Corra v2.0 User's Guide

INSTITUTE FOR Systems Biology

Revolutionizing science. Enhancing life.

BMC Bioinformatics

BioMed Central

Methodology article

Open Access

Corra: Computational framework and tools for LC-MS discovery and targeted mass spectrometry-based proteomics

Mi-Youn Brusniak[1], Bernd Bodenmiller[2,3], David Campbell[1], Kelly Cooke[1], James Eddes[1], Andrew Garbutt[1], Hollis Lau[1], Simon Letarte[1], Lukas N Mueller[2,3], Vagisha Sharma[1], Olga Vitek[4], Ning Zhang[1], Ruedi Aebersold[1,2,3,5] and Julian D Watts*[1]

Address: [1]Institute for Systems Biology, 1441 North 34th Street, Seattle, WA 98103, USA, [2]Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland, [3]Competence Center for Systems Physiology and Metabolic Disease, ETH Zurich, Zurich, Switzerland, [4]Department of Statistics and Department of Computer Science, Purdue University, West Lafayette, IN, USA and [5]Faculty of Science, University of Zurich, Zurich, Switzerland

Email: Mi-Youn Brusniak - mbrusnia@systemsbiology.org; Bernd Bodenmiller - bodenmiller@imsb.biol.ethz.ch; David Campbell - dcampbel@systemsbiology.org; Kelly Cooke - kcooke@systemsbiology.org; James Eddes - eddes@imsb.biol.ethz.ch; Andrew Garbutt - andga@microsoft.com; Hollis Lau - hlau@amgen.com; Simon Letarte - sletarte@systemsbiology.org; Lukas N Mueller - Lukas.Mueller@imsb.biol.ethz.ch; Vagisha Sharma - vsharma@u.washington.edu; Olga Vitek - ovitek@stat.purdue.edu; Ning Zhang - nzhang@u.washington.edu; Ruedi Aebersold - rudolf.aebersold@imsb.biol.ethz.ch; Julian D Watts* - jwatts@systemsbiology.org

* Corresponding author

Published: 16 December 2008

BMC Bioinformatics 2008, 9:542 doi:10.1186/1471-2105-9-542

Received: 20 May 2008

Accepted: 16 December 2008

Corra is an open source software Licensed under the Apache License, Version 2.0 and it's source code , demo data and this guide can be downloaded at the http://tools.proteomecenter.org/Corra/corra.html.

This user guide is written by Micheleen Harris (mharris@systemsbiology.org) and Mi-Youn Brusniak (mbrusniak@systemsbiology.org)

June 10, 2010

1. Introduction

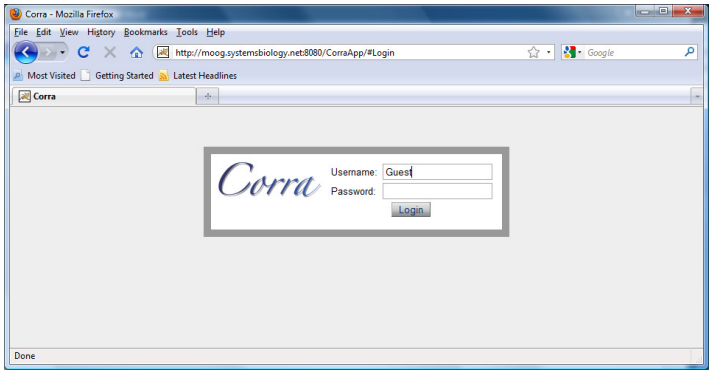
Corra is a single, user-friendly, informatic framework, that is simple to use and fully customizable, for the enabling of LC-MS-based quantitative proteomic workflows of any size, able to guide the user seamlessly from MS data generation, through data processing, visualization, and statistical analysis steps, to the identification of differentially abundant or expressed candidate features for prioritized targeted identification by subsequent MS/MS. In the first published version of Corra software with the paper was v 1.5 in 2008 and since then, there were more update in the pipeline. In detail, Corra v1.5 pipeline ended by generating target list from statistical analysis. Corra v2.0 added additional feature extracting alignment tools as well as customized target list generation and annotation step using target LS-MS run. This guide uses the yeast gene knock out example used in Corra paper to illustrate the step of using v2.0 extended pipeline steps.

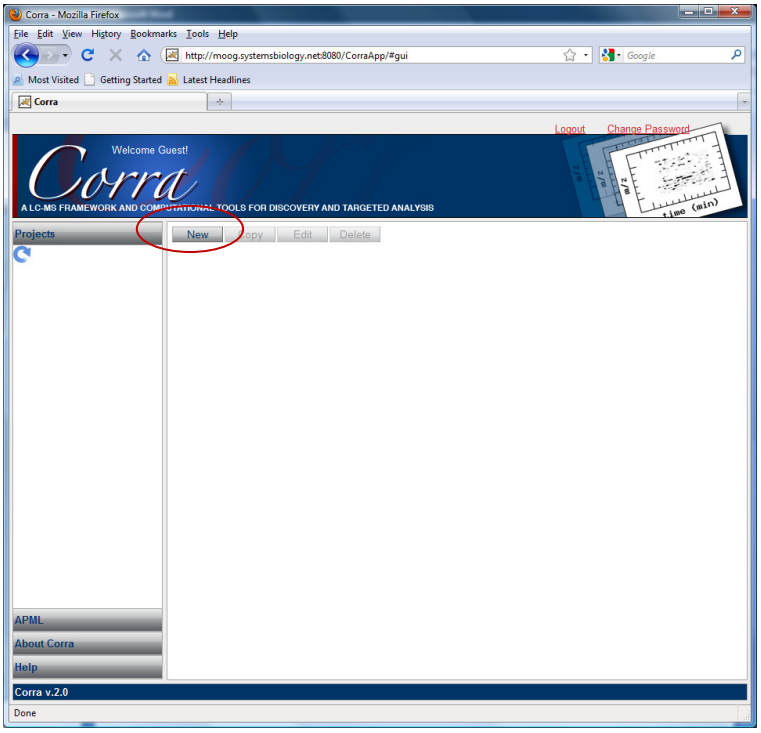
1.1 Login

Website: Ask administer in your institution which server the Corra is deployed to and ask Corra admin to add your account. For this guide, we will use guest account.

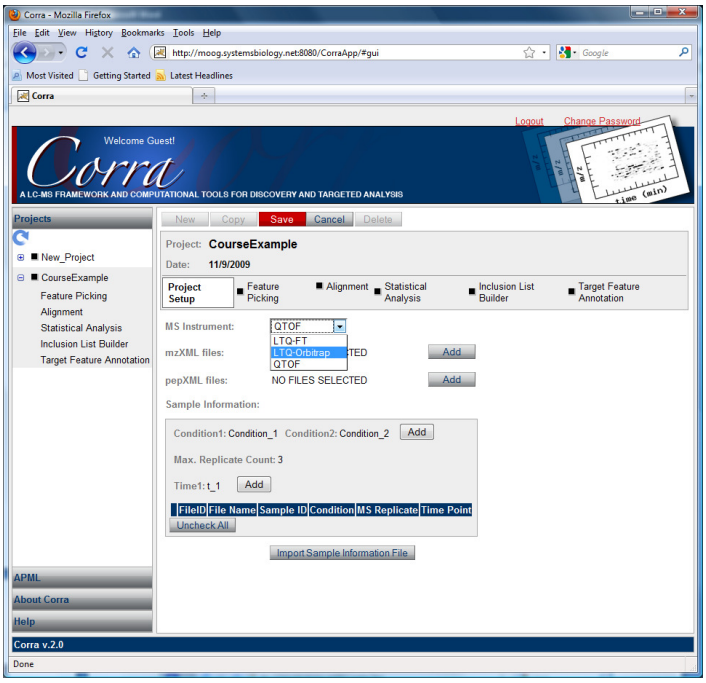
The URL should be something like the following.

http://corrademo.systemsbiology.net



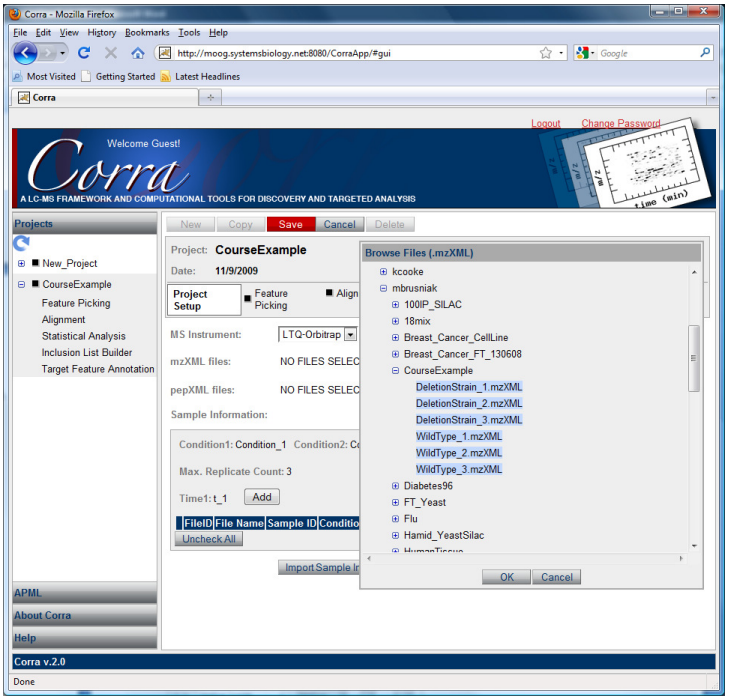


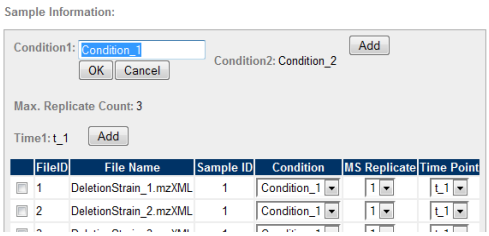
1.2 Once logged in, click “New” to create a new project and give it a name (here it is “CourseExample”).



1.3 Choose the instrument type under drop down menu “MS Instrument”

1.4 Adding Data. Your data must be in mzXML format (if not, there are several converters from RAW data to mzXML, such as ReAdW and mzWiff). Click “Add” next to mzXML files to add mzXML formatted data to the project (required before you save the project). Select the mzXML files from the drop-down menu with which you want to run Corra (you can hold down the Shift key to select a group of files). Then press “Save” and reopen Project Setup by pressing “Edit.”

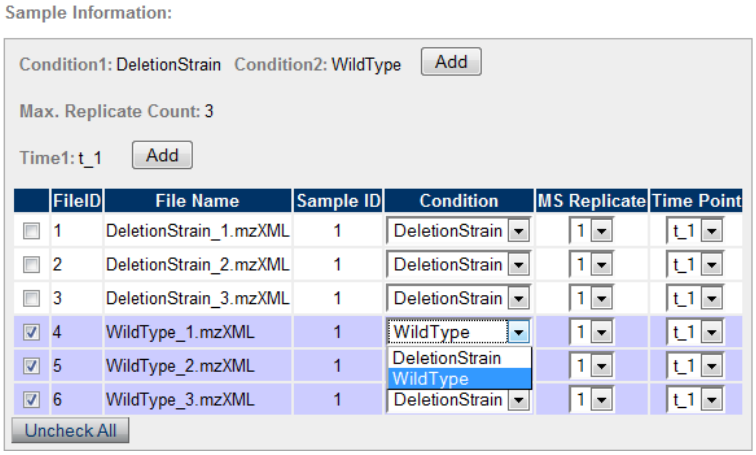


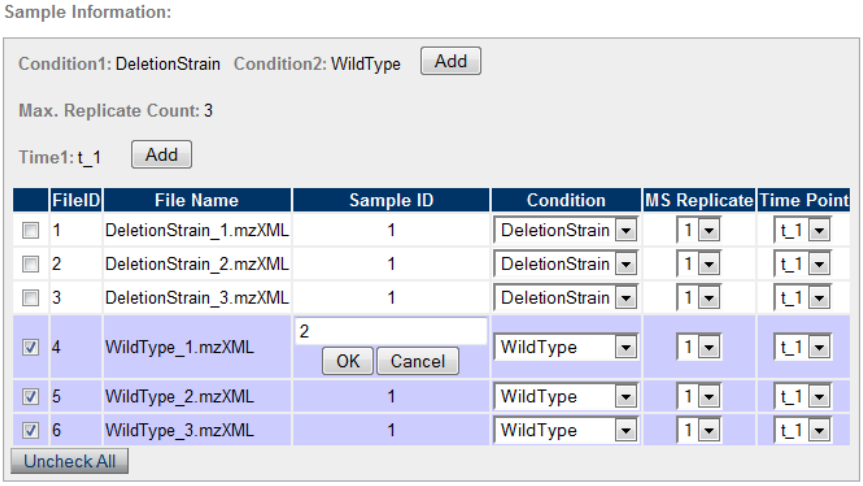


1.5 Defining Conditions, Sample IDs, Replicates and Time points. Make sure you have clicked "Edit" to continue setting up the project.

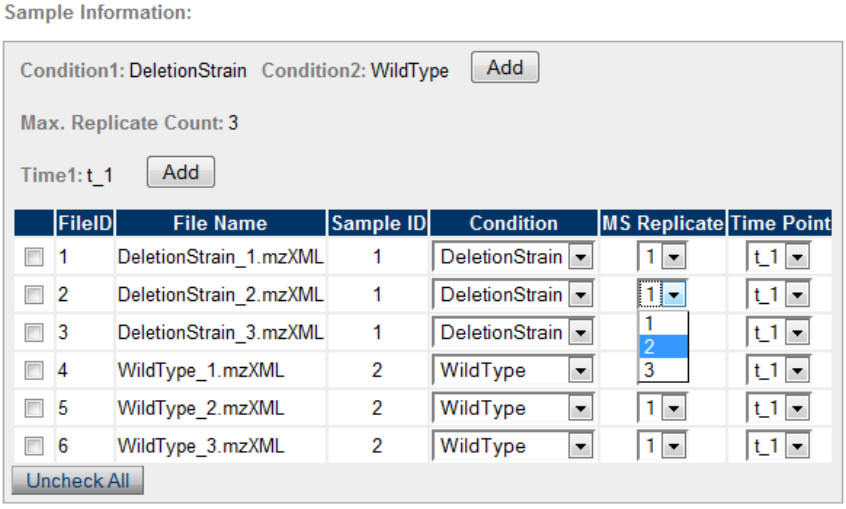
Click on "Condition_1" or "Condition_2" to rename these labels. If you wish to add any more conditions, click "Add."

Check the files to label and use the drop-down menu to select the condition label appropriate for this group of files.





Each group of files (e.g. replicates belonging to a particular biological group) should share the same "Sample ID." Assign a numerical ID by clicking on a number in the "Sample ID" column as shown left side.



Define replicates using the drop-down menu in the "MS Replicates" column. If you have more than 3 replicates increase the replicate count by clicking on the number next to "Max. Replicate Count."

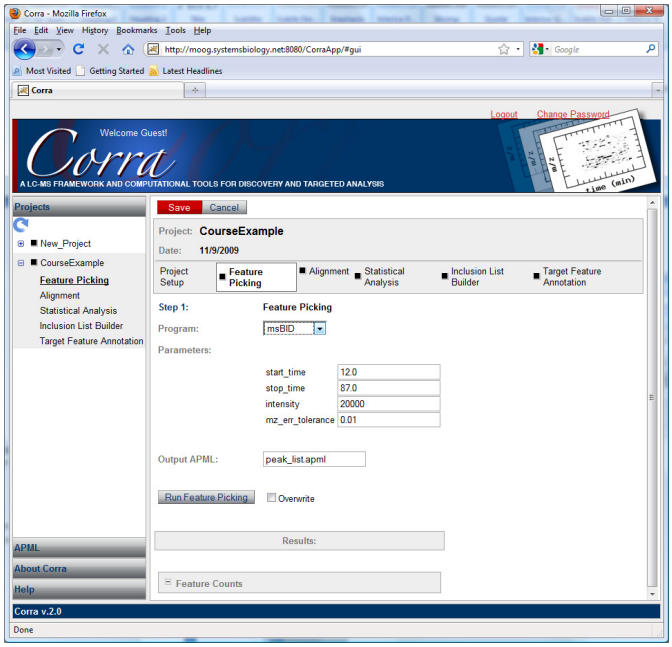
If you have more than one defined time point, add it by clicking the “Add” button next to “Time1” and rename by replacing “t_1” or “t_2” etc. Then specify them in the “Time Point” column using the drop-down menu.

Don’t forget to **Save your work**!

Alternatively, you could setup the project by importing a “Sample Information File.” This is useful if, say, you have a similar project with many mzXML files, as entering all of the setup information by hand could be a rather long process or you can use “Copy” project option which will create a new project with current project setup page. This “Copy” option can be used to analyze data using alternative Corra pipeline options.

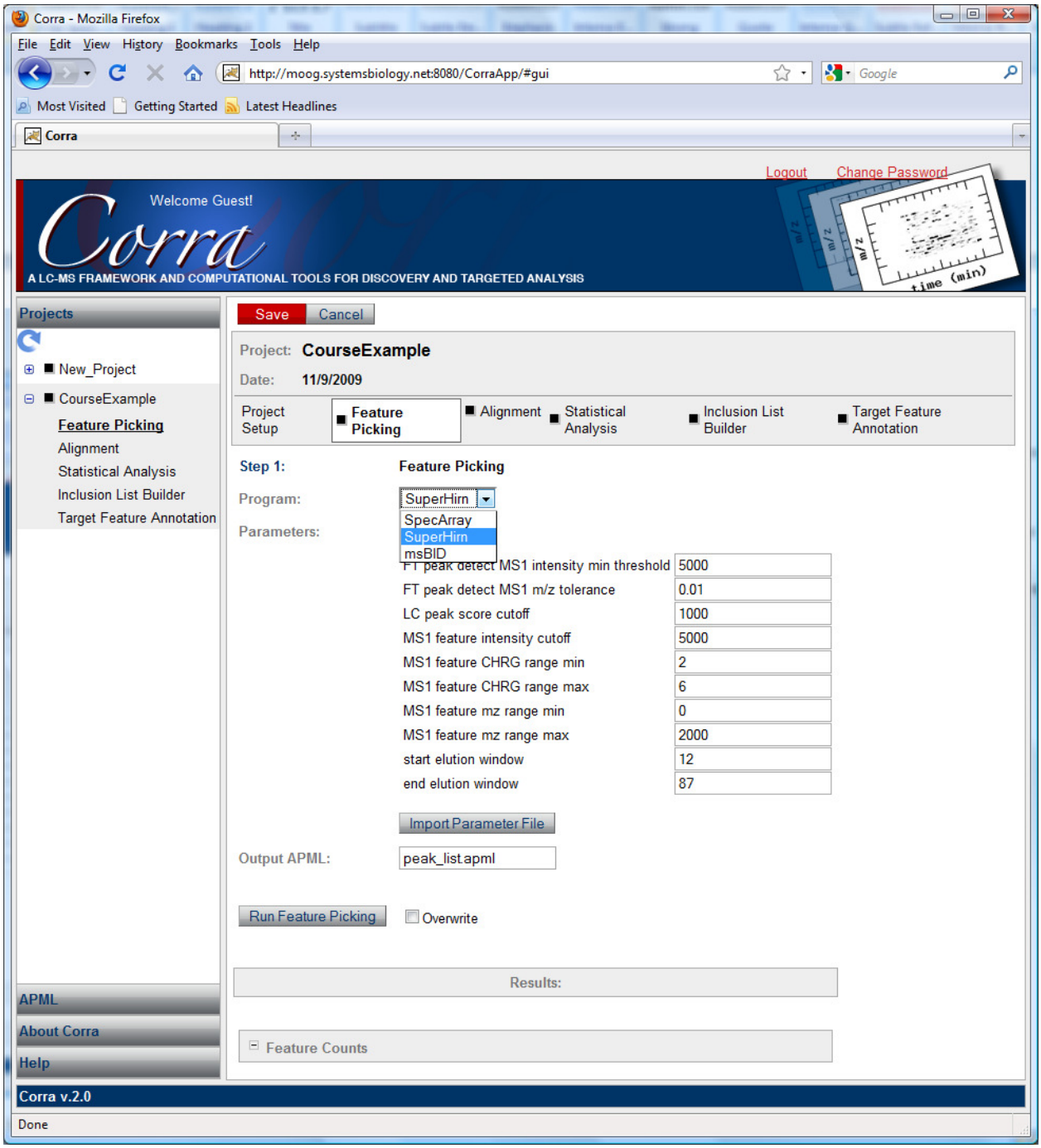
2. Feature Picking

2.1 Click on “Feature Picking”, then “Edit”



2.2 Program for feature picking

Select the desired program (e.g. SpecArray is used for TOF-MS data and SuperHirn /msBID for FT-MS/Orbitrap)



For the Feature Picking step it might be useful to view the mzxml file(s) data. Using Pep3D, a .png (image) file can be created and viewed in a generic graphic viewer. For example, using the SuperHirn program, the elution window is set by default to begin at 12 and end at 87 minutes. Viewing the mzxml file in a program like Pep3D can help you decide if you wish to exclude (or include) parts of the experiment based on how the elution profile looks (Pep3D is a viewer of LC-MS or LC-MS/MS data in a general 2D "gel-like" format).

2.3 Set “Parameters” or import a parameter file

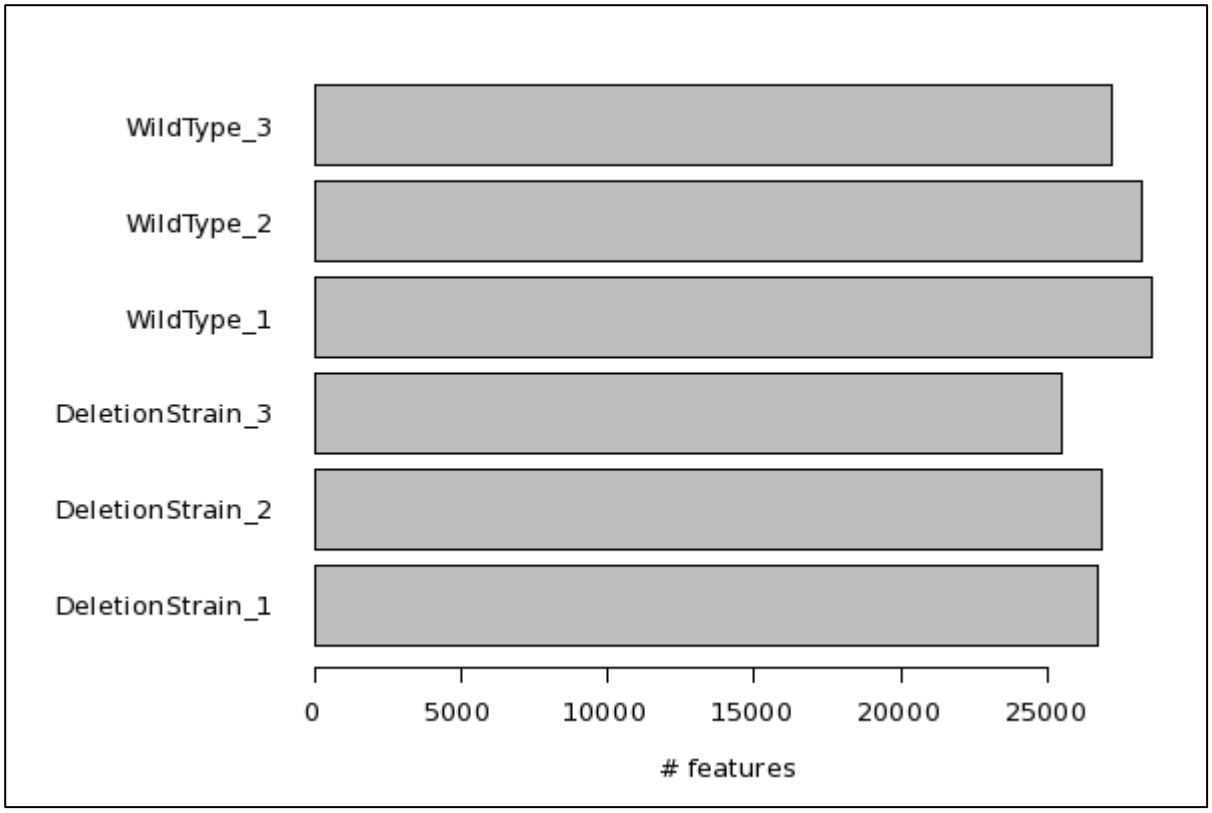
2.4 Click “Run Feature Picking”

> **Note**: Text in yellow indicates a process that is currently running and text in green is a process which has completed successfully. Text in red indicates an error has occurred and Corra log files may be referenced for further information.

When Feature Picking is done, you can scroll down to view the resulting feature counts for each input file.

> Note: The “FT peak detect MS1 intensity min threshold” could be increased in the case where there are too many features and/or you desire the subsequent runs to be faster (adjust the parameter and rerun Feature Picking).

Here is a picture of the result of the feature counting:

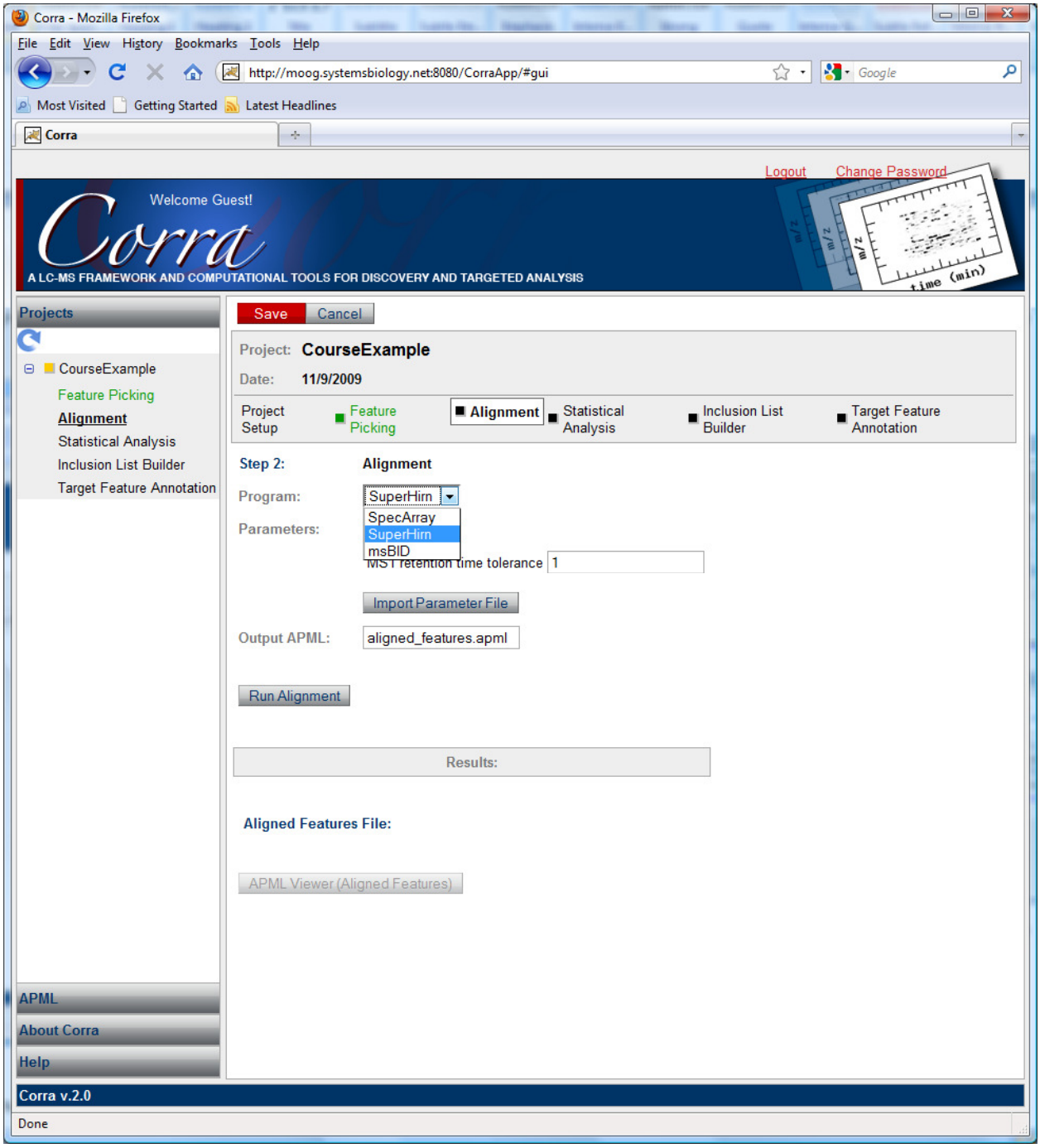


Note: These pictures can be downloaded as a .pdf file through the link below this graph.

3. Alignment

3.1 Click on the next step, “Alignment.” Click “Edit” to setup the Alignment parameters.

3.2 Select a program (this should correspond to the program selected during Feature Picking)

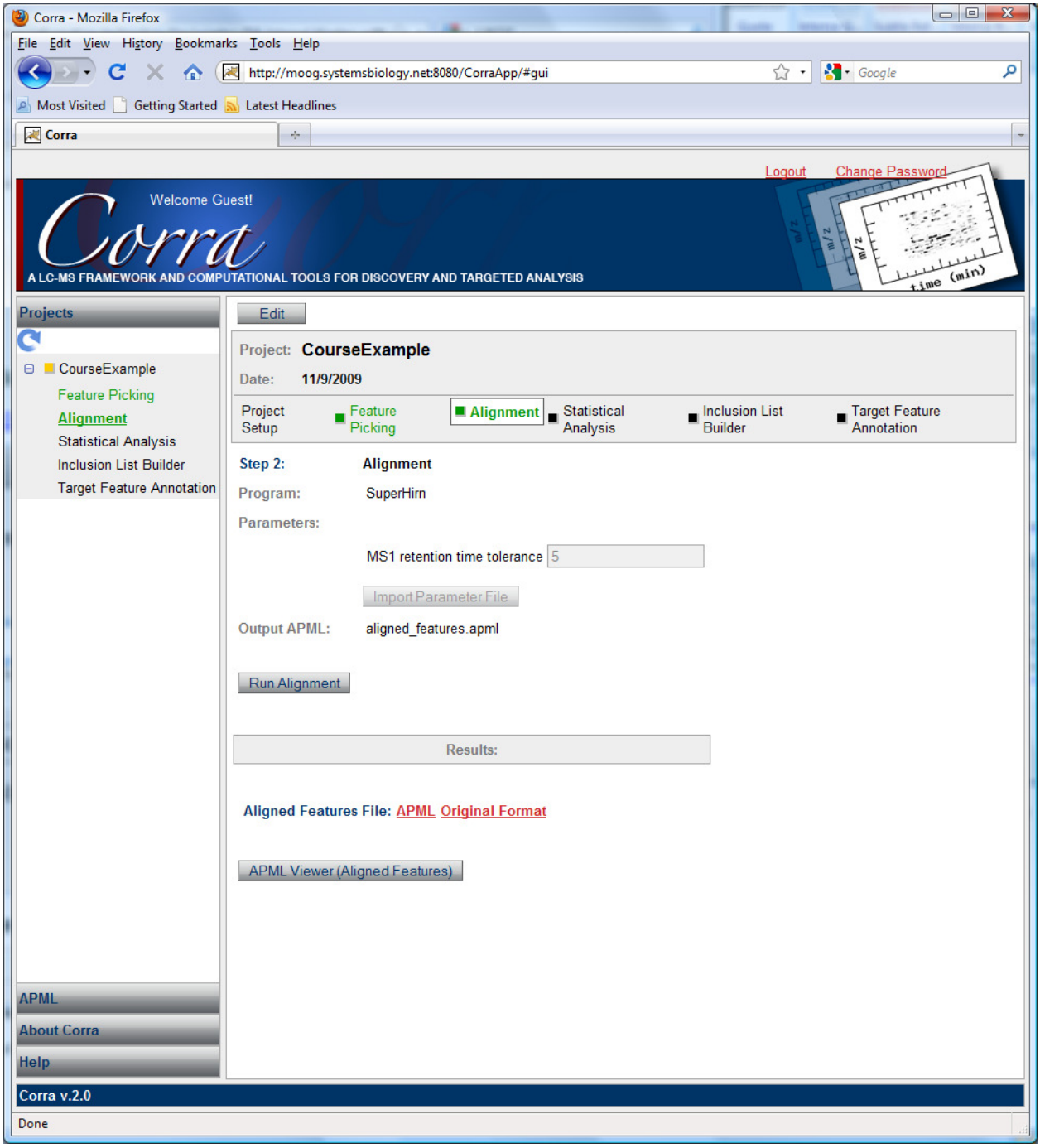


3.3 Parameters

Adjust parameters to meet the specifications of your analysis and then click “Run Alignment.”

Note: It might be a good idea to start with a value of 5 for the MS1 retention time tolerance.

3.4 Alignment Results

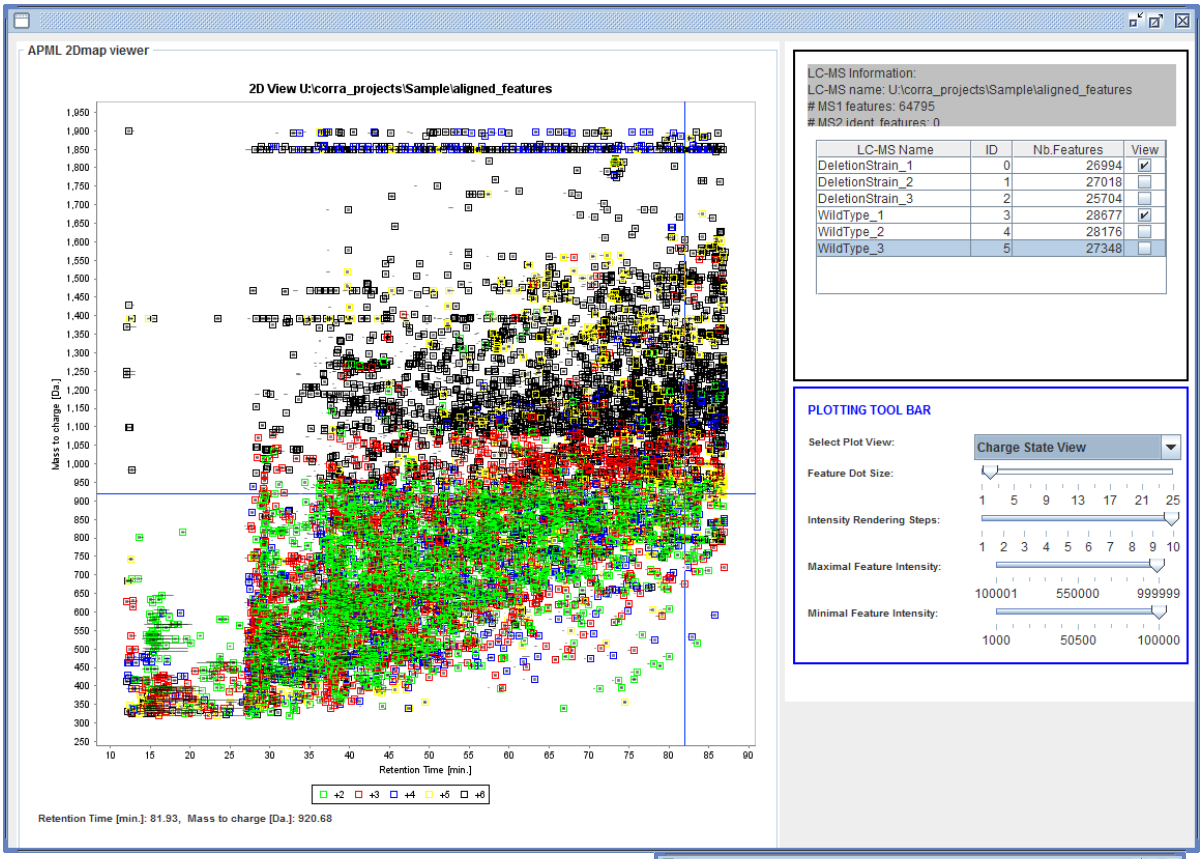


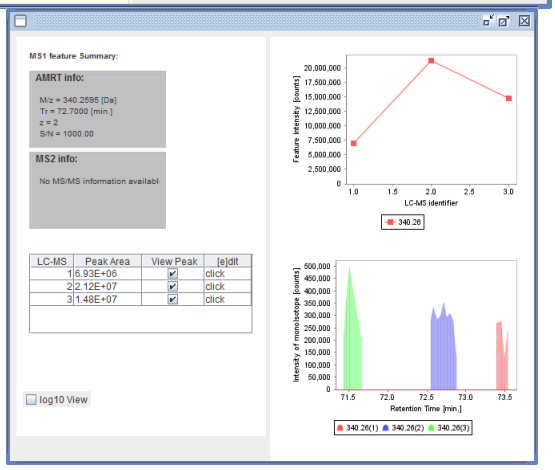
An APML (Annotated Putative Peptide Markup Language) file is created and maybe downloaded (by clicking the APML link in red) and viewed in an APML viewer comes with Corra (

http://sourceforge.net/projects/corra/files/Corra-APML/APMLv2.0.1/APMLv2.0.1.tgz/download). This will help the user to view in a graphical way, the amount of aligned features. See next section for details about the APML Viewer.

3.4.1 APML Viewer

Open an .apml file in the viewer to see the aligned features in a *m/z* vs. Tr plot.
Try this: In the "Plotting Tool Bar", go to "Selected Plot View" -> "Times Aligned View" and click on a point in the graph to get a dialog box which shows the aligned features for that point (in this case there are three features aligning):



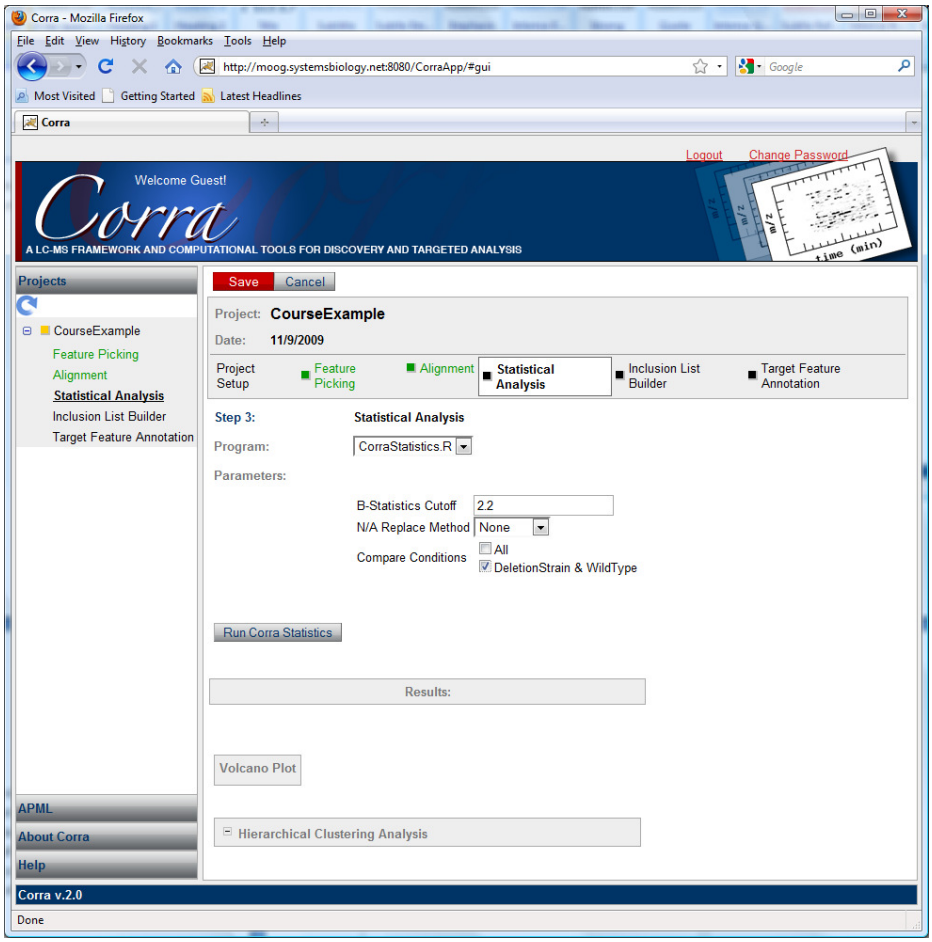


4. Statistical Analysis

4.1 Setup Statistical Analysis

Click on “Statistical Analysis” and “Edit.” The program in use is a collection of R modules called CorraStatistics.R. Set the “B-Statistics Cutoff” (B = -[log odds ratio]) or use the default of 2.2. Here we change it to 0. Usually, having a “N/A Replace Method” of “none” is satisfactory. The “N/A Replace Method” is to be used if you wish to fill in missing features with a value, either a minimum value or user-defined value. Use the drop-down menu to select the type of N/A Replace Method to use.

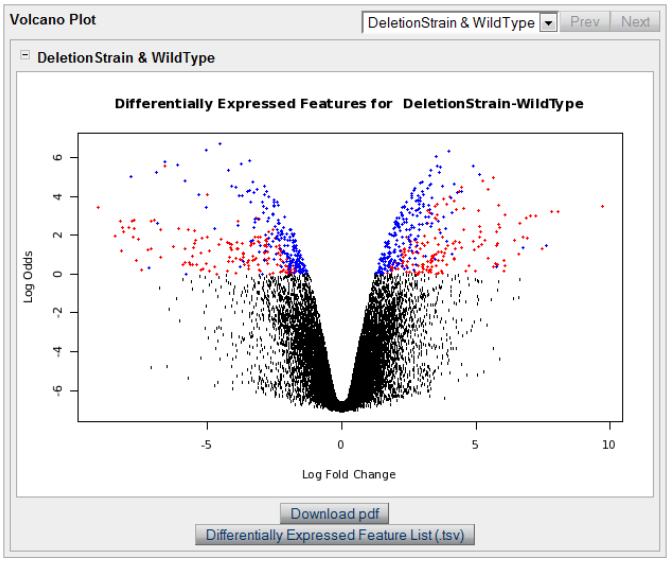
Select the comparisons to be calculated (red circle on figure below). Save the setup and the Statistics step will begin.



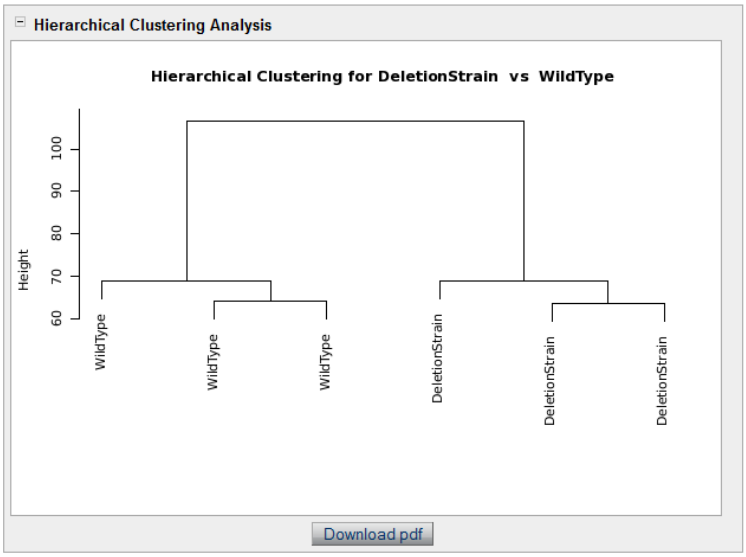
4.2 Results

The result of the statistical analysis is displayed in two plots, a volcano plot and hierarchical cluster (unsupervised) as shown below. The red dots are features which are found in some of the samples, but not all and have a Log Odds ratio greater than the value set in the “B-Statistics Cutoff” field (I set it to 0, here, but the default is 2.2). The blue dots represent features that were found in all of the samples with a Log Odds greater than the default or user defined limit. A Log Fold Change which is negative, indicates that a feature is more abundant in Condition 2 (here, WildType) whereas a positive Log Fold Change indicates that the feature is more abundant in Condition 1 (here, DeletionStrain).

The Volcano plot:



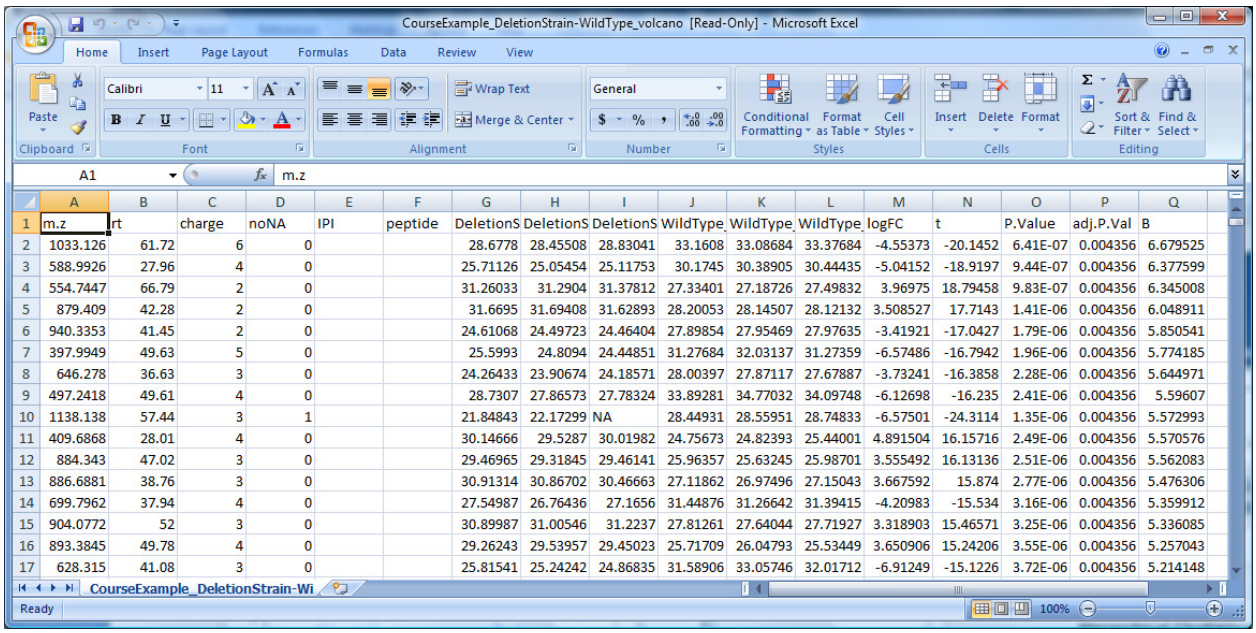
The Hierarchical cluster tree:



A tab delimited file (.tsv) is created for the aligned features and can be downloaded by clicking "Differentially Expressed Feature List (.tsv)" link below the Volcano Plot. The data contained in the .tsv file comes from an analysis using CorraStatistics.R (Bioconductor) as a backend. It is used as input in the following step, "Inclusion List Builder."

The "Differentially Expressed Feature List (.tsv)" file is shown below, opened in MS Excel. Note in addition to the data there is statistical information such as logFC (log fold change), p-value and B value.

5. Inclusion List Builder



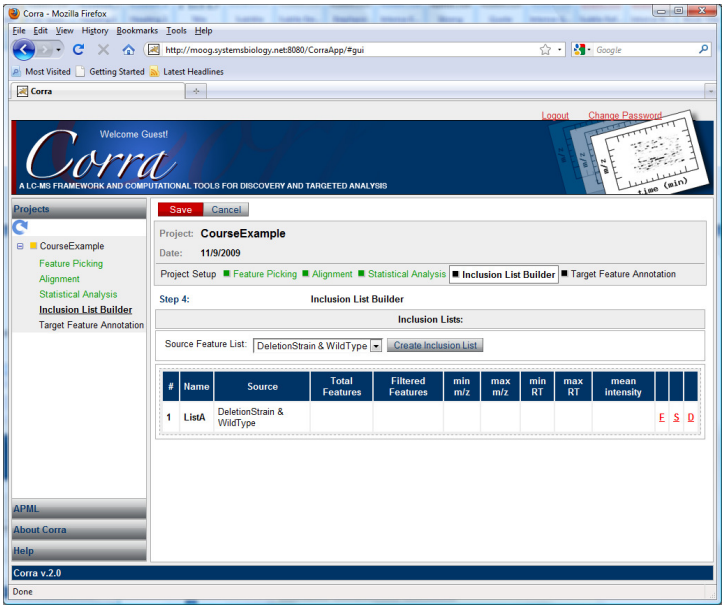
m.z	rt	charge	noNA	IPI	peptide	DeletionS	DeletionS	DeletionS	WildType_	WildType_	WildType_	logFC	t	P.Value	adj.P.Val	B
1033.126	61.72	6	0			28.6778	28.45508	28.83041	33.1608	33.08684	33.37684	-4.55373	-20.1452	6.41E-07	0.004356	6.679525
588.9926	27.96	4	0			25.71126	25.05454	25.11753	30.1745	30.38905	30.44435	-5.04152	-18.9197	9.44E-07	0.004356	6.377599
554.7447	66.79	2	0			31.26033	31.2904	31.37812	27.33401	27.18726	27.49832	3.96975	18.79458	9.83E-07	0.004356	6.345008
879.409	42.28	2	0			31.6695	31.69408	31.62893	28.20053	28.14507	28.12132	3.508527	17.7143	1.41E-06	0.004356	6.048911
940.3353	41.45	2	0			24.61068	24.49723	24.46404	27.89854	27.95469	27.97635	-3.41921	-17.0427	1.79E-06	0.004356	5.850541
397.9949	49.63	5	0			25.5993	24.8094	24.44851	31.27684	32.03137	31.27359	-6.57486	-16.7942	1.96E-06	0.004356	5.774185
646.278	36.63	3	0			24.26433	23.90674	24.18571	28.00397	27.87117	27.67887	-3.73241	-16.3858	2.28E-06	0.004356	5.644971
497.2418	49.61	4	0			28.7307	27.86573	27.78324	33.89281	34.77032	34.09748	-6.12698	-16.235	2.41E-06	0.004356	5.59607
1138.138	57.44	3	1			21.84843	22.17299	NA	28.44931	28.55951	28.74833	-6.57501	-24.3114	1.35E-06	0.004356	5.572993
409.6868	28.01	4	0			30.14666	29.5287	30.01982	24.75673	24.82393	25.44001	4.891504	16.15716	2.49E-06	0.004356	5.570576
884.343	47.02	3	0			29.46965	29.31845	29.46141	25.96357	25.63245	25.98701	3.555492	16.13136	2.51E-06	0.004356	5.562083
886.6881	38.76	3	0			30.91314	30.86702	30.46663	27.11862	26.97496	27.15043	3.667592	15.874	2.77E-06	0.004356	5.476306
699.7962	37.94	4	0			27.54987	26.76436	27.1656	31.44876	31.26642	31.39415	-4.20983	-15.534	3.16E-06	0.004356	5.359912
904.0772	52	3	0			30.89987	31.00546	31.2237	27.81261	27.64044	27.71927	3.318903	15.46571	3.25E-06	0.004356	5.336085
893.3845	49.78	4	0			29.26243	29.53957	29.45023	25.71709	26.04793	25.53449	3.650906	15.24206	3.55E-06	0.004356	5.257043
628.315	41.08	3	0			25.81541	25.24242	24.86835	31.58906	33.05746	32.01712	-6.91249	-15.1226	3.72E-06	0.004356	5.214148

Go to the “Inclusion List Builder” section. (Note: Inclusion List Builder depends upon the Statistical Analysis step which must be completed successfully).

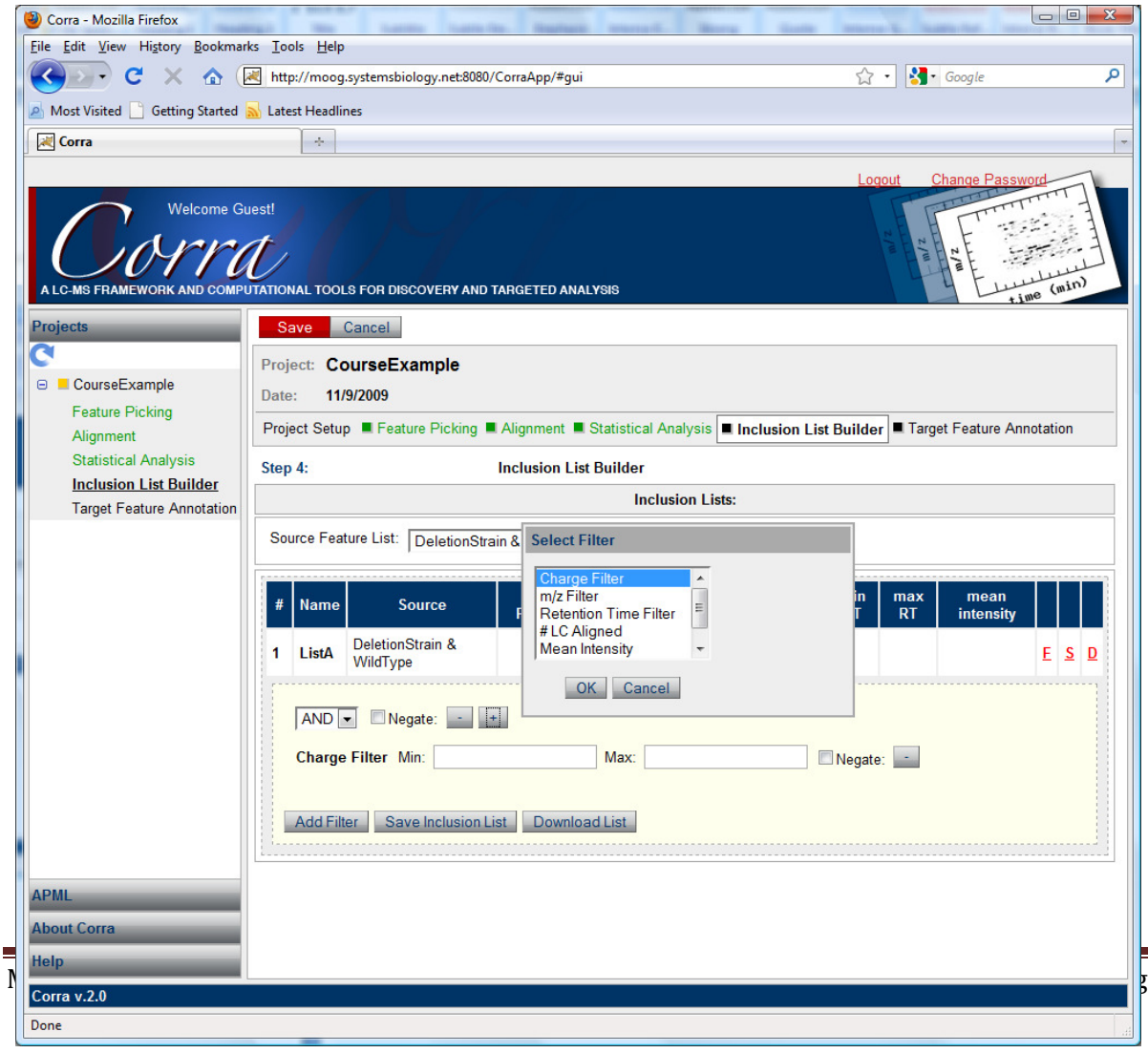
Click “Create Inclusion List” and give it a name.

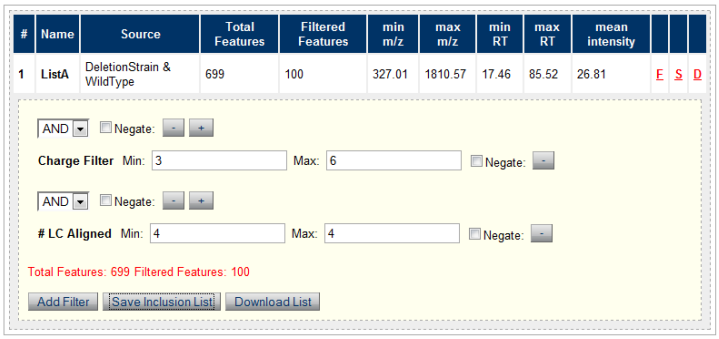
Click “F” to add a filter and “Add Filter.” Select a type of filter by clicking on the “+” sign (circled in blue in next figure) and using the drop down menu. To delete this filter click on the “-“ sign (circled in orange).

A “#LC aligned” filter might be useful if you wish to focus on the number of features aligned across all samples a certain amount of times.

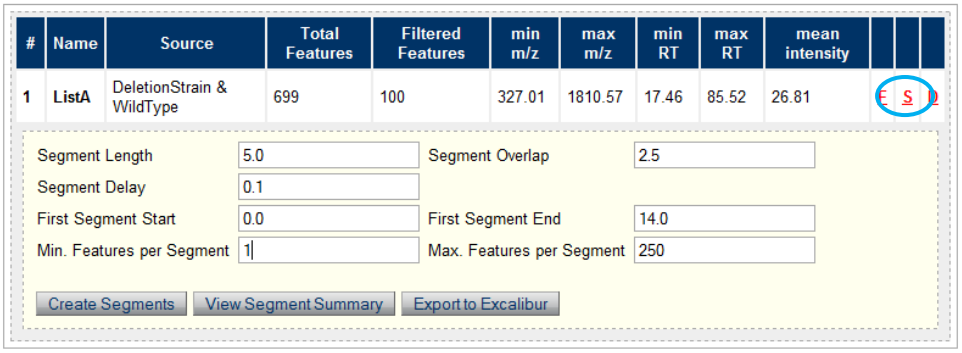


A “Mean Intensity” filter might be useful when features with low intensities wish to be excluded (Note: mean intensity is the $\log_2$ Intensity of a peak).





Hit "Save Inclusion List" and it will show how many total features and filtered features there are, plus some information such as min and max m/z (filters may be applied to limit these as well). Press "F" again to close the filter menu.



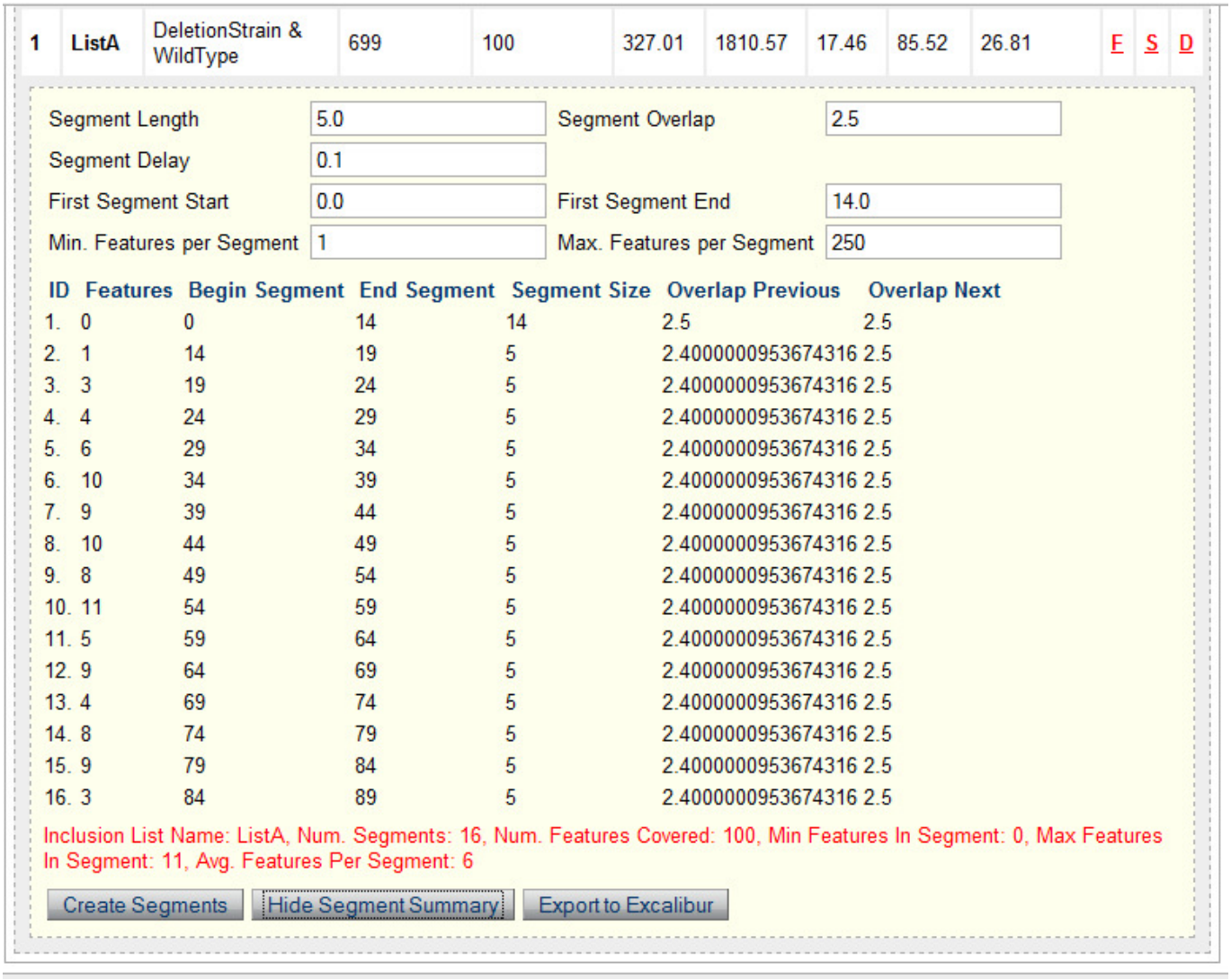
Click the red "S" to modify the segment settings. Segments can be useful when using the Thermofinnagin machine as these can be programmed into a target run. They can allow the machine to focus on certain parts of the run and not focus on others.

Segment Length is the "window" so to speak (minutes). The segment overlap is how many minutes one wishes to expand the window before and after the segment.

"First Segment Start" is usually just zero, but "First Segment End" is important as this first segment might capture parts of the run (usually at the beginning) where nothing very informative is happening. The "Min. Features per Segment" and "Max. Features per Segment" might be useful to play around with if there are too many features or too few.

Click "Create Segments."

To view the result of segmentation click on "View Segment Summary" and something like this should be displayed:



1	ListA	DeletionStrain & WildType	699	100	327.01	1810.57	17.46	85.52	26.81	F	S	D

Segment Length: 5.0 Segment Overlap: 2.5
Segment Delay: 0.1
First Segment Start: 0.0 First Segment End: 14.0
Min. Features per Segment: 1 Max. Features per Segment: 250

ID	Features	Begin Segment	End Segment	Segment Size	Overlap Previous	Overlap Next
1.	0	0	14	14	2.5	2.5
2.	1	14	19	5	2.4000000953674316	2.5
3.	3	19	24	5	2.4000000953674316	2.5
4.	4	24	29	5	2.4000000953674316	2.5
5.	6	29	34	5	2.4000000953674316	2.5
6.	10	34	39	5	2.4000000953674316	2.5
7.	9	39	44	5	2.4000000953674316	2.5
8.	10	44	49	5	2.4000000953674316	2.5
9.	8	49	54	5	2.4000000953674316	2.5
10.	11	54	59	5	2.4000000953674316	2.5
11.	5	59	64	5	2.4000000953674316	2.5
12.	9	64	69	5	2.4000000953674316	2.5
13.	4	69	74	5	2.4000000953674316	2.5
14.	8	74	79	5	2.4000000953674316	2.5
15.	9	79	84	5	2.4000000953674316	2.5
16.	3	84	89	5	2.4000000953674316	2.5

Inclusion List Name: ListA, Num. Segments: 16, Num. Features Covered: 100, Min Features In Segment: 0, Max Features In Segment: 11, Avg. Features Per Segment: 6

Create Segments | Hide Segment Summary | Export to Excalibur

In this case I only had 100 features to begin with so there are very few features in my segments so I might try to increase my segment length.

You may save list inclusion list by clicking “Export to Excalibur.”

6. Target Feature Annotation

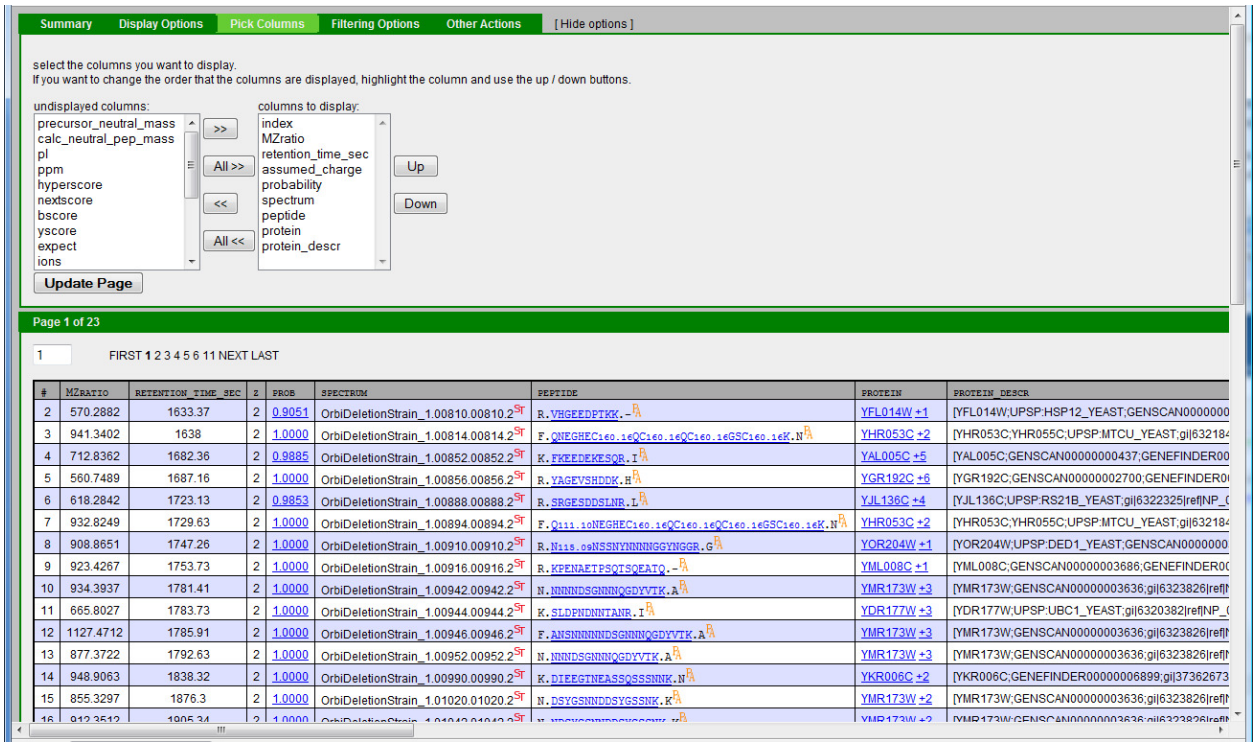
This module is to be used after MS/MS identification of peptide fragments to add sequence (and other protein descriptions) annotations to the original sample spectra, beginning the process of identifying proteins of interest. These could be the focus of future DDA or SRM analyses.

Target Feature Annotation annotates the statistical analysis output data (volcano plot data) based on the *m/z* values in a *.pep.xls* input file (provided by user).

6.1 Add an input file *somefile*.pep.xls

Note: This input file can be created using a pepxml viewer to convert a *.pep.xml* file to *.pep.xls* (e.g. PepXML Viewer – part of the TPP, see tutorial at http://tools.proteomecenter.org/wiki/index.php?title=TPP_Tutorial#PepXML_Viewer).

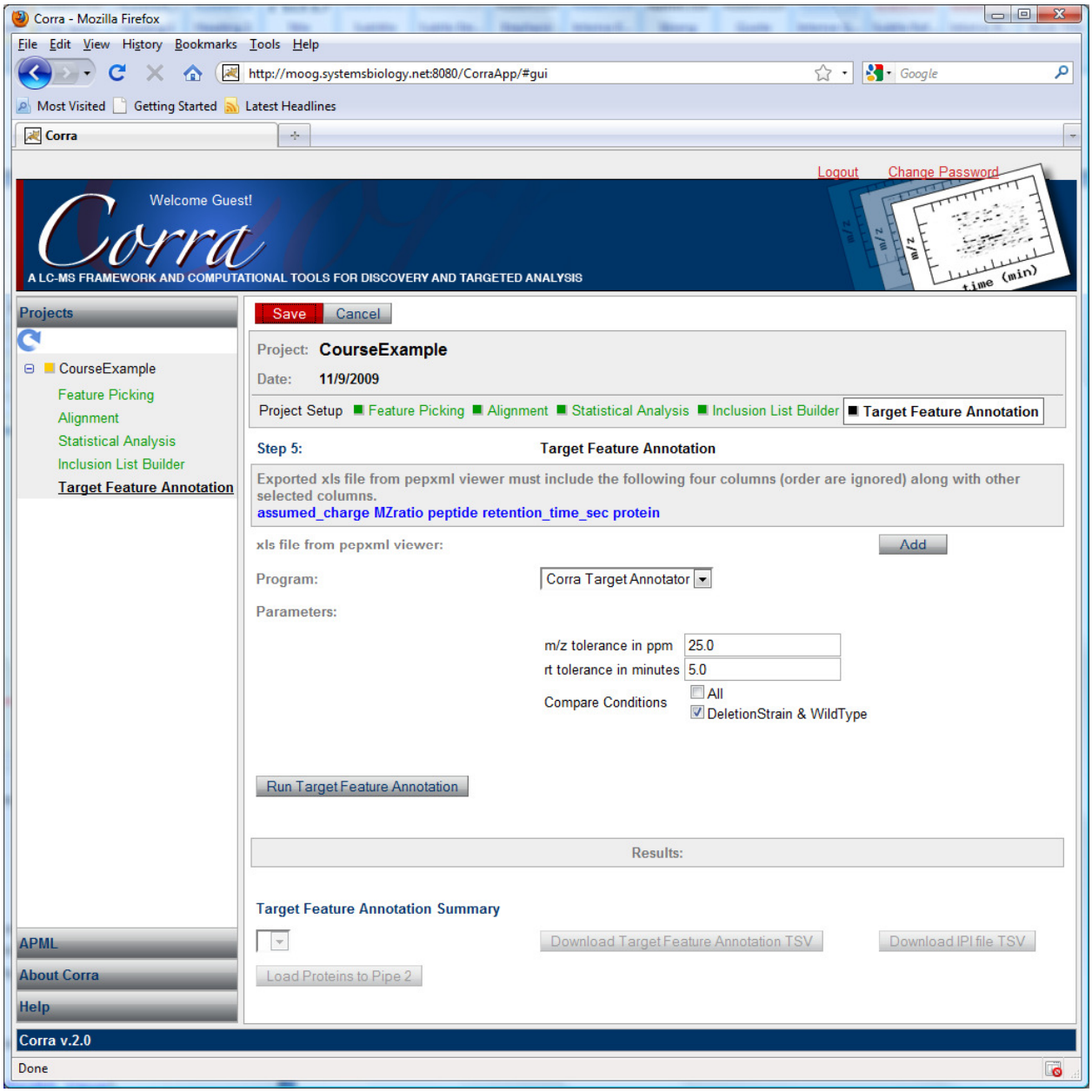
This is a screenshot from PepXMLViewer (uses PeptideProphet analysis):



The input interact.pep.xls file should have **at least** these headers (but you may add more, like the index and spectrum for instance):

assumed_charge MZratio peptide retention_time_sec protein

In order to run "Target Feature Annotation" you must Add an .xls (e.g., interact.pep.xls) file which has all of the possible annotations that may be queried and added to your data (volcano.tsv file actually). See next figure for adding a xls file.

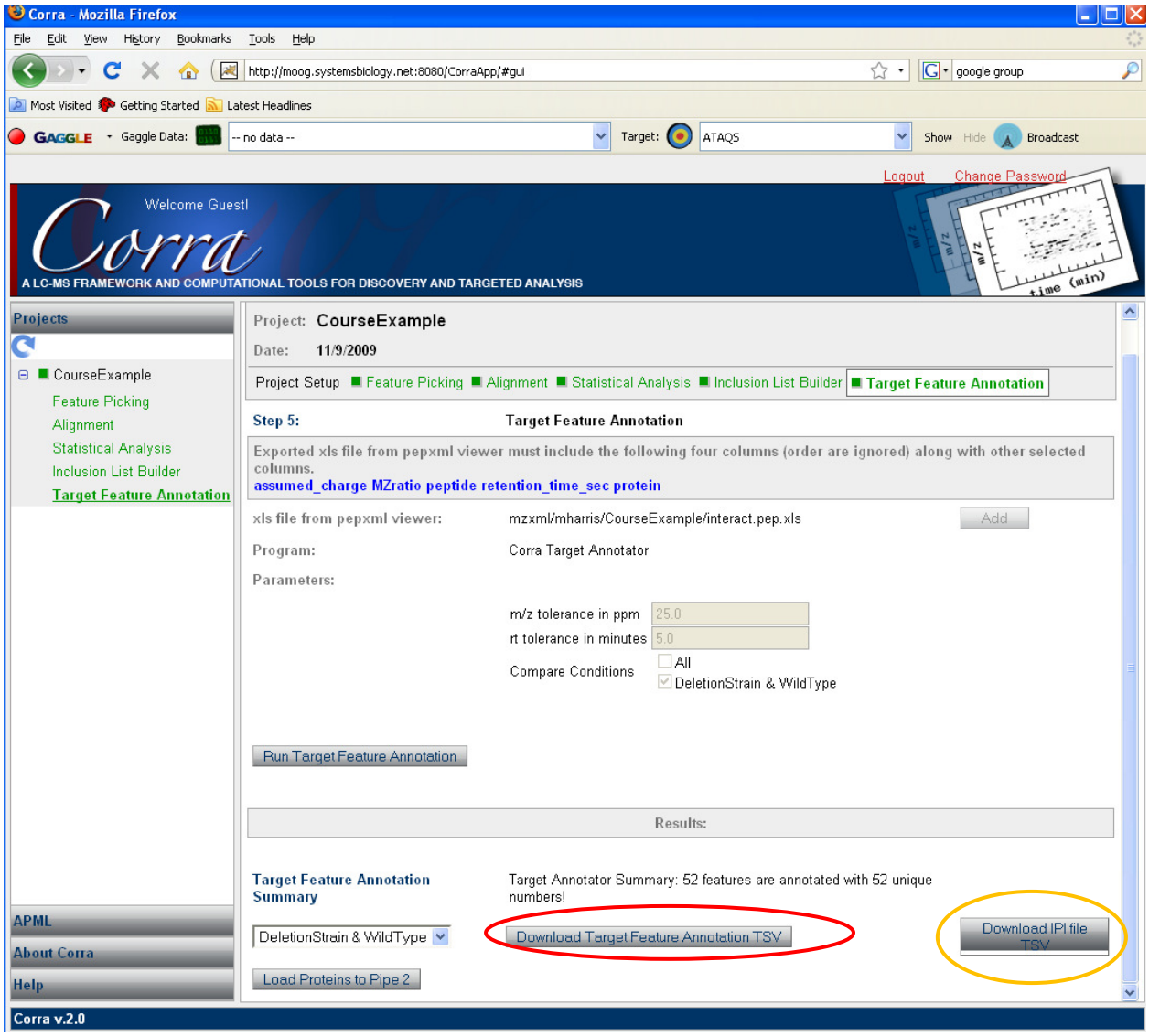


You may adjust the “m/z tolerance in ppm” which is set at a default to 25 ppm. Also, you may wish to adjust the” rt tolerance” in minutes. These are worth playing with if you do not get very many features annotated.

Then, hit “Run Target Features Annotation.”

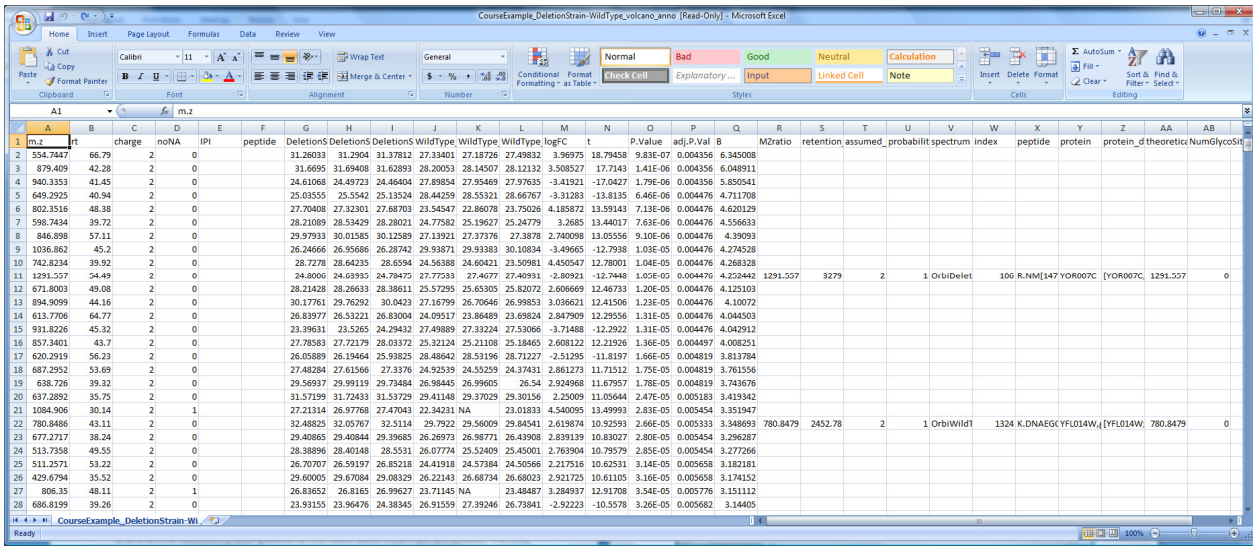
6.2 Results of Target Feature Annotation

Once Target Feature Annotation has run, you will have an annotated volcano.tsv file from section 4.



At this point, you may click “Download Target Feature Annotation TSV,” (circled in red) a file which looks similar to the “Differentially Expressed Feature List (.tsv)” from section 4, but with additional information including peptide descriptions for some of the features. Below, the annotated *.tsv* file is shown opened in MS Excel.

You may also download the “IPI file TSV” which just contains just the features which are associated with IPI(s).



6.3 PIPE2

You may load the results of Targeted Feature Annotation into PIPE2 by clicking “Load Proteins to Pipe2.” Alternatively, you can copy and paste your IPIs, ORFs or other feature identifiers into PIPE2 to map them to several other databases, providing additional information about these important features. The PIPE2 link is here:

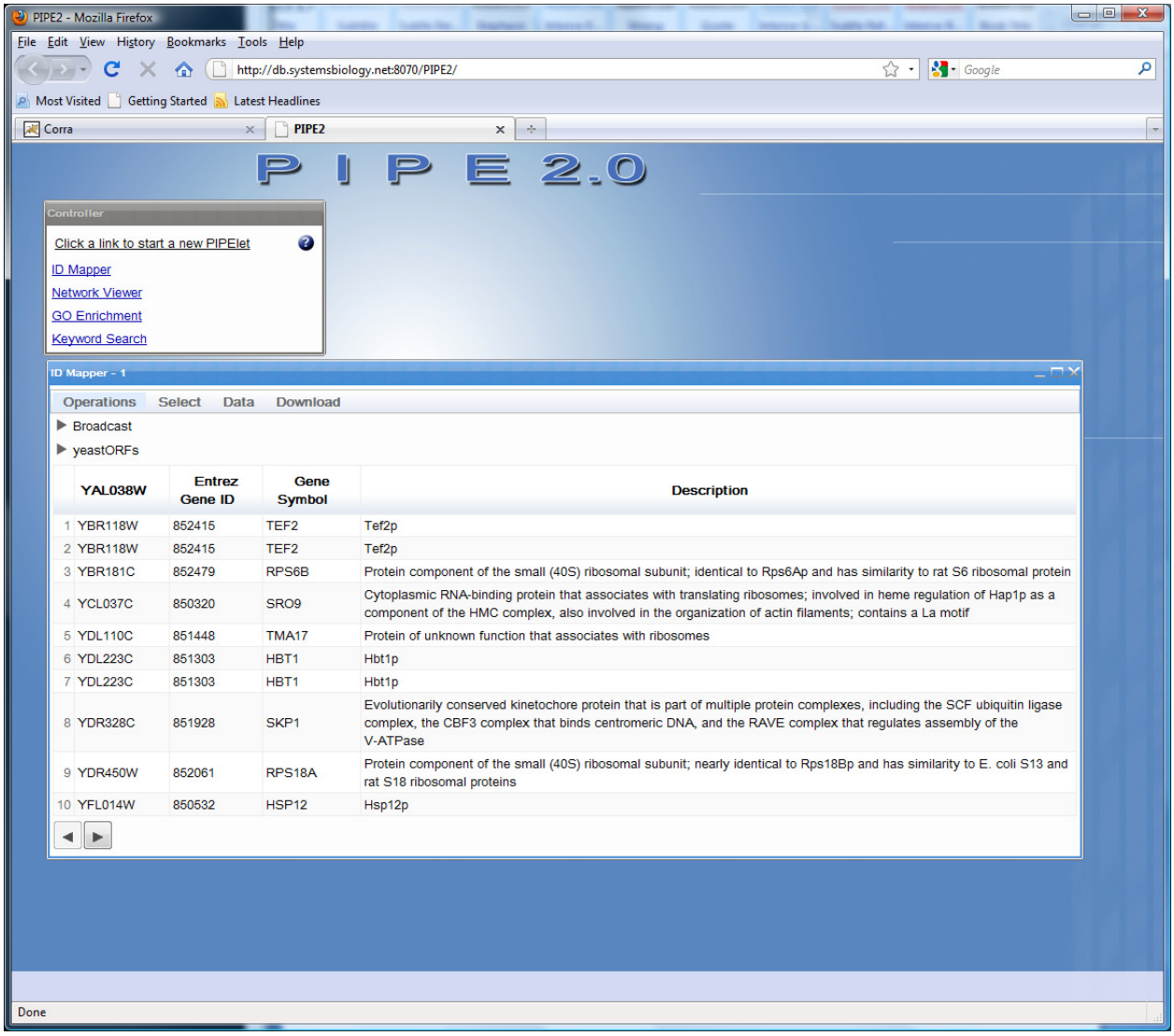
http://db.systemsbiology.net:8070/PIPE2/

Note: you must have firegoose extension installed in your computer when using Mozilla Firefox browser (http://gaggle.systemsbiology.org/docs/geese/firegoose/install/).

Here is a screenshot of our yeast DeletionStrain and Wildtype data into PIPE2 after Target Feature Analysis (remember we found 52 aligned and annotated features).

This links to a tutorial of PIPE2:

http://db.systemsbiology.net:8070/PIPE2/PIPE2/docs/PIPE2_tutorial.doc



	YAL038W	Entrez Gene ID	Gene Symbol	Description
1	YBR118W	852415	TEF2	Tef2p
2	YBR118W	852415	TEF2	Tef2p
3	YBR181C	852479	RPS6B	Protein component of the small (40S) ribosomal subunit; identical to Rps6Ap and has similarity to rat S6 ribosomal protein
4	YCL037C	850320	SRO9	Cytoplasmic RNA-binding protein that associates with translating ribosomes; involved in heme regulation of Hap1p as a component of the HMC complex, also involved in the organization of actin filaments; contains a La motif
5	YDL110C	851448	TMA17	Protein of unknown function that associates with ribosomes
6	YDL223C	851303	HBT1	Hbt1p
7	YDL223C	851303	HBT1	Hbt1p
8	YDR328C	851928	SKP1	Evolutionarily conserved kinetochore protein that is part of multiple protein complexes, including the SCF ubiquitin ligase complex, the CBF3 complex that binds centromeric DNA, and the RAVE complex that regulates assembly of the V-ATPase
9	YDR450W	852061	RPS18A	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps18Bp and has similarity to E. coli S13 and rat S18 ribosomal proteins
10	YFL014W	850532	HSP12	Hsp12p

7. Trouble Shooting

Problem	**Fix**
The mzxml file format is invalid	try using indexmzXML.exe to correct errors
PIPE2 button does nothing in firefox browser	Install firegoose extension