbahce, N., Johnson, J.R., McGovern, K.E., Clarke, S.C., Shales, M., Mercenne, G., Pache, L., Li, K., Hernandez, H., Jang, G.M., Roth, S.L., Akiva, E., Marlett, J., Stephens, M., D'Orso, I., Fernandes, J., Fahey, M., Mahon, C., O'Donoghue, A.J., Todorovic, A., Morris, J.H., Maltby, D.A., Alber, T., Cagney, G., Bushman, F.D., Young, J.A., Chanda, S.K., Sundquist, W.I., Kortemme, T., Hernandez, R.D., and Craik, C.S., 2011. Global landscape of HIV–human protein complexes. *Nature* 481:365-370.

MacLean, B., Tomazela, D.M., Shulman, N., Chambers, M., Finney, G.L., Frewen, B., Kern, R., Tabb, D.L., Liebler, D.C., and MacCoss, M.J. 2010. Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 26:966-968.

Maere, S., Heymans, K., and Kuiper, M. 2005. BiNGO: A cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21:3448-3449.

Martin, A., Ochagavia, M.E., Rabasa, L.C., Miranda, J., Fernandez-de-Cossio, J., and Bringas, R. 2010. BisoGenet: A new tool for gene network building, visualization and analysis. *BMC Bioinform.* 11:91.

Mellacheruvu, D., Wright, Z., Couzens, A.L., Lambert, J.-P., St-Denis, N.A., Li, T., Miteva, Y.V., Hauri, S., Sardiu, M.E., Low, T.Y., Halim V.A., Bagshaw, R.D., Hubner N.C., Al-Hakim, A., Bouchard, A., Faubert, D., Fermin, D., Dunham, W.H., Goudreault, M., Lin, Z.Y., Badillo, B.G., Pawson, T., Durocher, D., Coulombe, B., Aebersold, R., Superti-Furga, G., Colinge, J., Heck, A.J., Choi, H., Gstaiger, M., Mohammed, S., Cristea, I.M., Bennett, K.L., Washburn, M.P., Raught, B., Ewing, R.M., Gingras, A.C., and Nesvizhskii, A.I. 2013. The CRAPome: A contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* 10:730-736.

Mering, C.V. 2003. STRING: A database of predicted functional associations between proteins. *Nucleic Acids Res.* 31:258-261. Available at: *http://www.nar.oupjournals.org/cgi/doi/10.1093/nar/gkg034* [Accessed March 10, 2011].

Morris, J.H., Knudsen, G.M., Verschueren, E., Johnson, J.R., Cimermancic, P., Greninger, A.L., and Pico, A.R. 2014. Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions. *Nat. Protoc.* 9:2539-2554.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. 1999. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27:29-34.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. 2009. CORUM: The comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.* 38:D497-D501.

Shannon, P. 2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13:2498-2504. Available at: *http://www.genome.org/cgi/doi/10.1101/gr.1239303*.

Sowa, M.E., Bennett, E.J., Gygi, S.P., and Harper, J.W. 2009. Defining the human deubiquitinating enzyme interaction landscape. *Cell* 138:389-403.

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. 2006. BioGRID: A general repository for interaction datasets. *Nucleic Acids Res.* 34:D535-D539.

Su, G., Morris, J.H., Demchak, B., and Bader, G.D. 2014. Biological network exploration with cytoscape 3. *Curr. Protoc. Bioinformatics* 47:8.13:8.13.1-8.13.24.

Tanabe, M. and Kanehisa, M. 2012. Using the KEGG database resource. *Curr. Protoc. Bioinformatics* 38:1.12:1.12.1-1.12.43.

Vinayagam, A., Hu, Y., Kulkarni, M., Roesel, C., Sopko, R., Mohr, S.E., and Perrimon, N. 2013. Protein complex-based analysis framework for high-throughput data sets. *Sci. Signal.* 6:rs5-rs5.

Yu, X., Ivanic, J., Wallqvist, A., and Reifman, J. 2009. A novel scoring approach for protein co-purification data reveals high interaction specificity. *PLoS Comput. Biol.* 5:e1000515.

Internet Resources

https://github.com/everschueren/MiST/

The Github repository is the main online resource for this unit. We have opened the MiST repository to the public but currently it is only editable by approved collaborators. In the near future we will also update the webpage at http://modbase.compbio.ucsf.edu/MiST/ to reflect the protocols described in this manuscript.