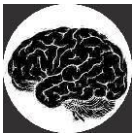
	SKILLS CENTER **STANDARD OPERATING PROCEDURE**	**A BIOFIZZ**  **PRODUCTON**
The CRAPome Database **Module Hours: 2**	**Effective Date: 4/12/2021** **Checked by A. Siclair**	**Revision # 1.0** **M. Guzie**

BACKGROUND

One of the most commonly used techniques for protein analysis is mass spectrometry. The approach of mass spectrometry involves generating ions from a substance of interest and then separating the ions by their mass-to-charge ratio and using the relative abundance of these ratios to determine molecular weight and structure (The Broad Institute, 2008). The earliest indications of such a technique were developed by J. J. Thomson in his work on the cathode ray tube. Thomson's studies aimed at determining the nature of cathode rays, streams of electrons observed in vacuum systems. To demonstrate the electron to be a particle, Thomson worked to build a tool which would measure both the mass and charge of the electron at the same time. Thomson was awarded a Nobel prize in 1906 for his studies showing the nature of the electron, and scientists now recognize his tool as the first form of a mass spectrometer (Griffiths, 2008).

Mass spectrometry was originally used in the early 1900s to measure the mass of individual atoms; in this endeavor, the technique was able to show the existence of isotopes (Griffiths, 2008). From then, the realization of the applicable uses of mass spectrometry in the analysis of larger, more complex molecules has brought the technique to the forefront of the study of biological macromolecules.

Applications of this technique are highly valuable in the field of proteomics. To study protein structure or protein interactions, scientists use affinity purification techniques coupled with mass spectrometry to identify the components of these purifications and the structure of isolated proteins (Mellacheruvu, et al., 2013). However, many proteins can bind non-specifically, or purifications which are done do not contain only the desired protein. Therefore, it is incredibly crucial yet often difficult for scientists to determine whether an isolated protein sample is binding specifically or is a non-specific contaminant.

The solution to this is to compile known data on negative control experiments. These are experiments containing antibodies or probes which should not bind and isolate specific proteins to determine if the proteins are binding and being isolated nonspecifically. The CRAPome is a database comprised of commonly found contaminant proteins when running protein isolations in biochemical affinity assays which was developed by the Nesvizhskii lab at the University of Michigan (Nesvizhskii, n.d.). While the name is seemingly fitting for a database compiling excess proteins in biochemical isolations, it is actually an acronym standing for Contaminant Repository for Affinity Purifications. The database compiles mass spectrometry negative control experiments with detailed protocols and methodologies cited by each contributor (Mellacheruvu, et al., 2013). The collection of these negative control mass

spectrometry experiments in one database can be extremely useful to scientists in the pursuit of running protein isolations and interpreting the purity of their purifications.

1. PURPOSE

The purpose of this procedure is to become familiar with the CRAPome, gain the ability to navigate the database, and understand why having a database of common contaminant proteins is useful.

2. SCOPE

This procedure applies to qualified skills center users.

3. RESPONSIBILITY

3.1. It is the responsibility of the user to understand and perform the procedure described in this document.
3.2. It is the responsibility of the user performing the procedure to fully document any deviations from the written procedure.
3.3. It is the responsibility of the user to become trained in the use of this application.

4. DEFINITIONS

4.1. Mass Spectrometry: An analytical technique which measures the mass-to-charge ratio of a sampled molecule(s). This method can be used to calculate the molecular weight. By calculating the molecular weight, unknown compounds can be identified. Other uses of this technique include quantifying the amount of a known compound present in a sample and determining the structure and chemical properties of a molecule (The Broad Institute, 2008).
4.2. Mass Spectrum: A graphical representation of the ion abundance versus the mass-to-charge ratio (m/z where m is the mass and z is the charge) of the ions separated in a mass spectrometer (Yury, et al., 2019).
4.3. Spectral Count: The total number of spectra identified for a specific protein; thus can be indicative of the relative abundance of the protein in a sample, if more spectrum are detected for a single protein (higher spectral count.)

4.4. Epitope: The part of an antigen (specific molecular structure which induces an immune response) which is recognized by antibodies, and thus allows for identification and tagging of proteins when using the corresponding antibody.
4.5. Epitope Tag: An epitope that is fused to a protein through recombinant genetic engineering so that the protein can be detected and localized using an antibody for the epitope.
4.6. Frequency of Occurrence: The quantitative value of how often the protein of interest is found to be present in negative control experiments, measured through the protein's spectral count. This indicates the frequency of how often the protein occurs as a contaminant. This value can be used to compare under what experimental conditions a contaminant protein is more or less likely to be found in the purification as a contaminant (for example, comparing the frequency of occurrence when using an agarose gel medium vs. a nano-magnetic medium may indicate which medium is more likely to be contaminated with that specific protein.)

5. MATERIALS/EQUIPMENT

5.1. A computer to perform the procedure.
5.2. The CRAPome database: https://reprint-apms.org/.

6. PROCEDURE

In this procedure, the CRAPome database will be detailed, including how to utilize the different workflows and find and extract useful information regarding contaminant proteins in biochemical purifications.

6.1. Using Workflow 1

6.1.1. Visit the CRAPome Database at REPRINT: https://reprint-apms.org/.
6.1.2. The REPRINT home page will display two large squares to either access the CRAPome database or to analyze data using other methods. Select the CRAPome square on the left to go to the database.
6.1.3. The two workflow options will be displayed. Select the desired workflow, beginning with Workflow 1, by selecting "Start." (**Figure 1**)

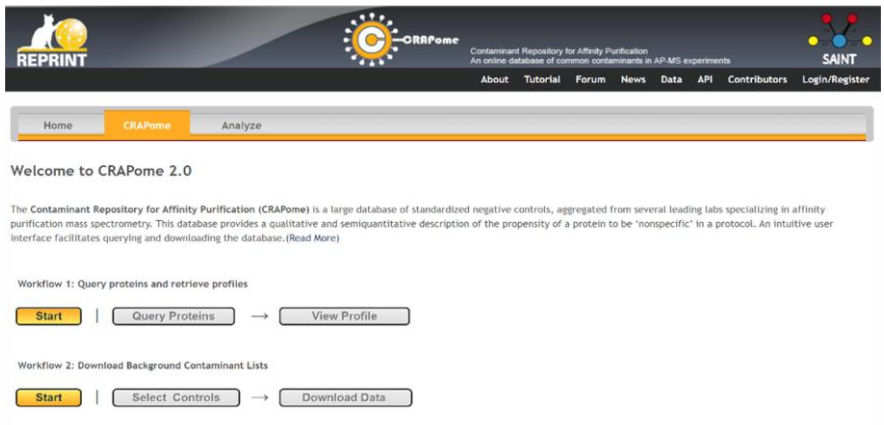


Figure 1

Figure 1: The home page for the CRAPome 2.0 database on REPRINT, displaying the two options for workflow.

6.1.4. Workflow 1 allows the user to search for proteins of interest and view their resulting profiles across the compiled negative control experiments. This can be done by pasting proteins or gene names into the search tab, separated by commas or new lines. Different formats are recognized for different organisms:
-For *Homo sapiens:* RefSeq protein ID, Ensembl protein ID, NCBI gene ID, Uniprot entry name, Uniprot entry ID, and HGNC gene symbol are recognized formats.
-For *Saccharomyces cerevisiae:* SGDID, Ensembl protein ID, RefSeq protein ID, and Uniprot ID/Accn are recognized formats.
-For *Escherichia coli:* Uniprot ID, associated gene name, and refSeqGI are recognized formats.
For this SOP, the Uniprot ID will be used for simplicity.

6.1.5. To find the UniProt ID for a protein, begin by opening another tab and going to the UniProt database at https://www.uniprot.org/.

6.1.6. Search the protein of interest by name in the search bar, making sure the UniProtKB (knowledgebase) is the database being searched. (**Figure 2**)

Figure 2

Figure 2: A UniProtKB search for the cytoskeletal protein actin.

6.1.7. Make sure that you are copying the entry ID for a protein from either *Homo sapiens, Saccharomyces cerevisiae,* or *Escherichia coli* because these are the three organisms which the CRAPome collects mass spectrometry experimental information for. Results can be filtered to a specific organism using the filter specifications on the left: under "Popular Organisms" for the protein searched, organisms with the most results for the query protein are displayed and can be selected, or a specific organism can be searched in the search bar under "Other organisms" by typing in the organism's name and selecting "Go." (**Figure 3**)

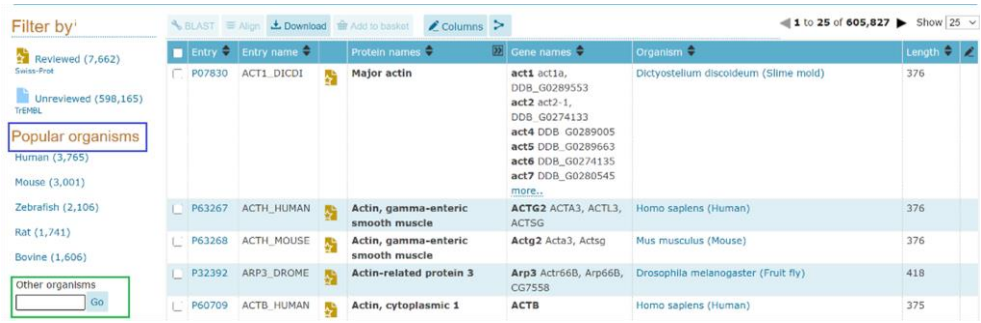


Figure 3

Figure 3: The UniProtKB search results for the cytoskeletal protein actin. Here, the results are unfiltered, displaying all available entries with some form of actin in the name, but results can be filtered on the left of the screen.

6.1.8. Once the results have been filtered to a specific organism from which the CRAPome contains data, copy the 6 character Entry code of the desired protein to be pasted into the CRAPome search box (**Figure 4**).

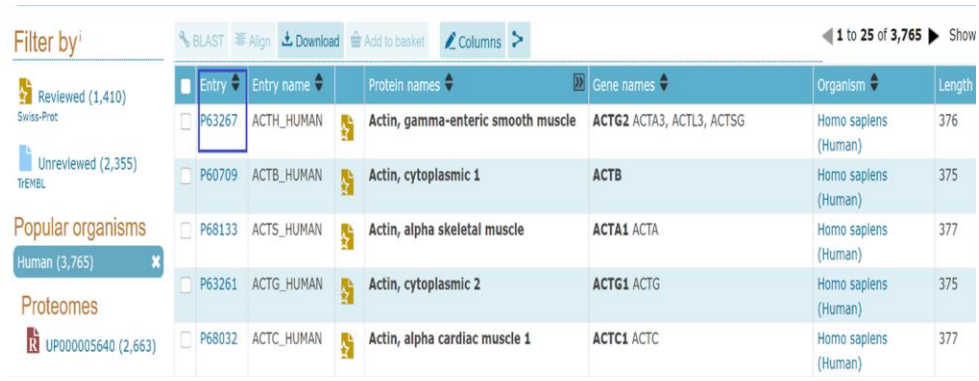


Figure 4

Figure 4: The UniProtKB search results for the protein actin, with results filtered to humans.

6.1.9. Return to the original tab with workflow 1 open and paste the entry code into the search box under “Enter Protein Identifiers.

6.1.10. Select submit.

6.1.11. The query results will be displayed. (**Figure 5**)

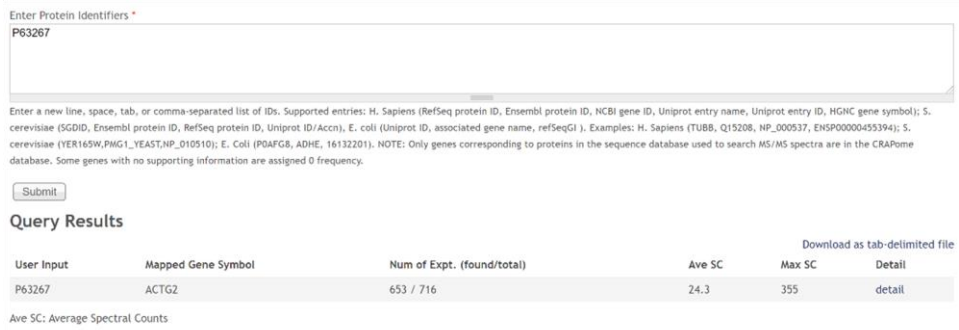


Figure 5

Figure 5: The query results for an actin protein in *Homo sapiens* (UniProt entry P63267). The first column lists the protein entries submitted in their respective format. The second column lists the associated gene symbol for the protein entry. The third column shows the fraction of the number of experiments where the protein was detected over the total number of experiments in the CRAPome database. The fourth column lists the averaged spectral counts. The fifth column lists the maximal spectral counts. The sixth and final column provides a link to a detailed profile of the protein entry.

Note what information about this protein can be taken away from the results. This version of actin is found in 653 of the total 716 mass spectroscopy experiments compiled in the CRAPome database (column 3). This indicates that this protein is found in many negative control mass spectroscopy experiments and therefore is likely a common contaminant. Identification of a protein's spectrum in a negative control experiment indicates the protein was found in the purification (as an undesired product) and is therefore likely a contaminant.

6.1.12. Select the "detail" link in the 6th column. This will display a profile of the protein comprised of CRAPome experimental data.

6.1.13. The top of the page shows two different graphical representations of data. (**Figure 6**)

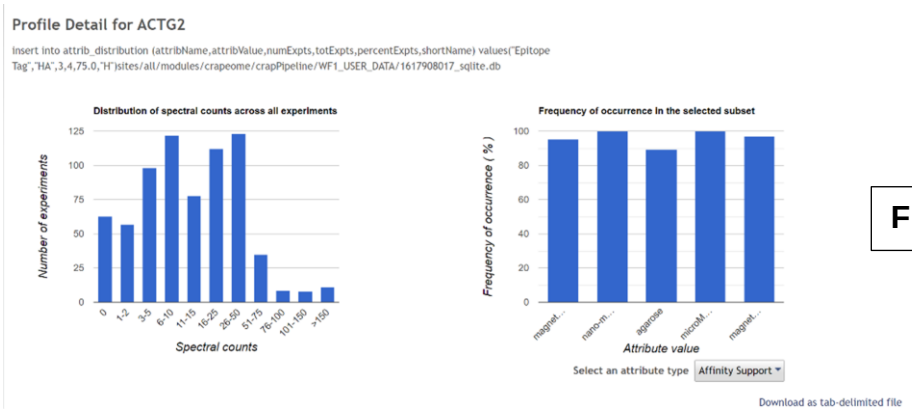


Figure 6

Figure 6: Two graphs representing the CRAPome data compiled for the ACTG2 actin gene. The graph on the left shows the protein's abundance in spectral counts (x) in different experiments (y). The graph on the right shows different experimental conditions/attributes (x) compared to the frequency that the protein occurs (y) in those experiments and can be altered to four different attribute types.

Note what information about this protein can be taken away from these results. The graph on the left shows how many CRAPome control experiments report different spectral count ranges (meaning ~75 experiments identified a spectral count of 11-15 for this actin protein.) While ~58 experiments showed 0 spectral counts for this protein, many other experiments showed enough abundance that this protein is likely to be a common contaminant. The graph on the right shows the frequency of occurrence of the protein over different attribute values; using different attribute values organizes the data in different ways which demonstrate what experimental conditions the protein is likely to occur in as a contaminant.

6.1.14. For the right graph, explore the different attribute types which can be used to view the data. The options are Affinity Support (shown in Figure 6), Cell Line, Epitope Tag, and Subcellular Fractionation.

-Affinity Support shows the frequency of occurrence in different mediums used for protein isolation experiments, such as agarose gel.

-Cell Line shows the frequency of occurrence in different cell culture types, such as HeLa cells.
-Epitope Tag shows the frequency of occurrence when using different epitopes for protein detection and isolation. (**Figure 7**)
-Subcellular Fractionation shows the frequency of occurrence in different isolated cellular components.

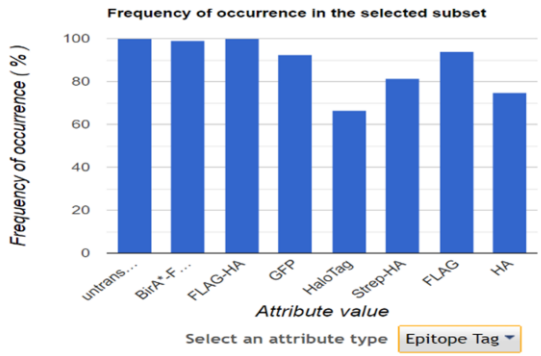
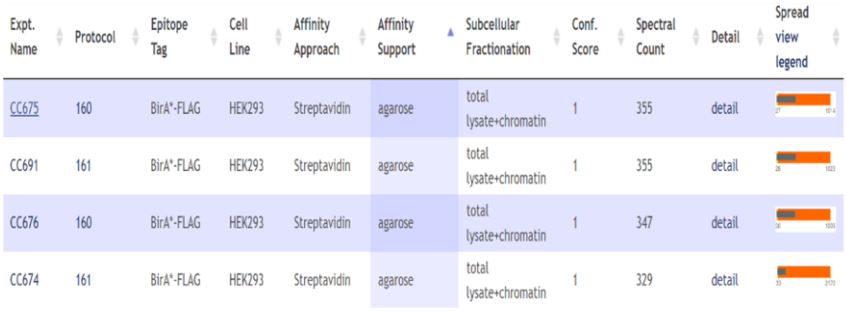


Figure 7

Figure 7: The frequency of occurrence of the ACTG2 protein relative to different epitope tag types. This shows that the FLAG-HA and BirA*-FLAG types have a higher occurrence of the protein, and therefore may be more prone to non-specific binding of this protein than other epitope tags such as GFP.

6.1.15. Scrolling down beneath the two summary graphs will show each individual experiment in the CRAPome database which identified the protein. This includes the details of the experiment such as what protocol used, different affinity attributes used, identification scores, spectral count, and a small figure “Spread” which summarizes the protein abundance distribution. Note that clicking on the experiment name, protocol, or details tab will provide additional information, but the user must have an account to view

additional information on the experiment and protocol. This is not true for details. (**Figure 8**)



Expt. Name	Protocol	Epitope Tag	Cell Line	Affinity Approach	Affinity Support	Subcellular Fractionation	Conf. Score	Spectral Count	Detail	Spread view legend
CC675	160	BirA*-FLAG	HEK293	Streptavidin	agarose	total lysate+chromatin	1	355	detail	
CC691	161	BirA*-FLAG	HEK293	Streptavidin	agarose	total lysate+chromatin	1	355	detail	
CC676	160	BirA*-FLAG	HEK293	Streptavidin	agarose	total lysate+chromatin	1	347	detail	
CC674	161	BirA*-FLAG	HEK293	Streptavidin	agarose	total lysate+chromatin	1	329	detail	

Figure 8

Figure 8: The detailed view of individual experiments found in the CRAPome repository. The first column identifies the specific experiment name. The second column identifies the protocol used. The third-seventh column identify different affinity attributes used in the experimental methodology. The eighth and ninth column give confidence scores that the protein identified is indeed the protein of interest and not another protein. The 10th column gives the number of spectral counts identified for the protein of interest in the specific experiment. The 11th column provides a link to more detailed information on the specific experiment. The 12th and final column provides a visual representation of the protein abundance distribution.

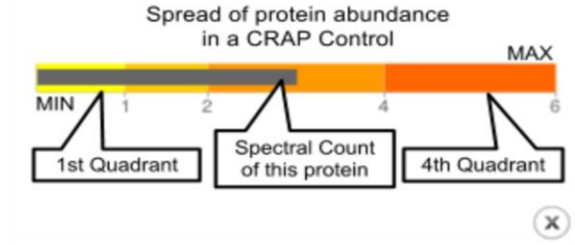


Figure 9

Figure 9: The graphical organization of the “Spread” protein abundance distribution graph. The grey bar represents the spectral count for the protein of interest and the colored portion represents the total amount of spectra in the sample (divided into quadrants of differing colors.)

6.1.16. Select the view legend option under the “Spread” column to see how the graph is organized. (**Figure 9**) This graph shows how much of the protein of interest is present

relative to the complete sample by comparing spectral counts of the protein of interest (represented by the grey bar) compared to total spectral counts in the sample, the sum of all the protein spectra present in the negative control experiment (colored portion). When the grey bar extends further right into the colored region, this indicates that the protein of interest is one of the most abundant in that specific CRAPome control experiment.

6.2. Using Workflow 2

6.2.1. Return to the CRAPome 2.0 database to use workflow 2: https://reprint-apms.org/?q=chooseworkflow. Workflow 2 can be used to download subsets of data from the CRAPome repository.

6.2.2. Choose workflow 2 by selecting the “Start” option.

6.2.3. Select a dataset (organism and experiment type) to download from. Here, the “H. sapiens – all” dataset will be used.

6.2.4. Select “Start” to view the dataset.

6.2.5. Use the green filter options on the left to narrow down the list of negative controls by selecting the green box for the specific experimental attribute and then checking the options to be included in the resulting control experiments. (**Figure 10**)

Figure 10

Figure 10: The CRAPome control experiments yielded from using filters for three different cellular attributes (experiments utilizing HeLa cells, cytosolic fractions, and the FLAG epitope tag) are displayed.

Note that multiple options can be selected under each attribute filter, but only one option is selected here. By selecting multiple options under a specific attribution, controls containing either one of the options will be shown (for example, selecting HeLa cells and Jurkat cells will show controls which utilize either cell type.) There are also other filter options below the ones shown which can be used.

6.2.6. Add each desired CRAPome control to a dataset by selecting the “Add” option. “Add all” can be used to add all the controls which match the filtered results to the dataset. The selected controls will appear as a dataset on the right side of the screen.

6.2.7. Select the blue “Next” button on the top right of the page to view and download the dataset. The data will display each individual gene found within the control experiments added.

6.2.8. Download the dataset by selecting “Download data matrix”. The data will be downloaded as an excel table.

6.2.9. Specific proteins can be elucidated from the dataset by typing in a gene name into the “Search” bar option. (**Figure 11**)

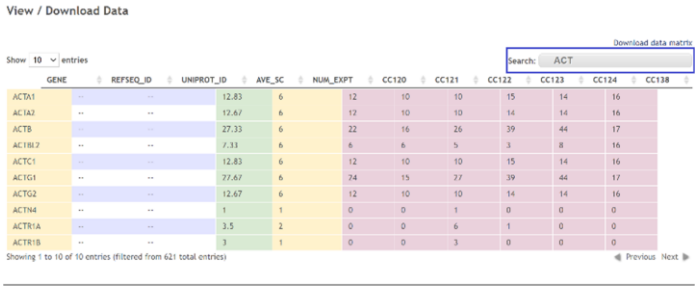


Figure 11

Figure 11: Genes containing “ACT” in the name are narrowed down from the dataset.

7. TROUBLE SHOOTING

7.1. The CRAPome site is under maintenance. The CRAPome database used to be available on the CRAPome website, www.crapome.org. If the CRAPome site remains under maintenance until further notice, it will provide an alternative link to use the database on a new site: REPRINT (reprint-apms.org).Use this link.

7.2. Protein of interest not detected in any negative control experiments (spectral count of 0.) This indicates that the protein is likely not a common contaminant in purification assays. You most likely don’t have to worry about this protein contaminating a sample. Choose a different protein (i.e. housekeeping genes or cytoskeletal elements) to continue the SOP with.

8. REFERENCES

Griffiths, J. (2008, July 31). A brief history of mass spectrometry. Retrieved from https://pubs.acs.org/doi/10.1021/ac8013065

Mellacheruvu, D., Wright, Z., Couzens, A. L., Lambert, J., St-Denis, N. A., Li, T., . . . Nesvizhskii, A. I. (2013, July 7). *The crapome: A contaminant repository for affinity*

Purification-mass spectrometry data. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3773500/.

The CRAPome: a Contaminant Repository for Affinity Purification Mass Spectrometry Data. Retrieved from https://reprint-apms.org/?q=chooseworkflow.

What is mass spectrometry? (2016, June 24). Retrieved from https://www.broadinstitute.org/proteomics/what-mass-spectrometry

Yury O. Tsybin, ... Anton N. Kozhinov, *Mass Spectrum.* (2019). Retrieved from https://www.sciencedirect.com/topics/chemistry/mass-spectrum

9. MODULE MASTERY TASK

Use the CRAPome database to find information about specific proteins and whether or not they are likely to function as a contaminant in biological purification assays and extract relevant data from negative control mass spectroscopy experiments.

9.1. Using Workflow 1

9.1.1. Submit 4 proteins of interest for analysis in workflow one using their UniProt ID or another accepted format. Rank them in order of what protein is most likely to be involved in contamination and non specific binding based on the CRAPome data to which protein is the least likely. What does this information indicate about these proteins? Can you draw any connection between these proteins structure/function to their contamination affinity?

9.1.2. Choose one of these proteins to investigate further. Go through each different affinity attribute and identify under what condition these proteins are most likely to contaminate/cause nonspecific binding. What does this tell you you should do if trying to run a purification assay and avoid this protein contaminating the assay?

9.1.3. How many individual experiments are available where this protein was present?

9.1.4. Choose 3 individual experiments. What does the spread column for each indicate?

9.2. Using Workflow 2

9.2.1. Select a dataset, use at least 4 filters to narrow the search results, and add at least five experiments to include in your results data matrix. Download the data matrix and submit the excel file. Submit with the excel data file a short summary of what filters you used to obtain the data and thus what this data is showing in terms of negative controls.