User manual – CRAPome version 1.0
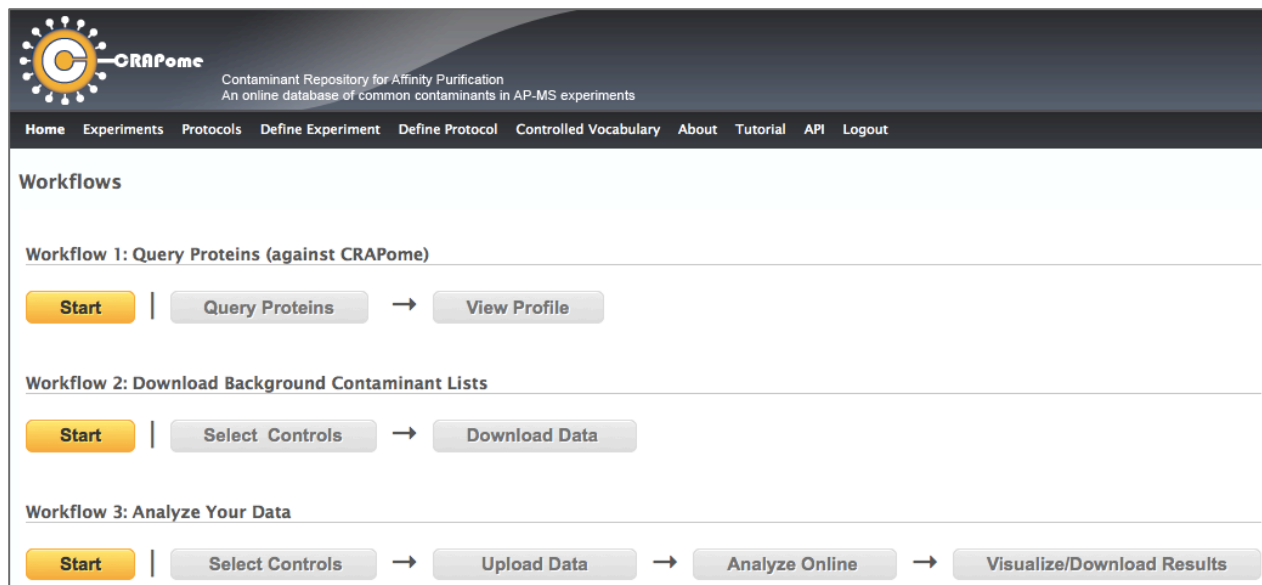
Prepared by Datta Mellacheruvu, Anne-Claude Gingras and Alexey Nesvizhskii, December 2012

1.   Introduction

This tutorial describes the Contaminant Repository for Affinity Purification, its web interface, and related tools, collectively referred to as CRAPome (www.crapome.org). The contaminant repository contains the lists of proteins identified from negative control experiments collected using affinity purification followed by mass spectrometry (AP-MS). Original MS data for each experiment are obtained from the data creator(s), generally as .raw or mzXML/mzML files (mgf files are also accepted if raw/mzXML data cannot be obtained for any reason). MS/MS data are processed by the repository administrator using a uniform data analysis pipeline consisting of an X!Tandem database search against the RefSeq protein sequence database, followed by PeptideProphet and ProteinProphet analysis (part of the Trans-Proteomic Pipeline). Each experiment in the CRAPome represents a biological replicate (technical replicates, i.e. repeated LC-MS/MS runs on the same affinity purified sample, or multiple fractions as in the case of 1D SDS-PAGE separation, are combined into a single protein list). Protein identifications are mapped to genes (see FAQ at the end of this document) and stored in a database along with their abundance information (spectral counts). CRAPome controls are associated with an experimental description via text-based protocols and controlled vocabularies. Users query the database via a web interface at www.crapome.org, using different user workflows (described below). Some functionalities in workflow 2 and workflow 3 require user registration. Note that until the manuscript is accepted for publication, access to the database is password-protected: login access can be requested at www.crapome.org. The database currently contains data from human AP-MS experiments only: as the database expands, additional species will be added. As of December 2012, the database contains ~350 experiments generated using ~75 unique protocols that were deposited by 11 laboratories.

2.   CRAPome welcome screen

The users of the repository access information stored in the database by first selecting one of the three workflows shown in **Fig. 2.1**, and described in detail below. A number of additional options are available from the menu bar; visible options depend on user status (e.g. end user, data contributor/annotator, admin).

Figure 2.1. CRAPome welcome screen. The three current user workflows are displayed; select the desired workflow by pressing "Start".

3. Workflow 1: Query selected proteins

This workflow allows the user to query for selected protein(s) of interest and view their profiles across different negative control experiments.

> **Step 1:** Paste a list of protein(s) or gene entries in a tab, comma or new-line separated format and click on 'submit' as shown in **Fig. 3.1** (compatible formats are shown in the figure legend).

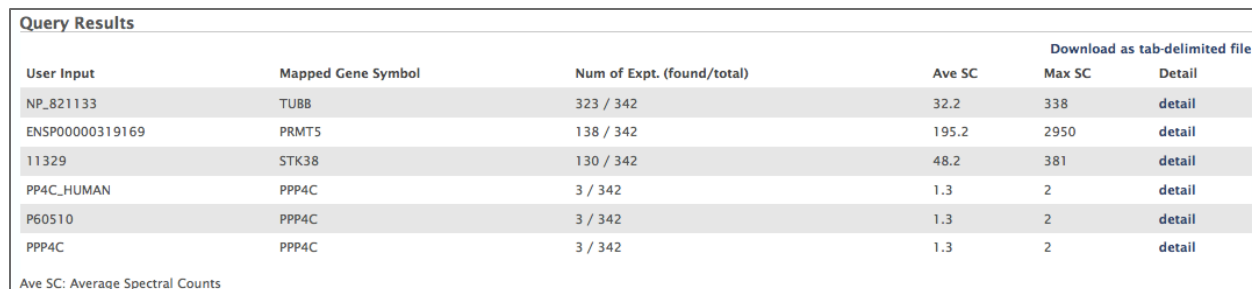**Figure 3.1. Protein/Gene identifiers compatible with the CRAPome.** *From top to bottom*: RefSeq protein ID, Ensembl protein ID, NCBI Gene ID, Uniprot entry name, Uniprot entry ID, HGNC gene symbol.

> **Step 2:** The query results are displayed as shown in **Fig. 3.2**.

User Input	Mapped Gene Symbol	Num of Expt. (found/total)	Ave SC	Max SC	Detail
NP_821133	TUBB	323 / 342	32.2	338	detail
ENSP00000319169	PRMT5	138 / 342	195.2	2950	detail
11329	STK38	130 / 342	48.2	381	detail
PP4C_HUMAN	PPP4C	3 / 342	1.3	2	detail
P60510	PPP4C	3 / 342	1.3	2	detail
PPP4C	PPP4C	3 / 342	1.3	2	detail

Ave SC: Average Spectral Counts

Figure 3.2. Query results. The first column shows the list of entries submitted by the user, while the second column lists the Gene Symbols mapped to the entries. The third column details the number of experiments in the database where the selected gene/protein was detected (with at least one peptide having PeptideProphet probability of 0.9 or higher); the total number of experiments in the CRAPome is also listed. The fourth and fifth columns list the averaged spectral counts and maximal spectral counts for the selected gene/protein across the experiments in which it was identified. The last column provides a link to the detailed profile for each of the selected genes/proteins. Note that this summary page can be downloaded as a tab-delimited file. From a quick survey of the results, PPP4C (protein phosphatase 4, catalytic subunit) does not appear in a high percentage of experiments (only 3 out of 242 in total), nor with high spectral counts, across the CRAPome, suggesting that it is unlikely to be a common contaminant in AP-MS studies. By contrast, TUBB (tubulin), PRMT5 (protein arginine methyltransferase 5) and STK38 (serine threonine kinase 38, also known as NDR1) are frequently detected, and often with a high number of spectral counts, indicating that they may be contaminants.

> **Step 3:** When the user clicks on the 'detail' link, a profile of the protein (in the CRAPome repository) is shown, as in **Fig. 3.3** and **3.4**. At the top of the page are graphical summary views of the data. The profile on the left shows the abundance distribution across experiments (i.e., how many CRAPome controls report this protein in the spectral count ranges 1-2, 3-5, 5-15, and so on; **Fig. 3.3**, left panel). If there are many experiments that report a protein in the higher spectral count ranges, one can suspect that it has a greater propensity to be a contaminant. Also shown (on the right) are the frequency of identification across groups of experiments selected based on the controlled vocabularies used to organize the data. This figure helps to provide an overview of (experimental) conditions in which this protein is likely to be a background contaminant (**Fig. 3.3**; right panel).
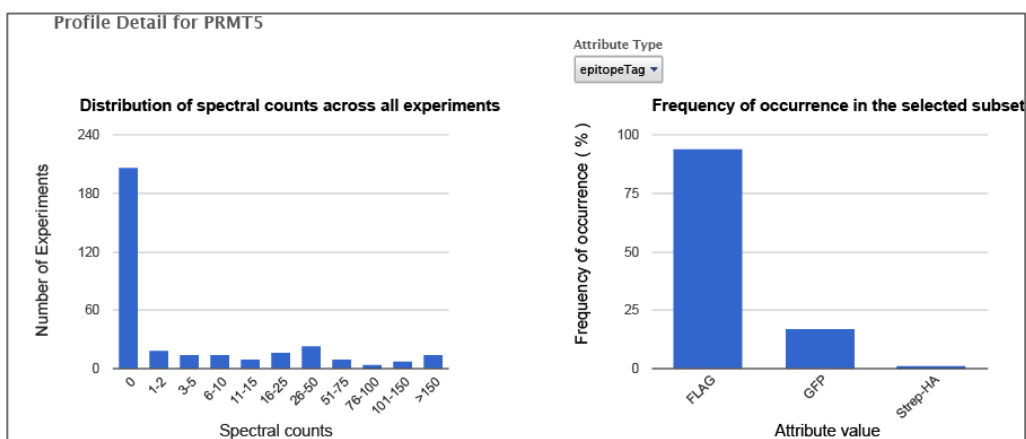
Figure 3.3. Detailed view of PRMT5 as a contaminant across the CRAPome. *Left*: abundance distribution of PRMT5 (spectral counts) across all experiments. While ~200 AP-MS analyses revealed no peptides for PRMT5, this protein was detected with a large spread of spectral counts across the remaining experiments. *Right*: distribution of the identification of PRMT5 across different epitope tag purifications. In this case, PRMT5 is frequently detected across FLAG purifications (94%; mouse over the bar to see the frequency listed as fraction of total, 124/132), but much less so across GFP (11/64) or Strep-HA (2/134) purifications, indicating that its contaminant propensity may be linked specifically to the FLAG epitope or the anti-FLAG antibody. Only those attribute values are shown on the plot that have at least one associated experiment in CRAPome where the queried protein was identified (e.g. PRMT5 was not identified in any TAP experiments, and thus TAP is not shown). Note that this graph can also be redrawn to show the distribution for the type of affinity support, cell line, or subcellular fractionation, to further explore contaminant behavior.

In addition to these summary views, the actual identifications of the protein in each experiment in the CRAPome repository (along with the identification scores and spectral counts) are listed below the figures in a tabular format (**Fig. 3.4**). The protein abundance distribution for each control (with protein abundances measured by their spectral counts) is also shown in a small box plot-like figure. The grey bar represents the spectral counts for the protein of interest. The background bands, from light yellow to dark orange colors, represent the 1st, 2nd, 3rd and 4th quartiles, respectively. When the grey bar representing the protein spectral counts is in the dark orange area, this protein is amongst the most abundant proteins in the corresponding CRAPome control.

Figure 3.4. Detailed view of the Experiments (column "Expt. Name") in which PRMT5 was detected, alongside its spectral counts, linked protocol and controlled vocabularies. The columns "Conf. Score" and "Conf. Score (unique)" refer to the identification of PRMT5 in the mass spectrometry experiment, and the confidence that the detected protein/gene is PRMT5 and not another protein (the values for these two scores are the maximum PeptideProphet probability across all peptides mapping to that protein, or only across all peptides unique to that protein, respectively). Lastly, the column "Spread" shows the protein identified with the maximal spectral counts as a colored bar (separated in quartiles; max counts are shown – click on "view legends" for details) and the spectral counts for PRMT5 are shown as a grey bar. Click on the "View legend" button to see the color mapping (**Fig. 3.5**).
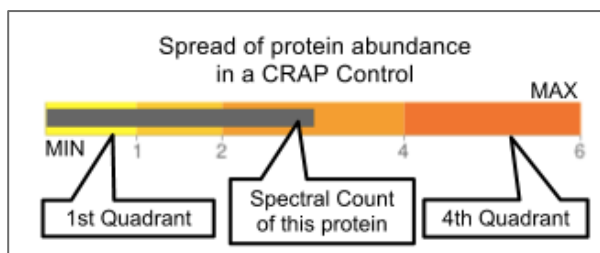
Figure 3.5. Color-coding map of the quartile information shown in the "Spread" column of Fig. 3.4.

The numbers in the Experiment Name and Protocol columns (as shown in **Fig. 3.4**) are hyperlinks. By clicking on these links, information about the experiment and the experimental protocol are shown in new windows as in **Fig. 3.6**.

Experiment #1

Experiment Name

CC1

Constituent Files

ACG_10435_control_benzonase_IP_1_LTQACG_25FEB2011

Admin Comments

ACG_10435_control_benzonase_IP_1_LTQACG_25FEB2011

Protocol Name

- 293 Flp-In FLAG mChip Orbitrap

Protocol #

15

Attributes

- Organism : human

Protocol #15

Protocol Name

293 Flp-In FLAG mChip Orbitrap

Protocol Comments

Gingras lab - version 1.0

Experiments(s)

- CC1 - ACG_10435_control_benzonase_IP_1_LTQACG_25FEB2011
- CC4 - ACG_11302_empty_3XFLAG_chAP_Orbi_11JUN2011

Attributes

- Organism : human
- Cell/tissue type : HEK293
- Cell/tissue subtype : HEK293 Flp-In T-REx
- Drug treatment : none
- Subcellular fractionation : total lysate+chromatin
- Epitope tag : FLAG
- Control protein : tag alone
- AP steps : single

Figure 3.6. Experiment (left) and Protocol (right) information. Only the top of the page is show; scroll down for additional information.

By clicking on the 'detail' link in the Spectral Count column, more detailed information about the protein is shown, including the list of identified peptides, along with their probabilities and spectral counts (**Fig. 3.7**).

Protein Detail

- GI Number: 6005814
- RefSeqID : NP_009202.1
- Gene Symbol(s): STK38
- Experiment: CC138
- Peptides:

Show 10 entries   Search: _____

Sequence	Charge	Peptide Probability	Number Spectra
DIKPDNLLLDSK	2	0.9984	6
DTGHVYAMK	1	0.9939	5
DTGHVYAMK	2	0.9985	4
DWVFINYTYK	2	0.9853	12
ETLTFPPEVPISEK	2	0.9973	22
FC[160]C[160]EWEHR	2	0.9925	4
IGAPGVEEIK	1	0.9432	1
IGAPGVEEIK	2	0.9989	20
LGLEDFESLK	2	0.9989	38
LSDFGLC[160]TGLK	2	0.9989	50

Showing 1 to 10 of 17 entries   Previous Next

Figure 3.7. Peptide summary view for a selected protein (STK38) in one experiment (CC138)

The summary table can be downloaded as a tab-delimited file (**Fig. 3.7**).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	id	ipName	protID	numSpec	exptSum	exptFreq	exptMin	expt2ndQ	expt3rdQ	exptMax	protocol	epitopeTag	cellLine	affinitySupport	subcellularFrac	proteinProb	maxUniqPepProb	maxIniProb
2	81608	CC189	PRMT5	2950	10077	137	1	60	24	2950	66	FLAG	HEK293	agarose	total cell lysate	0.9998	0.9998	0.9998
3	83271	CC192	PRMT5	2853	10744	137	0	68	26	2853	66	FLAG	HEK293	agarose	total cell lysate	0.9998	0.9998	0.9998
4	81480	CC190	PRMT5	2830	9861	137	1	70	26	2830	66	FLAG	HEK293	agarose	total cell lysate	0.9998	0.9998	0.9998
5	81060	CC191	PRMT5	2521	7887	137	0	46	15	2521	66	FLAG	HEK293	agarose	total cell lysate	0.9998	0.9998	0.9998

Figure 3.8. Summary table. The columns are as follows: A) ID (a unique identifier for the detection of a given protein across the database; B) ipName, the unique identifier for the experiment in the CRAPome; C) protID, the mapped Official Gene Symbol; D) numSpec, the spectral counts associated with PRMT5 in the experiment; E) exptSum, the sum of all spectra assigned in the experiment; F) exptFreq, how often PRMT5 has been detected across the CRAPome; G-J) quartiles (defined by their spectral count boundaries); K) protocol used for the experiment; L-O) Selected controlled vocabulary values; P-R) protein probability values, defined as described above.

4. Workflow 2: Create contaminant lists

This workflow allows the user to download subsets of data from the CRAPome repository. Each control experiment (CRAPome Control, or CC) is assigned a unique identifier (CC1-CCx), linked to a protocol, and annotated with standard vocabulary (such as the epitope tag type, cell line, affinity matrix, etc.). These attributes can be used to filter the list of available CRAPome controls. These filters are available on the left as shown in **Fig. 4.1**.

> **Step 1:** Use the filters on the left (**Fig. 4.1**) to narrow down the list of negative controls.
> **Step 2:** "Add" each desired CRAPome control of interest by clicking the button in the table. If desired, select the "Add All" button instead. Added controls will appear in the "Selected controls" box on the right. *There is a limit of 60 controls that can be selected at the same time*. A link at the top of the home page provides an option to download the entire database content as a tab-delimited text.
> **Step 3 (*optional*):** Give this list of selected controls a name and save it for future use (note that this option is restricted to registered users). If you wish to reload a previously saved list, you can do so by clicking the "load" link.

Figure 4.1. Selection of the controls in User Workflow 2. *Left panel*: available controlled vocabularies to filter the list of controls (selecting different options across categories is equivalent to an "*and*" function; selecting multiple boxes within a category corresponds to an "*or*" function, as shown here with the selection of HA tag and HaloTag experiments from the

database. *Middle*: table of the controls that passed the selected filters. Clicking on the control name (first column) or the protocol name (third column) displays extended information. The second column lists the number of proteins identified in each of the controls. A limited view of the controlled vocabularies is shown in columns 4-8. Controls are added to the list by clicking on the "Add" or "Add All" buttons in the last column (note that added controls can also be manually removed). Alternatively, controls can be added by loading a previously saved list (*right, bottom*). Once the desired controls are selected, the data table can be generated by pressing the blue "Next" button at the top right of the page.

Step 4: Click on "Next" button at the top of the page to view and/or download the data matrix (**Fig. 4.2**). The file will be downloaded as an Excel compatible table.

Step 5 (*optional*): Specific proteins can be queried in the data matrix by typing partial or complete gene name (wild cards are automatically added at the beginning and end). The data matrix can be downloaded as a tab delimited file using the "download data matrix" option.

Figure 4.2. Detailed table output for the User Workflow 2 – limited for search term "PRM". The complete list of the proteins identified across the selected CRAPome controls is shown by default. It can be also restricted to a selected search term (here: "PRM"). The occurrence of PRMT5 (NUM_OCCUR), 5 experiments out of the 5 selected, and PRMT1, in two out of the 5 experiments, are shown, alongside their averaged spectral counts (Avg SPC) and the spectral counts in each of the 5 selected experiments (CC35, CC36, CC38, CC40 and CC41; highlighted in pink cells). In this table, the "ID" column represents the unique identifier for the detection of a specific protein in a given experiment.

5.   Workflow 3: Use the CRAPome to analyze your data.

This workflow allows the user to process his/her data online using the CRAPome controls and the scoring tools implemented within the system. This workflow is only available to registered users. The minimum requirement is for the user to submit information regarding one bait (one sample), though we strongly advocate the use of biological replicates for the bait, and recommend that the user also uploads his/her own negative control runs.

Step 1: Select the CRAPome database controls that are most similar to the user data using controlled vocabularies and detailed protocols as shown for workflow 2 above (see **Fig. 4.1**). Selected controls can be saved as a list and reloaded as needed as in workflow 2. Press the blue "Next" button to navigate to the next page.
Step 2: Upload user data (See **Fig. 5.1**). The data should be formatted as per instructions on the webpage (also see **Fig. 5.2**). Once uploaded, the data appear in the 'user data' section below.

Figure 5.1. Upload user data. The top of the page enables browsing to upload the user data prepared in a comma-separated values (CSV) file (see **Fig. 5.2**). The table should not have headers and should consist of four columns: 1) Bait Name; 2) AP Name; 3) Prey Name; 4) Spectral Counts. Negative control analyses should be labeled "CONTROL" in the "Bait Name" column. The "Prey Name" can be either a RefSeq protein ID, Ensembl protein ID, NCBI Gene ID, Uniprot entry name, Uniprot entry ID or HGNC gene symbol. For mapping purposes, we strongly suggest also using one of these identifiers for the "Bait Name". Different "AP Names" will automatically be merged for analysis if they are assigned to the same "Bait Name". The bottom left of the page lists the User data that was uploaded while the bottom right lists the selected CRAPome controls.

Figure 5.2. Sample column format for upload. Column A is the BAIT name, column B the identifier for the experiment (AP name), column C the PREY identifier (here RefSeq protein ID) and column D is the spectral count for the PREY in the AP.

Step 3 (optional): If the user would like to exclude some of his/her data from the analysis, it can be done at this stage by clicking on 'remove' button. Similarly, one can go back and add/remove CRAPome controls. For a quick preview of the data matrix, click on 'Preview Data Matrix'. After the analysis is complete, the data can be deleted by clicking on 'clear uploaded user data' (See **Fig. 5.1**).

Step 4: Proceed to the analysis section by clicking on 'Next'. Here, Fold Change calculations and SAINT probability scoring can be used to generate ranked lists of bait-prey interactions.

Step 5: Select desired scoring options for Fold Change calculations (**Fig. 5.3**). Two different Fold Change calculations are generated by default. The first one (FC-A; standard) estimates the background by averaging the spectral counts across the selected controls while the second one (FC-B; stringent) estimates the background by combining the top 3 values for each prey. Combining scores from biological replicates of a bait purification is performed in FC-A by a simple averaging, while FC-B performs a more stringent geometric mean calculation. These parameters are preselected by default, but may be modified by the user as required. The user can also specify what set of controls to use (user controls alone or in combination with selected CRAPome controls).

Step 6 (optional): The user can specify whether to run SAINT or not, and which SAINT options ('lowMode', 'minFold', 'norm') should be employed. For details regarding SAINT and the options, please

refer to the Choi et al., Current Protocols in Bioinformatics (PMID 22948730). As with the Fold Change calculations, the user may select which controls to use, and how replicates should be combined. Note that if the number of controls is greater than 10, SAINT generates 10 "virtual controls" by selecting the 10 highest counts for each protein.

Figure 5.3. Interface for data analysis using CRAPome. The left panel displays the interface for running the Fold Change empirical score; CRAPome automatically generates two FC-scores as part of the analysis, the second one being designed as a more stringent score by default. Parameters can be modified as described above. The right panel displays the optional SAINT scoring, associated with model options.

Step 6: Once the desired options are selected, press "Run Analysis". The new entry will appear at the top of the 'Analysis Results' list (the list includes all previous jobs run by the user). Initially, the Status (last column) of the new job will be shown as 'queued' (followed by 'running' and then 'complete'). The column "Score Options" lists selected options for the Fold Change calculations for both the primary (FC-A; here labeled S1) and the secondary, more stringent (FC-B, here labeled S2) scores. SAINT options (when applicable) are listed in the next column.

Figure 5.4. Job status. This table lists all analyses performed by the user, alongside the selected options for the scoring.

Step 7: Refresh the web page periodically by clicking on 'Refresh'. When the job is finished and the results are ready to be viewed, the Status will change to 'complete'. A link called 'view results' will appear. The user will also receive an email notification with a link to the results page.

Step 8: Click on 'view results' link to view the results. At the top of the page, you will see graphical views of the data that summarize the results for each of the baits analyzed, or for all baits at once. The left panel compares SAINT (when run) to FC-A; when SAINT is not used, this panel displays a comparison between FC-A and FC-B (see **Fig. 5.5**). In both cases, the left panel describes the Receiver Operating Characteristic (ROC) analysis of the scoring (benchmarked to the interactions reported in iRefIndex). This visualization can assist in deciding which scoring function to use on the data. The middle panel displays a histogram of the interactions reported in iRefIndex versus those not reported, at different bins of SAINT probability or FC-A score when SAINT was not run. Finally, the panel on the right compares two different scores (by default, SAINT and FC-B if SAINT is used; FC-A vs. FC-B otherwise) at the level of individual proteins. Mousing over any of the graphs will display relevant information (e.g. gene names).
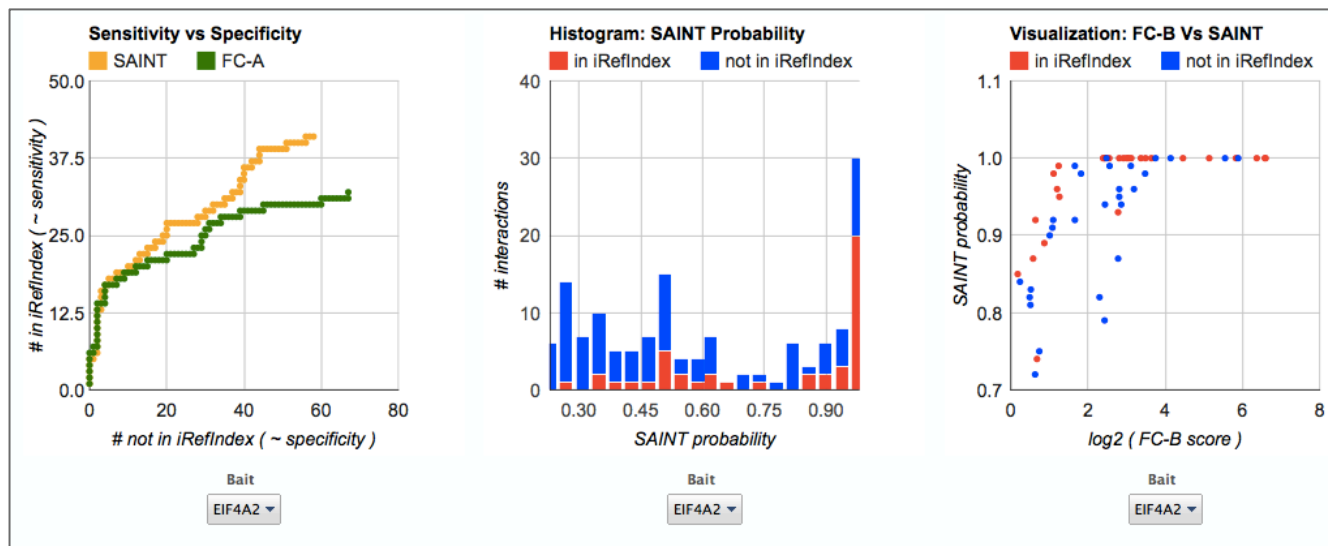
Figure 5.5. Graphical views of the data. Three different graphs are drawn for each analysis. Color coding on the middle and right graphs corresponds to the information present (red) or absent (blue) from the iRefIndex. When more than one bait was employed, use the dropdown menu below each graph to generate figures for the individual baits, or for all baits combined.

Step 9: The results can be viewed online in a matrix form or downloaded in a tabular format (see **Fig. 5.6**)

Baits in this data: B1 = EIF4A2, B2 = RAF1, B3 = WASL, B4 = MEPCE

Bi_SP: SAINT Probability of proteins co-purifying with bait *i*

Bi_FC_A: Primary FC Score (FC-A) of proteins co-purifying with bait *i*

Bi_FC_B: Secondary FC Score (FC-B) of proteins co-purifying with bait *i*

Bi_iREF: Interactions report (1: in iRefIndex, 0: not in iRefIndex)

Bi_Rj: Spectral counts of proteins co-purifying with bait *i* in replicate *j*

UCx: Spectral counts of proteins in user control *x*

CCy: Spectral counts of proteins in CRAPome control *y*

TIP: Mouse over the header names to see the full bait names.

Download data matrix
Download raw SAINT results

Show 10 entries   Search:

ID	PROTID	GENENAMES	B1_FC_A	B1_FC_B	B1_PROB	B1_IREF	B2_FC_A	B2_FC_B	B2_PROB	B2_IREF	B3_FC_A	B3_FC_B	B3_PROB
1	NP_001958	EIF4A2	450.06	107.32	0	1	0	0	0	0	0	0	0
2	NP_006752	YWHAE	0.71	0.19	0	0	127.55	34.13	1	1	0	0	0
3	NP_002871	RAF1	0	0	0	0	481.12	481.12	0	1	0	0	0
4	NP_001407	EIF4A1	149.9	35.2	1	1	0	0	0	0	0	0	0
5	NP_001017963	HSP90AA1	0.21	0.06	0	0	15.3	4.49	1	1	0.69	0.2	0
6	NP_003932	WASL	0	0	0	0	0	0	0	0	1280.35	1280.35	0
7	NP_055521	SART3	0	0	0	0	0	0	0	0	0	0	0
8	NP_003741	EIF3A	26.93	12.49	1	1	0.22	0.1	0	0	0.94	0.44	0
9	NP_006073	TUBA1B	0.12	0.05	0	0	2.34	1	1	0	0.52	0.22	0
10	NP_821133	TUBB	0.1	0.06	0	0	2.41	1.33	0.99	0	0.83	0.46	0.06

Showing 1 to 10 of 798 entries   ◀ Previous Next ▶

Figure 5.6. Results table. Preys are listed in rows. The columns are as follows: ID, row identifier; PROTID, the protein name in the upload file; GENENAMES, the mapped Gene Name. The rest of the columns describe the data: B1 indicates Bait 1 (in this case, EIF4A2). FC_A and FC_B are Fold Changes A and B, respectively; PROB is the SAINT probability, and IREF denotes an interaction reported in iRefIndex (1 = reported; 0 = not reported). The table has many more columns than can fit in this window, including spectral counts for every replicate (R1 is replicate 1) of a bait purification, and for every selected control.