# General info

**See** [**Broad Institute proteomics tutorials**](https://www.broadinstitute.org/scientific-community/science/platforms/proteomics/tutorials-and-workshops)**, BWS protein interaction methods.docx, BWS gel and blot protocol.docx**

## Experimental design

* **“Biological replicates are almost always necessary, but technical replicates are not.”** ([December 2012 Broad Proteomics workshop: Everything you always wanted to know but were afraid to ask](https://www.broadinstitute.org/partnerships/education/broade/proteomics-everything-you-always-wanted-know-were-afraid-ask), 06 20121206\_Proteomics\_Wrokshop\_Day2\_GlobalProteomics\_Namrata.pdf)
* **Controls are very important.**
  + wild-type vs. knockout
  + Virus encoding tagged form of protein vs control virus
  + IP for protein of interest vs IgG or other protein (useful if knockout is not available).
  + IP for tag (FLAG, GFP, etc) in samples with and without the tag.
    - HA tag is not typically recommended[1](#_ENREF_1). I have not gotten clean results.
    - FLAG or 3xFLAG has high affinity and is difficult to elute from beads. This is fine with on-bead digestion.

## Sample prep

* Protein complex isolation
  + Non-denaturing buffers like RIPA and ME will not disrupt protein complexes, but denaturing urea or SDS-based buffers can. See [Broad proteomics lectures](https://www.broadinstitute.org/partnerships/education/broade/proteomics-everything-you-always-wanted-know-were-afraid-ask).
  + Cross-linkers can be used to keep protein complexes together. The Sabatini lab uses DSP (Di(N-succinimidyl) 3,3′-dithiodipropionate), which is a homobifunctional crosslinker (has identical reactive groups). DSP is reversible after affinity purification. Note that the tandem mass tags used for quantification react with primary amine groups. Most cross-linkers also react with amine groups. Even if reversible, this could affect the ability of the peptide to bind a TMT for quantification. We could try a carboxy-reactive cross linker.
* Enzyme digestion
  + **On-bead digestion is preferred because it eliminates elution and gels.**

# Taplin Mass Spectrometry Facility

The [Taplin Mass Spectrometry Facility](http://taplin.med.harvard.edu/) (TMSF), run by Ross Tomaino, analyzes gel bands and spots. See publications for more info on methods and results[2-4](#_ENREF_2). Mass spectrometry analyses in the Sha and Goldberg paper were also performed at the TMSF.

## Sample prep for TMSF

* **Sample submission**
  + Run an SDS-PAGE gel and stain with colloidal blue. See BWS gel and blot protocol for full details. Colloidal blue stained bands should be clearly visible to have enough protein for mass spectrometry.
  + Label 1.5 mL tubes with lane and band number on top, initials and date on side.
  + On a white surface like the white tray from the ProteinSimple imager, cut bands out of gel with a razor blade and add to 1.5 mL tubes. A light box is helpful.
  + Gel bands can be kept in di-H2O at 4°C for 1 week or frozen at -20°C without di-H2O.
  + Contact Ross, fill out submission form from website and bring gel bands to Ross in C-building 517B.
  + Turnaround time: 2 weeks

## TMSF methods

### Enzyme digestion

* Trypsin is the standard enzyme used for protein digestion. It cleaves peptide bonds at the C terminal side of RK.
* Chymotrypsin cleaves at C terminal side of WYF, less specific than trypsin.
* AspN cleaves specifically at N terminal side of D.
* Adding other enzymes may increase sequence coverage, but also increases the complexity of the sample. For interacting proteins, Ross recommends only using trypsin. Multi-enzyme digests are useful for evaluating post-translational modifications or protein cleavage products.

### Mass spectrometry

* The Taplin Center uses LC-MS/MS with electrospray ionization. The have an Orbitrap Elite and Pro.
* No mass tags are used, so only relative quantification of proteins within a run (gel band).
* Data are analyzed using the SEQUEST algorithm (developed by the Gygi lab).
* Protein sequences are used to query the FASTA database for identification.

## Data analysis

The GFY web search interface was created by the Gygi lab. A brief document is provided (“TMSFcore\_Info”).

**Browse data**

**Sessions**

* **Browse available sessions:** This shows the title of each run so you can tell which band is which. This information should also be provided in the “search report” email.

**Search data**

* **Browse search data**
  + Each listing corresponds to a run (gel band).
  + The **Saved Sets** column has a dropdown box that allows selection of report, report with contaminants, or none. Contaminants include keratins that have been filtered out of the dataset.
    - From TMSFcore\_Info document: “If you load the data set Report w/contaminant you will also see the FDR (false discovery rate) for your data set. In the example given there is a 1.31% FDR at the peptide level. This means that there are 26 peptides that were matched to the reversed database out of the 3958 total peptides that were matched with the XCorr and ∆Corr values we used. 26/3958 X 100 = 0.6568%. This means that in this data set that approximately 0.6568% (26 peptides) of the forward peptides are incorrect. The reverse peptides are also included in the data set so the FDR is 0.6568% X 2 = 1.31%. The 26 reverse peptides match to 23 different proteins therefore the protein FDR is 6.73% (23 reverse proteins/683 total proteins X 100 = 3.367%. Again, since the reverse proteins are listed the FDR is twice this value (2 X 3.367% = 6.73%). This means that out of the 683 proteins listed there 23 reverse proteins that are incorrect matches and potentially 23 more forward matches that are incorrectly listed in the data set. **Reverse proteins are always listed with two hash tags (##).** “
    - “If you load the data with the None saved set this will bring in all the tryptic peptides (in almost all cases trypsin will be the enzyme used to digest samples) regardless of the database search scores. If you do this you will bring in a lot of incorrect peptides matches. You can see in this example that the peptide and protein FDR went from 1.31% and 6.73% respectively to 39.12% and 79.35% respectively. Again, be aware the contaminants and carry over proteins will also be brought back into the data if you load the data as None. Feel free to input your own XCorr and ∆Corr values and reload the pages.”
  + Click on the search ID to bring up the search details screen. **Save as PDF.**
  + Check box, then click export to save SEQUEST parameters. Probably not important for the end user.
  + **Check the box next to the sample of interest, then click view**
    - **Columns** allows you to select the data that are shown for each peptide and protein.
      * **Select “Start Position” and “End Position”** to display the location of each peptide in the protein.
      * **Under Protein Table, select “Coverage %”** to display protein sequence coverage. For purified preparations, 70-80% coverage is common. Good coverage does not necessarily indicate high abundance.
      * **Select “protein sum max” and “protein sum max %”** to display chromatography peak intensity, which approximates protein abundance.
      * **Under Annotation, select Protein MWT (kDa) and Annotation.**
    - Click **load peptides**.
      * The Max column in the peptide list shows the signal intensity for each peptide.
      * The # Ions column shows the observed and predicted ions. **Click on the number to view the mass spectrogram.**
      * The correlation score describes the match of observed and predicted ions. Not as important for the end user. The “XCorr” gives the average correlation score for all peptides that match the protein. A ΔCorr >0.1 is considered a significant match.
      * Each peptide listed is an independent scan. The same peptide can come up multiple times. It could recur within the protein, and this is easily determined by searching against the full peptide sequence. Peptide recurrence can also be due to chromatography. Peptides can sometimes be retained and resequenced. Assignment of peptides that repeat within a protein is a common problem[5](#_ENREF_5).
    - Click **load proteins**. The browser displays a list of proteins in the sample, with the number of “unique” and “total” peptides that map to each protein sequence. Click on a protein for more info. Multiple samples can be viewed simultaneously. “One Hit Wonders” are proteins with one peptide match.
    - **Export:** **export csv files of proteins and peptides.** For clarity, rename files with date sample was run, Taplin search name, gel lane and band, and description.
      * **Sort protein csv results by “intensity%” (chromatography peak intensity), highest to lowest.**
    - **Post-translational modifications:** If modifications are detected, there will be a separate report for peptides on which modifications were detected. This is indicated in “Saved Sets” on the Browse Search Data screen.
      * The start and end of the peptide are marked with a period. Typically, the amino acid adjacent to the start and end will also be shown.
      * **Phosphorylation is denoted with # to the right of the residue.**
      * **Acetylation is denoted with @ to the right of the residue.**
      * Glycosylation and ubiquitination: It is true that PTMs induce a mass shift that can be detected with MS. However, glycosylation can be challenging because there are many different glycosylation PTMs with different masses, unlike phosphorylation or ubiquitination, which are always the same. You have to know the mass of the modification you are looking for. Ross can mainly identify monomer sugars, such as GlcNAc. Longer glycosylations are less stable.
  + Other sections
    - Select LDA Scores
    - **Plot:**
      * I haven’t found this feature to be useful. It displays histograms and scatter plots.
      * I haven’t found a way to display the LC chromatograms. I asked Ross about this. It’s not really this simple, because the proteins are fragmented into peptides, and it would be difficult to determine which peptides correspond to which protein based on the chromatogram. See the chromatogram in 031616\_Report\_tc-d278g\_tc-d279\_Smith\_B.pptx for an example.
    - Peptide Filters: allows you to filter peptide results list.
    - Protein Filters: allows you to filter protein results list.
    - ProteinSieve
    - Protein Assembler
    - SQL Queries

# Thermo Fisher Center for Multiplexed Proteomics

The [Thermo Fisher Center for Multiplexed Proteomics](http://tcmp.hms.harvard.edu/), run by Ryan Kunz, analyzes protein lysates and eluates. They primarily perform whole proteome profiling, and also do affinity purification-mass spectrometry for analysis of interacting proteins. **Others in the lab, including, Alex Bartelt and Ana Paula Arruda have performed proteomic analyses with the TCMP. Ling Yang and Takahisa Nakamura**[**6**](#_ENREF_6) **worked with Ryan in an academic collaboration.**

## Sample prep

* For AP-MS interaction proteomics, samples are submitted in elution buffer after elution from the affinity matrix. It is not necessary to cut out gel bands, though a silver stained gel is recommended to assess protein yield.
* When preparing samples for submission to the Thermo Fisher Center for Multiplexed Proteomics (TCMP), run 10% of the eluate on a gel. The gel run can be as little as 15 minutes. Do a silver stain and/or colloidal blue stain (See BWS gel and blot protocol for full details).
  + **Gel bands must be even and clearly visible with colloidal blue staining. Silver stained bands should show up within the recommended 2-3 minutes.** If a Western blot is needed to see the protein, it is not enough.
* Then give Ryan Kunz (TCMP director) 80-90% of the eluate. Reserve some sample for blots if desired. See guidelines below, and gel and blot protocol for information on gel staining.
* Fill out the sample submission form online, notify Ryan, and bring samples to C-building room 528.

### Information and guidelines from [website](http://tcmp.hms.harvard.edu/analytical-services)

***General Information***

* The Center provides protein profiling analytical services to the HMS community and to external users.
  + If you are a member of the HMS community, please fill out the sample submission form, answer the questions listed, and provide a Harvard 33-digit billing number.
  + If you are an external user and would like to purchase these analytical services from the Center, please fill out the sample submission form and provide a purchase order number (include the amount of money on the P.O. number) or a credit card or instructions for bank transfer. Please provide a billing address for sending the invoice. Samples will not be analyzed unless your P.O. or credit card information is received. All P.O. numbers should be made payable to Harvard University.
* The more information we have about the nature of the samples, the better we will be able to serve you.
* **Turnaround times** vary between four to five weeks for proteome analysis (three to four weeks for quantitative IPs) depending on our backlog.

***Guidelines***

* 1 - Guidelines for protein profiling from whole cell lysates: Users should submit cell pellets or tissues samples on dry ice. Each pellet/tissue should contain 100-300 ug of protein.
* 2 - Guidelines for label checks: Same as protein profiling from whole cell lysates (see above)
* 3 - Guidelines for quantitative IP analysis: For the best multiplex IP results we recommend the following:
  + Use crosslinked antibodies
  + Have a control IP
  + For each condition we recommend triplicate IPs
  + Bait proteins should have epitope purification tags, e.g. Flag, HA, etc
  + Take 10% of elution add it to SDS loading buffer and run a PAGE gel. Silver-stain the gel and send it to TCMP @ HMS before submitting samples. The remaining 90% we will TCA precipitate and then digest.

## TCMP

### Gel

* The TCMP does a brief gel cleanup before mass spectrometry. Ryan claims this is helpful because the gel is agnostic to the elution buffer used to obtain the sample. However, a high protein concentration is required because of the small gel loading volume.
* After mass tagging, he then also does a 3 hour column separation prior to MS.
* **If samples are divided among multiple runs, include an internal “mix” standard for comparison.** This is basically a small amount (5 μL) of all the samples mixed together.
* 30 μL of each sample is loaded into a 10% Bis-Tris gel and run at 120V for 12 minutes.
* Gels are stained for 2 hours with Coomassie and destained overnight in water.
* Additional gels are run and stained with the remaining sample.
* Gel bands are cut out, destained, reduced and alkylated.

### Enzyme digestion

In-gel trypsin digestion is performed.

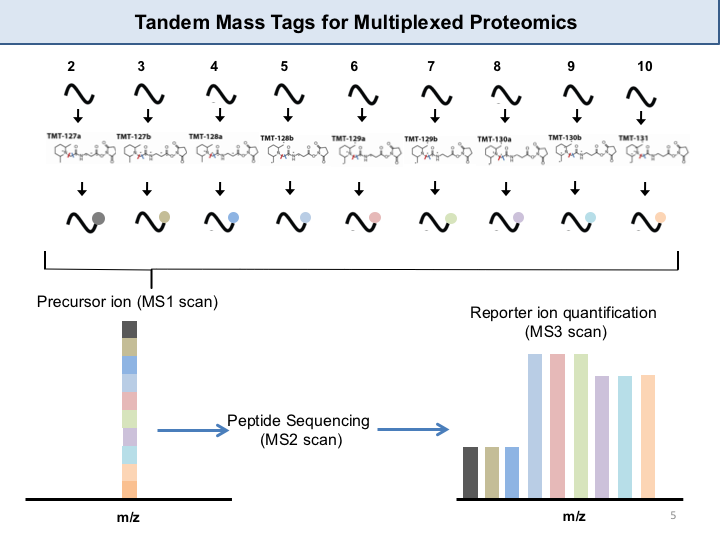
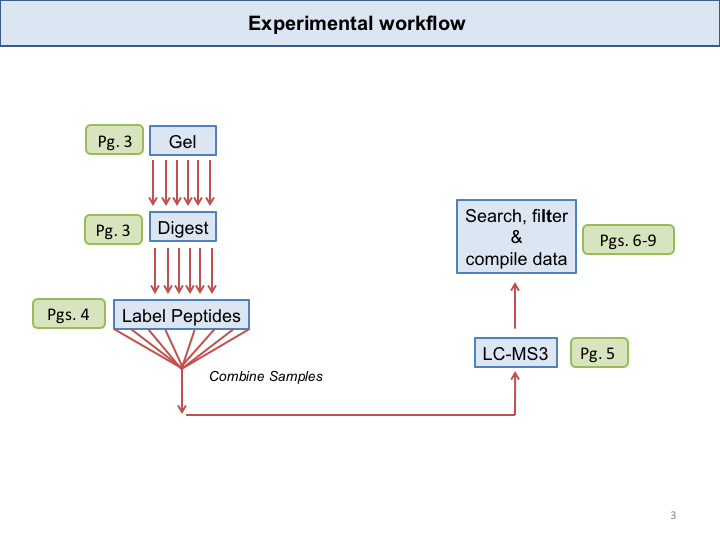
**Note that trypsin digestion can also be performed “on bead” (prior to elution from magnetic beads)**[**7**](#_ENREF_7)**,**[**8**](#_ENREF_8)**. On-bead digestion is preferred because it eliminates elution and gels.**

### Tandem Mass Tagging

[Tandem Mass Tags](https://www.thermofisher.com/us/en/home/life-science/protein-biology/protein-mass-spectrometry-analysis/protein-quantitation-mass-spectrometry/tandem-mass-tag-systems.html) (TMTs) are used to label primary amine groups. Ten different tags are available, allowing ten samples in the same mass spectrometry run. The tags are isobaric, meaning that they elute at the same time during LC, and have the same mass during MS1 acquisition, but after MS2 peptide sequencing, they fragment into unique ion masses during MS3 reporter ion quantification.

### Mass spectrometry

* Peptides are resuspended in 5% Acetonitrile, 5% formic acid
* Peptides were separated using a gradient of 6 to 28% acetonitrile in 0.125% formic acid over 180 minutes.
* Half of the sample is shot on an [Orbitrap fusion tribrid](http://planetorbitrap.com/orbitrap-fusion) mass spectrometer
  + MS3: MS1 (ion detection in orbitrap), MS2 (peptide sequencing in ion trap), MS3 (quantification in orbitrap)
  + The whole cycle takes 2 seconds in an orbitrap fusion tribrid.



### TCMP data analysis

* MS2 spectra are searched using the SEQUEST algorithm against a Uniprot composite database derived from the mouse proteome containing its reversed complement and known contaminants. **Peptide spectral matches are filtered to a 1% false discovery rate (FDR)** using the target-decoy strategy[9](#_ENREF_9), for determination of incorrectly identified proteins, combined with linear discriminant analysis.
  + From GraphPad “[What it means to control the FDR](http://www.graphpad.com/guides/prism/7/statistics/index.htm?stat_approach_3_false_discovery_rat.htm)”: FDR is basically Type I error rate.
* Proteins are quantified only from peptides with a summed SN threshold of ≥200 and MS2 isolation specificity of 0.5, and do not include contaminants or reverse hits.
* The TCMP provides an Excel workbook containing peptide counts (not sure if they are unique or total peptides), absolute and relative abundances of all proteins identified in the samples, and a PowerPoint report with methods and preliminary data analysis (hierarchical clustering performed in GENE-E, see below). They typically do not provide further assistance with data analysis.
* Protein Quant data are provided in two forms, **total summed signal to noise** and **normalized relative abundance**.
  + **Total summed signal to noise ratio (SNR)**: signal intensity.
  + **Normalized relative abundance (NRA):** normalization factors are calculated based on the total summed signal to noise in each TMT channel. Normalization is performed within each run/gel, but not between runs. **Note that this value assumes the same amount of protein in all samples, which might not be appropriate when considering the lacZ controls, which do not have Nrf1 and would therefore be expected to have different amounts of protein.** Total summed signal to noise may be more appropriate.

### Notes

* Does fee-for-service work in the TCMP or Taplin center ever lead to a more involved collaboration with the lab? -> Ryan: Rarely. Usually it happens the other way around: if a group is collaborating with the Gygi lab, they may realize they can get their work done more quickly through one of the two centers. For collaborations, usually the interested PI will talk directly to Steve Gygi. Steve will then say yes or no, and assign the project to one of his postdocs. The project is usually not high priority for them, and so it progresses very slowly. They are approached with many collaborations. Often the work ends up not being published at all, or if it is, the Gygi postdoc only gets middle author, which is not helpful to them.

# Farese Walther lab

## Sample prep

Perfuse mouse with PBS

Beads:

Does tween affect? Wash with IP lysis buffer, then detergent free wash at end

Freeze?

## Data analysis

They use Andromeda instead of sequest. In his old lab he also used trans proteomics pipeline

Ttest on untransformed values

R lima package benjamini transformations to control false discovery rates

>20% coverage adequate

ASK HIM ABOUT CHYMOTRYPSIN

# Data analysis

## Dataset management in Excel

**The major elements of data management in Excel are normalization, background subtraction, filtering, and sorting.**

**Always double check the data after a function is performed in Excel.**

### Normalization

To compare results across multiple runs, divide total summed signal to noise or normalized relative abundance for each sample by the corresponding value for the mix standard within that run. **Note that this creates a ratio, and log2 transformation should be performed before further analysis.**

* **Copy** column headers containing sample names into a new set of columns to the right. Delete the mix columns.
* **Normalize** databy dividing the value for the first sample by the value for the mix standard in that run. Include a $ before the column name to keep the column constant. Example: =G5/$F5. Modify the formula if columns are added to the left. If a mix standard was not included, it may be possible to normalize to a common protein.
* **Fill** to the right for the rest of the samples in the run (Edit -> Fill right, or button in ribbon), then fill down to calculate for all proteins in the run. **Repeat for any additional runs in the worksheet.**
* **Add columns** to the right as desired, to calculate HA cholesterol/HA control ratio for example. **Integrate the new column headers with the rest of the dataset.** Select the row of column headers, then filter (Data -> Filter, or use the button in the ribbon). If filters are already in place, click filter twice. You should see a dropdown box in each column header. This will ensure that all information in each row of the worksheet moves together when filters are applied to the columns. **Rearrange columns as desired before filtering.**
* **Calculate treatment/control ratio.**

### Filtration

* **Filter the hit list to exclude proteins with ≤2 peptides quantified in all runs.** Including only proteins with ≥2 peptides means the identification of the protein itself is confident, because multiple peptides corresponding to it have been identified, and that the identification of the protein is repeatable, because ≥2 peptides were detected in each mass spectrometry run.
* Filtering can easily be accomplished in Excel by clicking the dropdown box in the bottom right of the column header (worksheets from TCMP are in pivot table format), and selecting filter by >1. **Note that copying and pasting ranges of cells becomes more complicated after filtering.** Some cells will be hidden. Selecting a range of visible cells can skip over hidden cells, and Excel will read this as making multiple selections. **If cells need to be copied and pasted after filtration, unfilter and then refilter after.**
* **Proteins can also be filtered by % protein coverage by the corresponding detected peptides.** There is no set standard for this, but coverage >20% is generally adequate according to the Farese/Walther lab.

### Transformation

* **log2 transformation is useful when testing ratios.** 
  + **“Genes upregulated by a factor of 2 have an expression ratio of 2, whereas those downregulated by the same factor have an expression ratio of (–0.5). The most widely used alternative transformation of the ratio is the logarithm base 2, which has the advantage of producing a continuous spectrum of values and treating up- and downregulated genes in a similar fashion.”**[**10**](#_ENREF_10)
  + [Cluster 3.0 manual 2.3.1](http://bonsai.hgc.jp/~mdehoon/software/cluster/manual/Data.html#Data): “**The results of many DNA microarray experiments are fluorescent ratios. Ratio measurements are most naturally processed in log space.** Consider an experiment where you are looking at gene expression over time, and the results are relative expression levels compared to time 0. Assume at timepoint 1, a gene is unchanged, at timepoint 2 it is up 2-fold and at timepoint three is down 2-fold relative to time 0. The raw ratio values are 1.0, 2.0 and 0.5. In most applications, you want to think of 2-fold up and 2-fold down as being the same magnitude of change, but in an opposite direction. In raw ratio space, however, the difference between timepoint 1 and 2 is +1.0, while between timepoint 1 and 3 is -0.5. Thus mathematical operations that use the difference between values would think that the 2-fold up change was twice as significant as the 2-fold down change. Usually, you do not want this. **In log space (we use log base 2 for simplicity) the data points become 0,1.0,-1.0. With these values, 2-fold up and 2-fold down are symmetric about 0. For most applications, we recommend you work in log space.**”
  + See Takahisa Nakamura[6](#_ENREF_6) and Alex Bartelt's datasets. Also see Salk HOMER tutorial page. Alex was performing statistical tests on fold change (KO/WT and Cold/thermoneutral), so log2 was appropriate.
  + Log2 can also be useful in SILAC experiments (heavy/light)[11](#_ENREF_11).
  + Normalization to the mix standard creates a ratio.

### Background subtraction

* The experimental design should include control groups, such as knockout mice. If infecting with adenovirus encoding a tagged protein (such as Nrf1-HA) for isolation, include a lacZ control virus group. IgG IP control can also be used.
* Background proteins could result from nonspecific binding to agarose resin or magnetic beads. Ribosomal pathways are frequently enriched because ribosome proteins have strong affinity for magnetic beads. They are part of the “bead-ome[7](#_ENREF_7).” See Kazak *Cell* 2015[12](#_ENREF_12) for example of a proteomic analysis from the Gygi lab with abundant ribosome proteins. They did not enrich with magnetic beads in this analysis, but ribosome proteins may also be generally abundant in the cell. A subtraction method is needed to determine which proteins are nonspecific and which are true interactors.
* Background subtraction with Tandem Mass Tags: A TMT signal of 0 across all the channels indicates that there were no peptides corresponding to the protein detected in the run. Sometimes, a few peptides will be detected in one run, and none in another run. When peptides corresponding to a protein are detected, there will be signal across all the TMT channels. There is always a small amount of background signal, but samples in which the protein is present at higher relative abundance will have substantially stronger signal. I’m not aware of an established cutoff for background signal.

## Statistical analysis

### Basic statistics

#### Excel

* **Fold change: A fold change of 1.5 (log2=0.58) is commonly used as a threshold for biological significance**[**11**](#_ENREF_11)**,**[**13**](#_ENREF_13)**.**
* T-tests can be performed in Excel: =TTEST(group1,group2,tails,type).
* Sort ascending to see significant hits.
* Highlight significant results with conditional formatting. Format -> conditional formatting -> Style: classic, Format only cells that contain -> Cell value -> less than or equal to -> =0.05, Format with: green fill with dark green text.

### Hierarchical clustering and heat maps

* Hierarchical clustering, as the name implies, is a method of grouping data into a hierarchy of clusters. Samples are grouped based on their protein abundance profiles. Animals in the same treatment group would be expected to cluster together.
* The “one minus Pearson correlation method” correlates samples based on their protein expression profiles.
* Other examples of clustering:
  + Cao *Cell* 2008[14](#_ENREF_14)

#### Data input

* **Paste log2 transformed data into a new Excel worksheet, using “paste values”** to omit formulas and rows that have been filtered out (such as proteins with ≤2 peptides). Retain just the sample names as column headers, gene names in left column, and log2 data.
* **Make sure the data are sorted the same way each time, so that the heat maps correspond.**

#### GENE-E

* [Broad Institute](http://www.broadinstitute.org/cancer/software/GENE-E/). Requires Java. See GENE-E documentation for basics. This is my preferred program.
* Import data
  + Open the Excel file with File -> open file or Cmd+O.
  + Select the appropriate cell to delineate row and column titles.
  + The dataset can also be transposed to have each sample be a row instead of a column. However, note that in this setup, the hierarchical clustering may not display on the same side as the data labels (sample names).
* Heat map:
  + Remove the text of the protein names by right clicking the column header and deselecting Display->Text.
  + View -> Preferences -> Color scheme. Relative compares samples among columns within a given row (protein), global calculates a single color intensity scale based on all values in the dataset.
  + **Fit heat map to window** (button at top of screen). Click again to change back.
* Hierarchical clustering dendrogram:
  + Use “one minus Pearson correlation,” do not group columns.
  + The program will display a cluster tree, also called a dendrogram, above the heat map.
  + Drag the dashed line to modify the cluster threshold.
* Other functions
  + Data can be output to R.
  + The dataset can be filtered by search terms. Enter a term in either the columns or rows box. Click “select all matches” to select search results.
  + Row annotations provide advanced filtering capabilities. Double click on a function to add it to the formula builder. “MAD” is median absolute deviation, a measure of variation in the dataset.

#### Cluster 3.0

* Access [here](http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm#ctv). **Highly informative instruction manual** [**here**](http://bonsai.hgc.jp/~mdehoon/software/cluster/manual/index.html)**.** Also see [Salk HOMER clustering tutorial](http://homer.salk.edu/homer/basicTutorial/clustering.html).
* Based on the same original program as TreeView3, developed by Michael Eisen at Stanford[15](#_ENREF_15).
* Not updated as frequently as TreeView.
* Import data:
  + Export each worksheet from Excel workbook as tab delimited text (.txt)
  + Import into Cluster.
  + when trying to import .xlsx or .csv, you will get the error message “Error in data file. Less than two columns found in data file.”
* On the Hierarchical tab, clustering by gene will cluster by rows (gene/protein name), and clustering by array will cluster by column (sample name).
* Calculations performed in Cluster will be exported as corresponding .cdt and .gtr files. These can be imported into TreeView for visualization.

#### TreeView3

* Access [here](https://www.princeton.edu/~abarysh/treeview/) from the Baryshnikova lab at Princeton. Originally developed by Michael Eisen in David Botstein’s lab at Stanford[15](#_ENREF_15). No documentation currently available.
* Import data: export each worksheet from Excel workbook as a .csv file.
* Less features than GENE-E. Only has global color scheme for heat map. I couldn’t get clustering to work.

#### Perseus

* [Matthias Mann group Max Planck](http://www.coxdocs.org/doku.php?id=perseus:start) (Matthias Mann group, Max Planck)
* Perseus reads output from MaxQuant
* Import data
  + Export from Excel as tab-delimited text?

#### R

#### Prism

* Prism 7 has a new heat map feature. It colors globally and is not as flexible as GENE-E.

### Other

#### Perseus

* batch statistical significance
* See Meissner 2014[16](#_ENREF_16) Figure 4C (doesn’t provide methodological details)

## Pathway analysis

* See pathway analysis protocol

## Data analysis

* Data reporting standards: Minimum Information About a Proteomics Experiment ([MIAPE](http://www.psidev.info/node/91))[17](#_ENREF_17),[18](#_ENREF_18)
* Spectral counting is a technique that uses peptide intensity values to approximate protein abundances by mapping the peptides to the appropriate protein[19](#_ENREF_19),[20](#_ENREF_20). This is useful in label-free AP-MS analyses[21](#_ENREF_21), including Taka’s[6](#_ENREF_6). “Statistical methods hard to generate… not recommended for serious quantitative proteomics” ([December 2012 Broad Proteomics workshop: Everything you always wanted to know but were afraid to ask](https://www.broadinstitute.org/partnerships/education/broade/proteomics-everything-you-always-wanted-know-were-afraid-ask), 04 BroadE - Quantification in Proteomics.pdf).

# References

* [Broad Institute proteomics tutorials and workshops](https://www.broadinstitute.org/scientific-community/science/platforms/proteomics/tutorials-and-workshops)
* [Mass spectrometry-based proteomics at Nature Methods](http://blogs.nature.com/methagora/2014/09/mass-spectrometry-based-proteomics-at-nature-methods.html)
* BWS gel and blot protocol.docx
* BWS protein interaction methods.docx

1. Kwan JH, Emili A. Simple and Effective Affinity Purification Procedures for Mass Spectrometry-Based Identification of Protein-Protein Interactions in Cell Signaling Pathways. *Methods Mol. Biol.* 1394:181-187 (2016).

2. Yang T, Espenshade PJ, Wright ME, Yabe D, Gong Y, Aebersold R, Goldstein JL, Brown MS. Crucial step in cholesterol homeostasis: sterols promote binding of SCAP to INSIG-1, a membrane protein that facilitates retention of SREBPs in ER. *Cell* 110:489-500 (2002).

3. Fu S, Yang L, Li P, Hofmann O, Dicker L, Hide W, Lin X, Watkins SM, Ivanov AR, Hotamisligil GS. Aberrant lipid metabolism disrupts calcium homeostasis causing liver endoplasmic reticulum stress in obesity. *Nature* 473:528-531 (2011).

4. Sha Z, Goldberg AL. Proteasome-mediated processing of Nrf1 is essential for coordinate induction of all proteasome subunits and p97. *Curr. Biol.* 24:1573-1583 (2014).

5. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* 43:D512-520 (2015).

6. Nakamura T, Kunz RC, Zhang C, Kimura T, Yuan CL, Baccaro B, Namiki Y, Gygi SP, Hotamisligil GS. A critical role for PKR complexes with TRBP in immunometabolic regulation and eIF2alpha phosphorylation in obesity. *Cell Rep.* 11:295-307 (2015).

7. Keilhauer EC, Hein MY, Mann M. Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS). *Mol. Cell. Proteomics* 14:120-135 (2015).

8. Kronke J, Udeshi ND, Narla A, Grauman P, Hurst SN, McConkey M, Svinkina T, Heckl D, Comer E, Li X *et al.* Lenalidomide causes selective degradation of IKZF1 and IKZF3 in multiple myeloma cells. *Science* 343:301-305 (2014).

9. Elias JE, Gygi SP. Target-decoy search strategy for mass spectrometry-based proteomics. *Methods Mol. Biol.* 604:55-71 (2010).

10. Quackenbush J. Microarray data normalization and transformation. *Nat. Genet.* 32 Suppl:496-501 (2002).

11. Margolin AA, Ong SE, Schenone M, Gould R, Schreiber SL, Carr SA, Golub TR. Empirical Bayes analysis of quantitative proteomics experiments. *PloS one* 4:e7454 (2009).

12. Kazak L, Chouchani ET, Jedrychowski MP, Erickson BK, Shinoda K, Cohen P, Vetrivelan R, Lu GZ, Laznik-Bogoslavski D, Hasenfuss SC *et al.* A creatine-driven substrate cycle enhances energy expenditure and thermogenesis in beige fat. *Cell* 163:643-655 (2015).

13. Blagoev B, Ong SE, Kratchmarova I, Mann M. Temporal analysis of phosphotyrosine-dependent signaling networks by quantitative proteomics. *Nat. Biotechnol.* 22:1139-1145 (2004).

14. Cao H, Gerhold K, Mayers JR, Wiest MM, Watkins SM, Hotamisligil GS. Identification of a lipokine, a lipid hormone linking adipose tissue to systemic metabolism. *Cell* 134:933-944 (2008).

15. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95:14863-14868 (1998).

16. Meissner F, Mann M. Quantitative shotgun proteomics: considerations for a high-quality workflow in immunology. *Nat. Immunol.* 15:112-117 (2014).

17. Taylor CF, Paton NW, Lilley KS, Binz PA, Julian RK, Jr., Jones AR, Zhu W, Apweiler R, Aebersold R, Deutsch EW *et al.* The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* 25:887-893 (2007).

18. Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, Ball CA, Binz PA, Bogue M, Booth T *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* 26:889-896 (2008).

19. Liu H, Sadygov RG, Yates JR, 3rd. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 76:4193-4201 (2004).

20. Zybailov B, Coleman MK, Florens L, Washburn MP. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal. Chem.* 77:6218-6224 (2005).

21. Varjosalo M, Sacco R, Stukalov A, van Drogen A, Planyavsky M, Hauri S, Aebersold R, Bennett KL, Colinge J, Gstaiger M *et al.* Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nat. Methods* 10:307-314 (2013).