Experimental Design and MS Workflows for Omics Applications

Daniel Cuthbertson
Application Scientist
Agilent Technologies
Denver, CO

Acknowledgements: Support Team

AFO Metabolomics Team

Mark Sartain, Sumit Shah, Anne Blackwell, David Weil

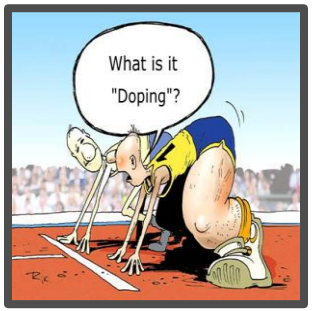
- Yuqin Dai — Santa Clara, CA
- Steve Madden — Software Product Manager
- Rick Reisdorph — National Jewish, Denver, CO
- MPP Support Team — Global

Outline

- Basics of Experimental Design and Statistics in “Omics” Research
- Example Multi-Omic Study with MPP 13
- Update What’s New in MPP and Profinder
  - Sample Correlation, Pathways, Metadata, etc.

Bio/Pharma

Life Sciences



Clinical/Diagnostics

Environment

Small Molecule Profiling

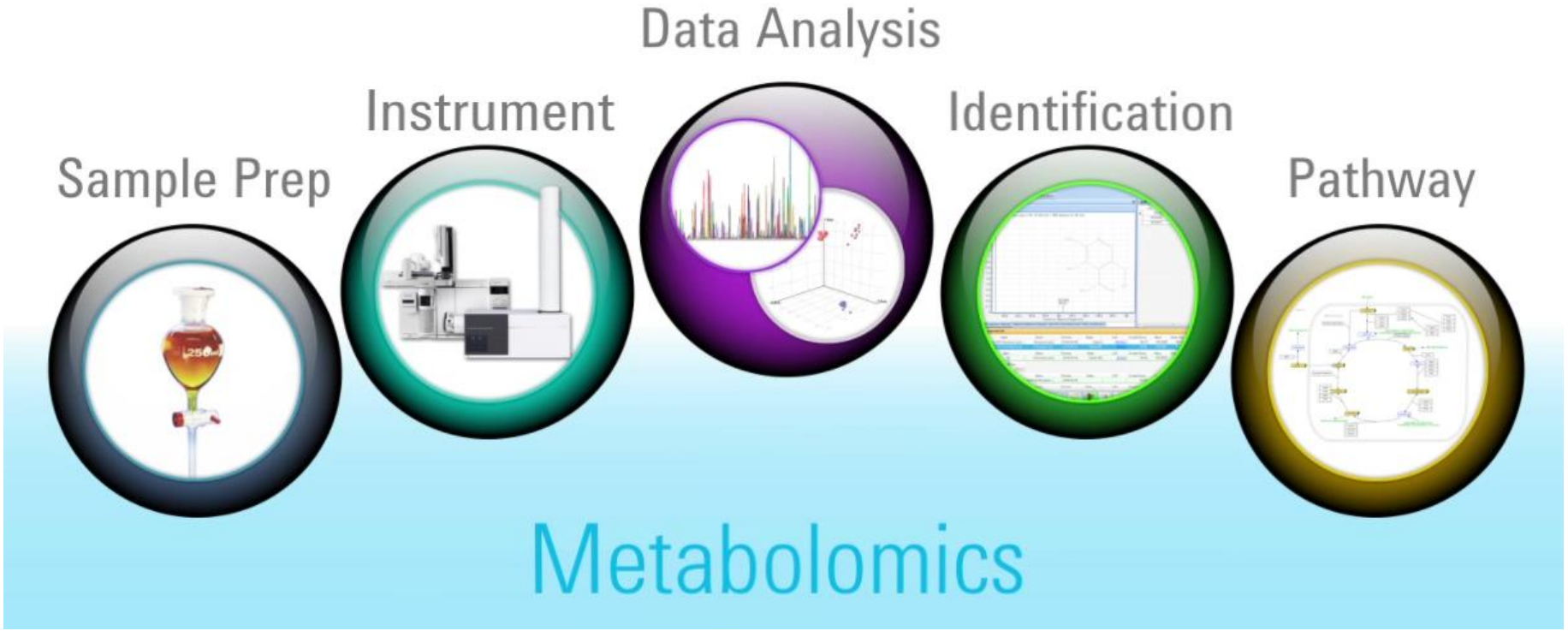
Forensics/Toxicology

Petrochemicals

Agriculture

Nutrition

Consumer



Experimental design is critical for Omics experiments.
All steps are critical from sample preparation to identification.
Your experimental design will be statistically driven!

Why Use Statistics?

- To Avoid jumping to conclusions.
- To Avoid finding patterns in random data
- Understand multiple comparisons
- Consider alternative explanations.

What Will Statistics Do and Not Do?

Will not:

- Give you absolute answers. Statistical conclusions give only probabilities.
- Give you scientific or clinical conclusions. It is your job to interpret the statistics and draw conclusions in context of the hypothesis.

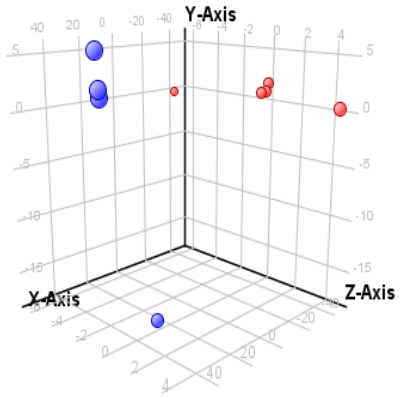
Will:

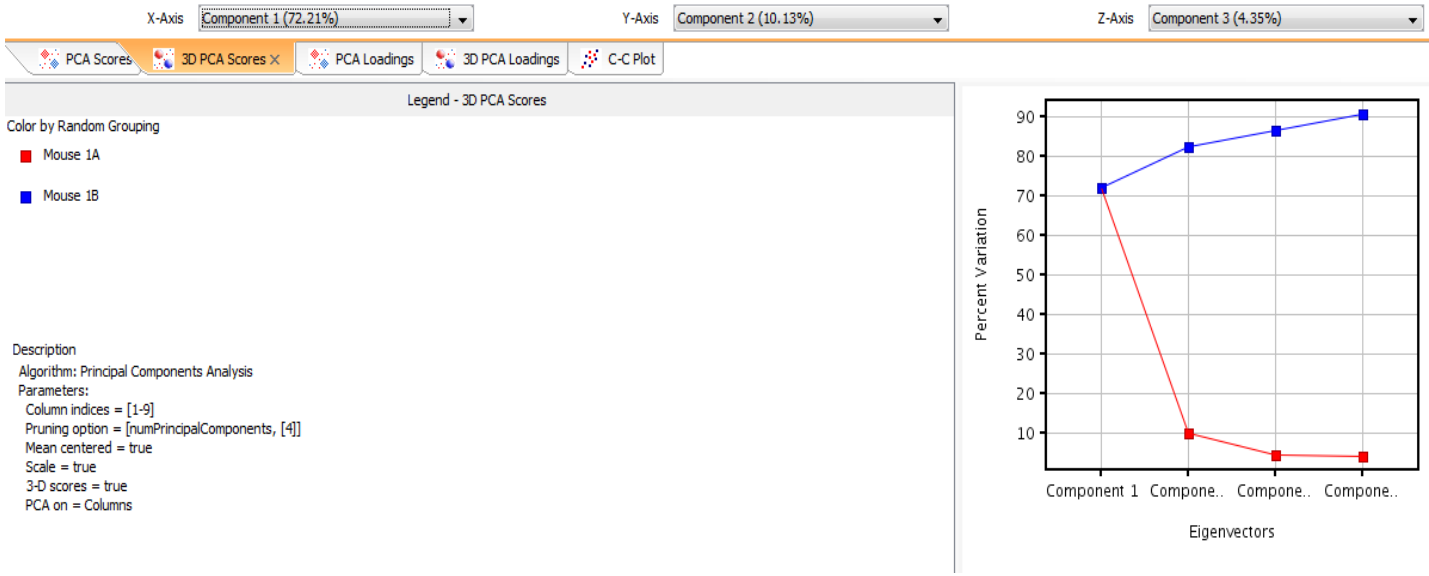
- Help you assess variability.
- Help you extrapolate from a sample to a population.
- Help you uncover relationships between variables.

What's Wrong With "Fishing"?

Workflow:

- Filter on Frequency
- T-Test 0.05 P-value cutoff
- PCA





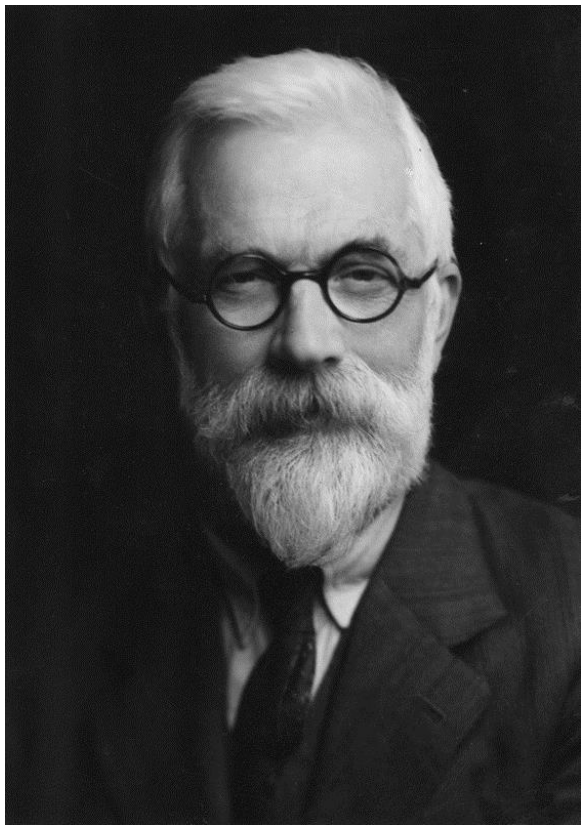
These are technical replicates from the same mouse!

What's Wrong With "Fishing"?

- The first thing you should do is determine goals and hypothesis for the experiment.
- Develop a statistical plan. Will the statistics I use help me answer the question?
- Use the hypothesis and use your *a priori* knowledge to evaluate whether to statistical results make sense or not.
- If they results don't fit with you're *a priori* knowledge then there may be some factor or error you haven't considered yet
- It's possible to discriminate groups arbitrarily on noise.
- Always consider the null or alternative hypothesis

These Basics are Fundamental to the Success of Your Statistical Design

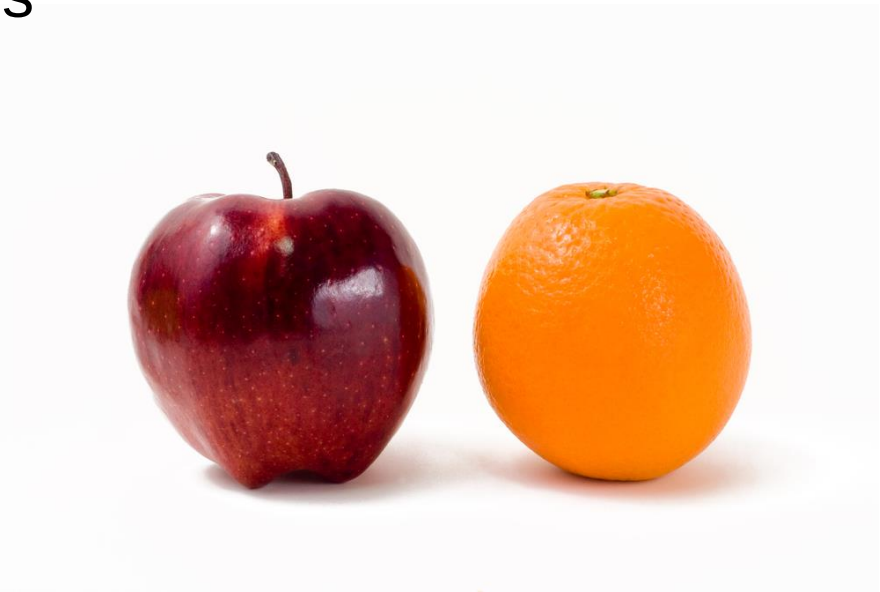
- Comparison
- Randomization
- Blocking
- Replication
- Factorial study design



R.A. Fischer

Metabolomics and Lipidomics are Fields of Comparison

- “Omics” applications often involve the profiling of hundreds or thousands of compounds.
- We often cannot get standards for each compound, thus we cannot do independent comparisons.
- Thus quantitation is a relative comparison between groups.
- We need statistics to discern differences between groups



Randomization is Essential to Control for Variance Related to Sample Batch

- Sample run order, sample processing order, solvent lots, different columns, different people prepping samples can all effect the data.
- To best distribute this error randomize the order in which samples are processed and acquired on the instrument.
- Failure to do so can lead to false discovery just by order in which samples were run!

Use appropriate randomization techniques, don’t haphazardly reorder your samples!

Blocking Further Reduces Error Due to Non-Experiment Variance

- Blocking arranges the experiment into lots that are similar to one another.
- Blocks contain mixtures of different sample types
- Use ‘Randomized Block’ designs for large studies

Example Randomized Block

Block 1:

- Fuji 3
- Gala 2
- Red Delicious 4
- Honey Crisp 2

Block 2:

- Red Delicious 2
- Gala 1
- Fuji 1
- Honey Crisp 4

Block 3:

- Gala 3
- Red Delicious 1
- Honey Crisp 3
- Fuji 2

Block 4:

- Fuji 4
- Honey Crisp 1
- Gala 4
- Red Delicious 3

What Should You Consider When Determining Number of Replicates

- **Type of replicate**:
  - **Technical Replicates**: Instrument and/or Sample Prep

    - *ONLY tell you about the variability in the mean measurement of a single sample*
  - **Condition Replicates**: Biological, Raw Material Lots, etc.

    - *Must be independently sampled from population*
    - *Will tell you about variability in the population*
- **Statistical Power**
  - Asks how many replicates you need to see the desired effect?
  - You must have a hypothesis:
    - *You must estimate the effect size or group means and their variability*

What is Statistical Power?

A measure of confidence in statistical results

Key Terms to know

α : Error in finding something statistically significant when the null hypothesis is true (set typically to desired p-value i.e.. 0.05) -False Positive rate

β: Error in finding something not statistically significant when the null hypothesis is false – False Negative rate

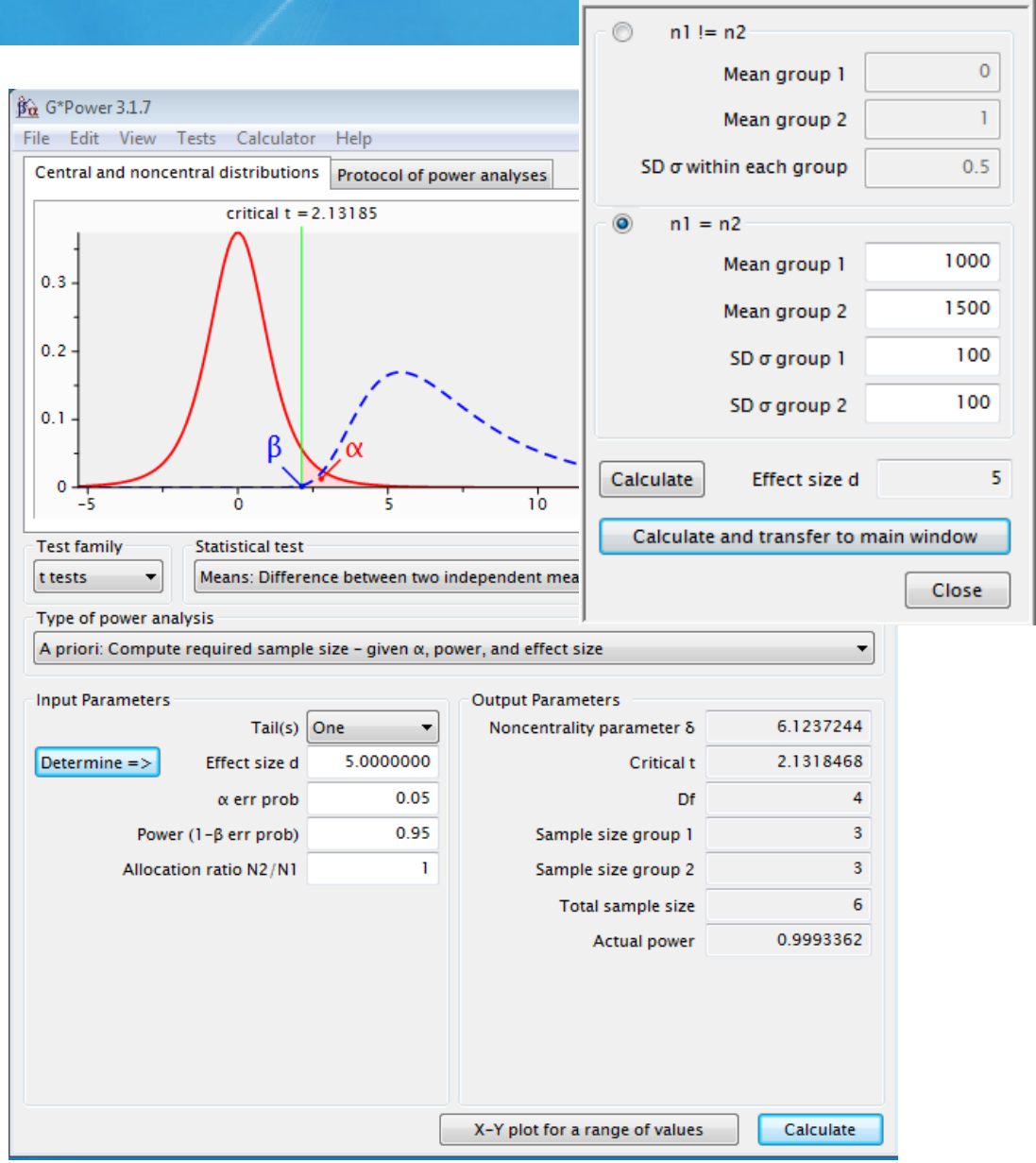
Power = 1- β

- If your experiment is underpowered you cannot be confident in the differences you see (or don't see)

Power Analysis

- Power Analysis Helps you determine the appropriate number of replicates to see a desired effect at a specified power
- You must estimate effect size or group means and Standard Deviations. (Hypothesis)
- Comparing more groups (i.e., ANOVA) typically requires more replicates
- Many online resources or you can download Gpower for free

http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/



Full Factorial Study Designs Are Used to Detect Interactions Between Treatments

- Used to investigate each of the primary variables and their interactions.
- Can be analyzed using ANOVA, Regression and Multivariate Statistics.
- Number of groups to test uses the (# Levels)^#Factors
- Full factorial designs can be laborious for studies complex studies.

2 Factors X 2 Levels
Ex. Drug A 100 um; Drug B 100 um

	A	B
Group 1	-	-
Group 2	+	-
Group 3	-	+
Group 4	+	+

3 Factors X 2 Levels
Ex. Drug A 100 um; Drug B 100 um; Drug C 100 um

	A	B	C
Group 1	-	-	-
Group 2	+	-	-
Group 3	-	+	-
Group 4	-	-	+
Group 5	+	+	-
Group 6	+	-	+
Group 7	-	+	+
Group 8	+	+	+

2 Factors X 3 Levels
Ex. 80C (10 min, 20 min, 30 min)
100C (10 min, 20 min, 30 min)

	A	B
Group 1	0	0
Group 2	1	0
Group 3	2	0
Group 4	0	1
Group 5	0	2
Group 6	1	1
Group 7	2	1
Group 8	1	2
Group 9	2	2

Fractional Factorial Studies May Reduce Effort

Example 1:

- Use when an interaction is not expected (or possible) between treatments. Ex: Time 0. Time 1, Time 2, Time 3
- Use one way ANOVA.

	A	B	C
Group 1	-	-	-
Group 2	+	-	-
Group 3	-	+	-
Group 4	-	-	+

Example 2:

- Use to estimate interactions of many variables with a specific factor.
- Other interactions are not expected or important to study. Otherwise may be confounded by two factor interactions.
- Ex: Interaction of drug 1 with 3 other coadministered drugs

	I	A	B	C	AB	AC
Group 1	+	-	-	-	+	+
Group 2	+	+	-	-	-	-
Group 3	+	-	+	-	-	+
Group 4	+	+	+		+	-
Group 5	+	-	-	+	+	-
Group 6	+	+	-	+	-	+
Group 7	+	-	+	+	-	-
Group 8	+	+	+	+	+	+

Other designs may be used for even more complex studies. Ex: Plackett-Burman

Paired Studies Can Increase You Power When Samples Are Related

Unpaired tests are used when the samples being compared are independent or unrelated

- Typical of most experiments

Paired tests are used when the samples are related

- Most common paired design is one in which one variable represents different individuals and the other variable represents “before” and “after” treatment
- If variability between individuals is expected to be large and the effect of treatment is small, than a very large sample size is needed to detect effect of treatment using t-test
- Using a paired t-test gives much more statistical power when difference between groups is small relative to variation within groups

Example of paired data: patient data before and after treatment

	FEV (%) Pre-treatment	FEV (%) Treatment day 3	FEV (%) Treatment day 6
Patient 1	60	70	75
Patient 2	80	85	82
Patient 3	50	60	75
Patient 3	100	110	90
Patient 4	65	70	65

Sampling and Sample Preparation

- Sample Preparation is the primary source of non-experimental variance in differential analysis
  - Starting sample amount
  - Inconsistent pipetting
  - Operator/ chemist
  - Sample degradation and freeze thaw cycles
  - Mixing and fractionation

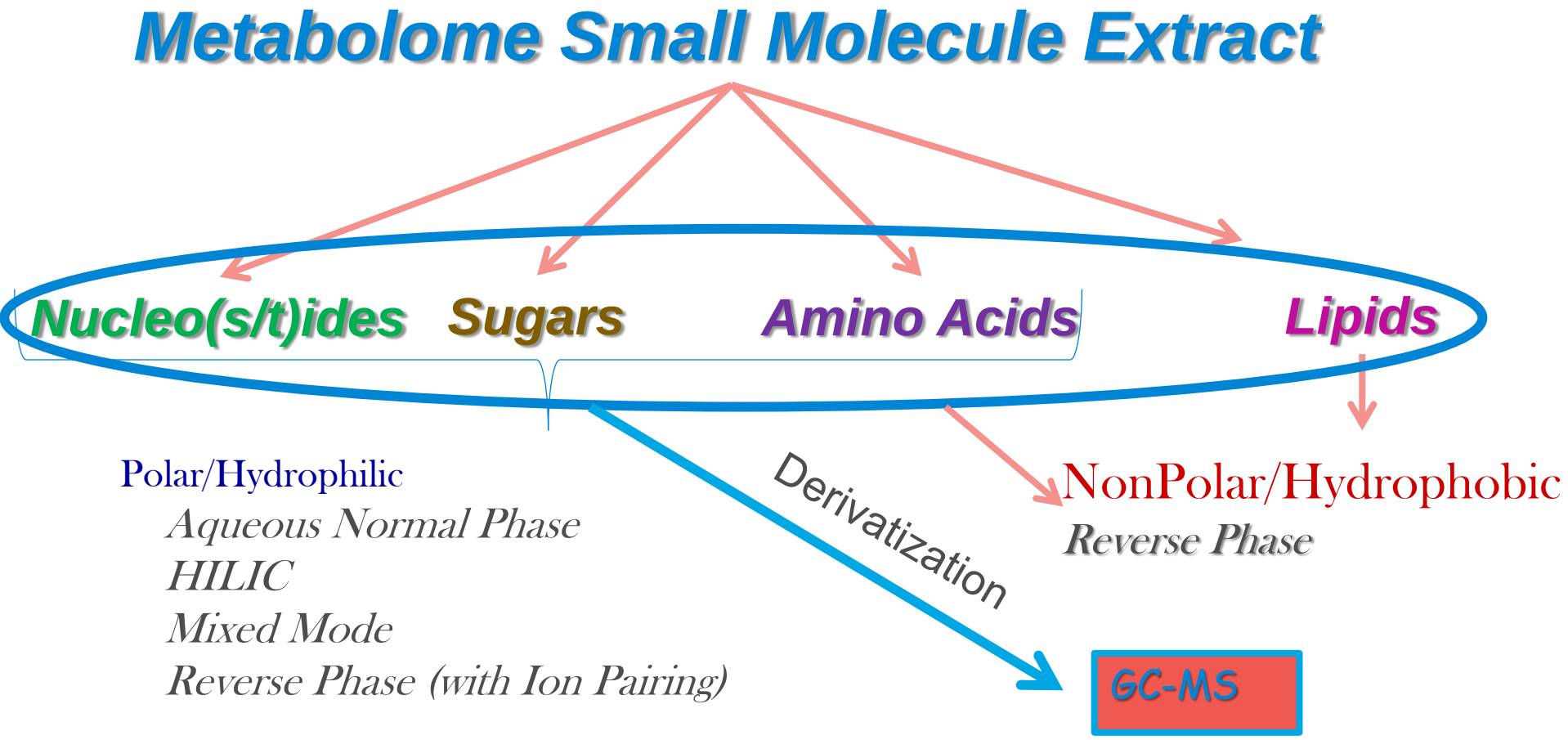
- Reduce impact of non-experimental variance
  - Careful method validation will help estimate variability
  - Use controls and QCs to account for variability
  - Normalization will only get you so far

Sources of Variation: Sample Preparation and Handling

Sample preparation best practices

- Try different sample preparation workflows.
  - Consult literature.
  - Use appropriate chemistries, techniques.
- Practice sample preparation and assess reproducibility via MS runs.
- Plan ahead. Make sure you won't run out of reagents mid-experiment.
- Be consistent in all aspects.
  - Sample collection
  - Storage
  - Freeze/thaw cycles
- Be precise in pipetting. Make sure pipetman are calibrated and functioning properly.
- Pre-label all tubes.
- Avoid mixing of phases during fractionation.

No universal separation method exits to profile all the classes of metabolites in a single LC-MS run



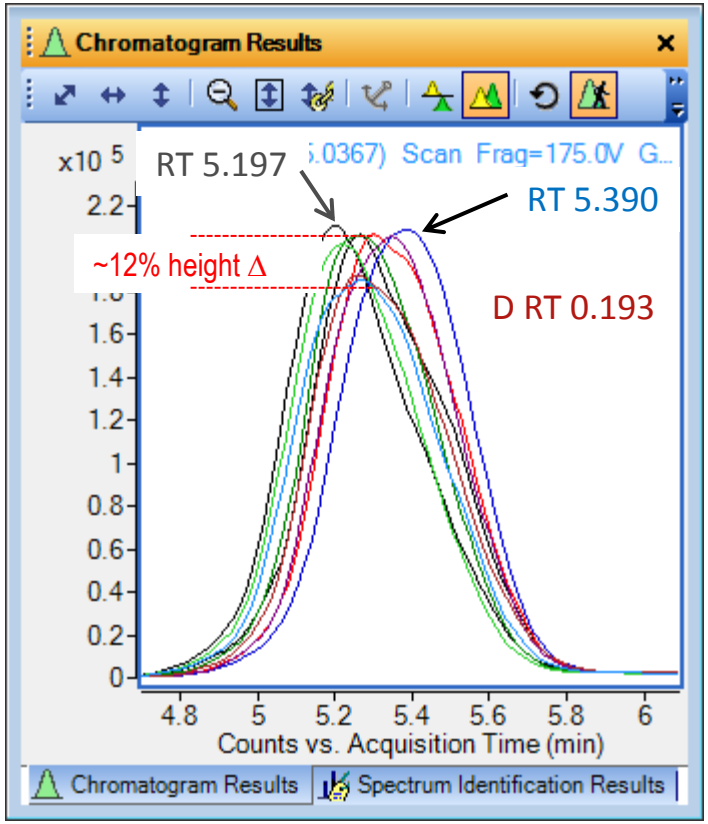
Factors to Consider When Choosing a Chromatography Method

- Profiling versus Targeted?
  - Profiling compromises settings for any specific compound class for greater overall coverage
  - Targeted Optimizes for specific pathways or compound classes
- What's your hypothesis?
  - Use your biology/chemistry to guide you. What types of molecules do you want to separate. (You can't have it all)
  - Choose a separation type most compatible with your molecules of interest.
- What is your need for speed versus resolution?
  - Time is solvent and money
  - Chromatographic resolution increases coverage, reduces interferences, reduces CV's and helps feature extraction.

Chromatography Is A Source Of Variability

- Variability in retention time and peak shape are sources of error
- Each phase and each compound has it's own inherent variability
- How often will you need to change columns during an experiment?
  Plan for enough columns!

  Column changes are a source of variability!
- MPP can help with alignment and RT correction!



What Instrument Will You Use?

LC - QTOF

- **Targeted data mining: Pathway Directed**
  - Analyze data with a database of known compounds
  - Project results onto pathways
- **Untargeted data mining:**
  - Find all compounds
  - Naïve data mining: Discovery based approach
  - Track metabolites using mass, or mass spectra and retention time

LC - QQQ

- **Hypothesis Driven; you select likely metabolites**
- Known metabolites only
- Higher sensitivity than profiling approach
- Absolute quantitation
- Typically develop methods for *tens/hundreds* of metabolites
- Project results onto biological pathways for interpretation

GC - QTOF

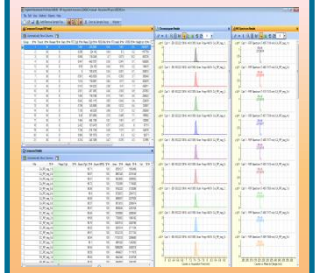
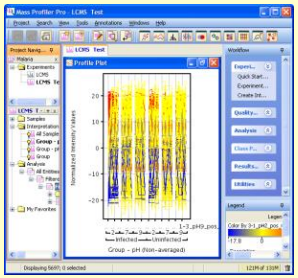
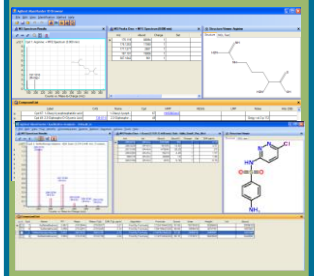
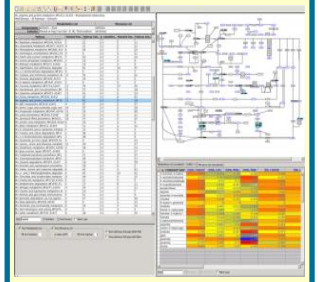
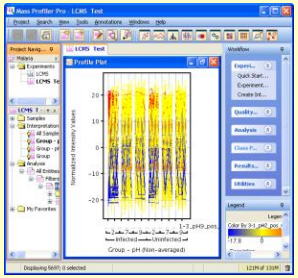
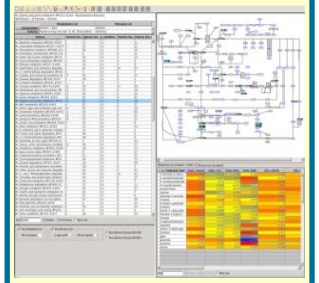
- **Analysis of Volatiles**
  - Headspace analysis
  - Derivitization of metabolites
- Can provide an excellent sampling of many difference compounds classes in a single run.
- Robust and predictable chromatography
- Extensive libraries for identification

Sources of Variation: Mass Spectrometry Data Acquisition

Mass Spectrometry Data Acquisition

- Be thorough in method development. Make sure MS and LC method parameters are optimal for buffer system, sample type, etc.
- Validate your method. Know what to expect!
- Use fresh mobile phase preparations.
- Change guard columns regularly.
- Purge LC pumps regularly. Retention times must be stable.
- Define and adhere to appropriate schedule for cleaning the source.
- Utilize QC samples to monitor system performance, especially for long term experiments involving many samples.
  - Use to assess RT and abundance reproducibility
  - Can provide useful false discovery information (MPP)
- Randomize sample run order.

Untargeted/Pathway Directed Workflow

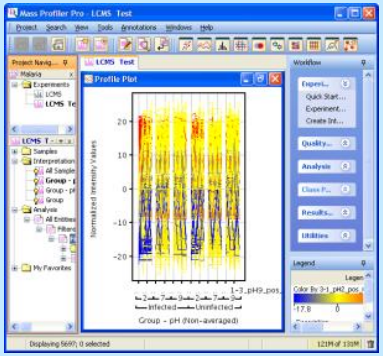
	Separate & Detect	Feature Finding	Alignment & Statistics	Identify	Pathways
Untargeted	 LC-TOF/QTOF	 Profinder	 Mass Profiler Professional	 ID Browser	 Pathway Architect
Pathway Targeted	 LC-TOF/QTOF	 Profinder	 Mass Profiler Professional		 Pathway Architect

LC-QQQ Targeted Metabolomics Workflow

1. Optimize LC/MRMs for metabolite standards using LC/QQQ and Optimizer
2. Acquire MRM data using LC/QQQ and **Study Manager**
3. Quantitation of metabolites using MassHunter Quantitative Analysis exporting project into MPP
4. Statistical analysis using Mass Profiler Professional
5. Choose pathway database and species
   - Analyze differential data and visualize the results on pathways

In Major Studies Instrument Types Work Together

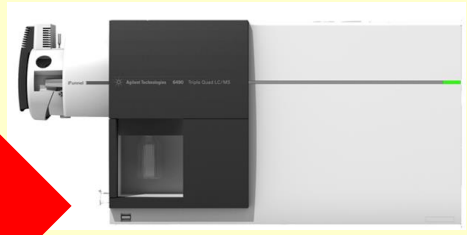
Discovery and Identification

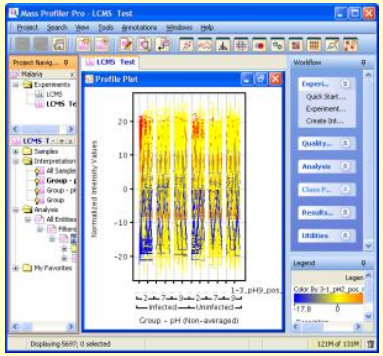


- Initial Pilot Study
- Acquisition of profiling data using appropriate chemistry's

- Statistical Analysis
- Identification of putative biomarkers
- Structural Annotation
- Power Analysis for validation study

Validation and Translation



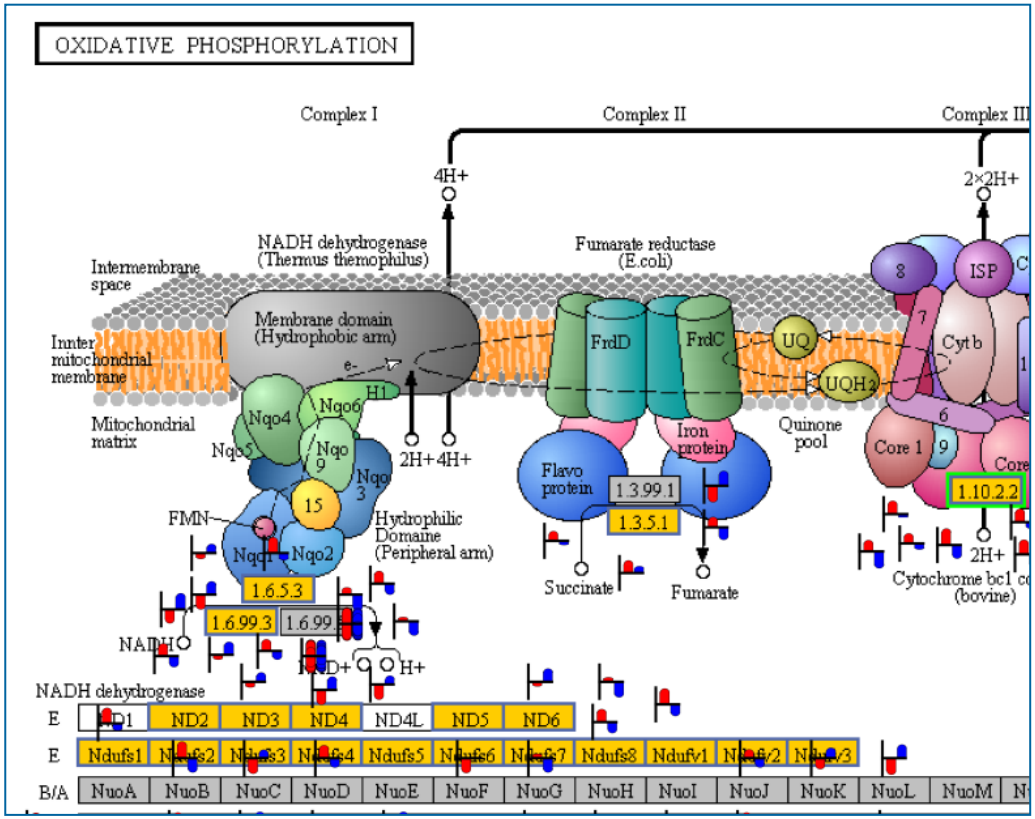


- Validation of putative biomarkers
- Replicate numbers determined by power analysis

- Final Statistical Analysis
- Publication
- Translational Model Building

Design of Integrated Biology Experiments

- Be sure to have ‘equivalent’ conditions for each “omic”
- Be especially aware of confounding variables and artifacts that are unique to each data type.
- Leverage Proteomics and Genomics to reduce noise in Metabolomics
- Correlate results in multiple “omics” to increase confidence



Adjusting for Unwanted Variability – Normalization

Normalization:

- External Scalars
  - Osmolality, Protein Content, Cell Count, Etc.
- Algorithmic
  - Total "Useful" Current and Percentile Shift:
    Assumes that the sum concentration of all analytes is the same in every sample.
  - Quantile:
    Assumes that the distribution of intensities in the samples is the same

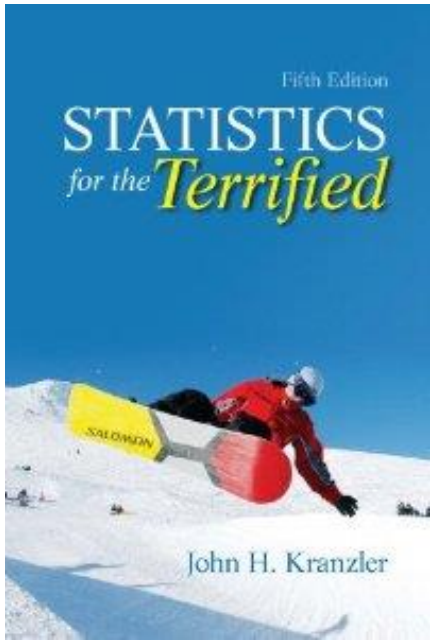
Key Point: Each normalization makes different assumptions about the sample. Violating these assumptions can introduce more error than you hope to correct.

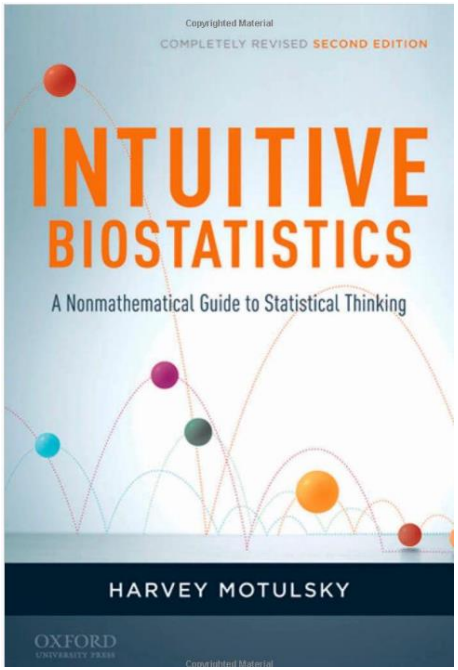
Key Point: Normalizations can be very sensitive to missing values

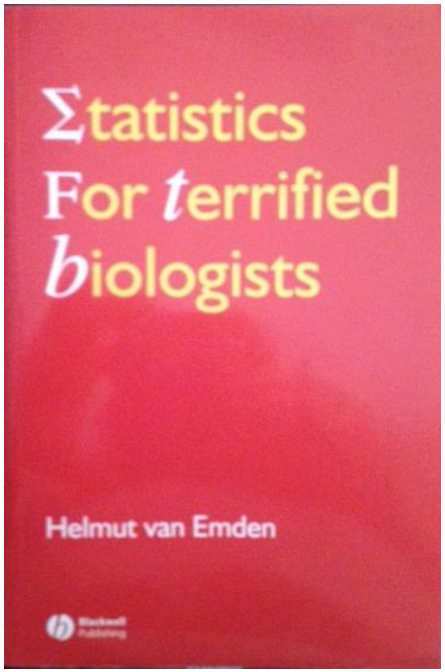
Key Point: Normalization is not always necessary

Just a brief overview….

Lots of books and online learning..



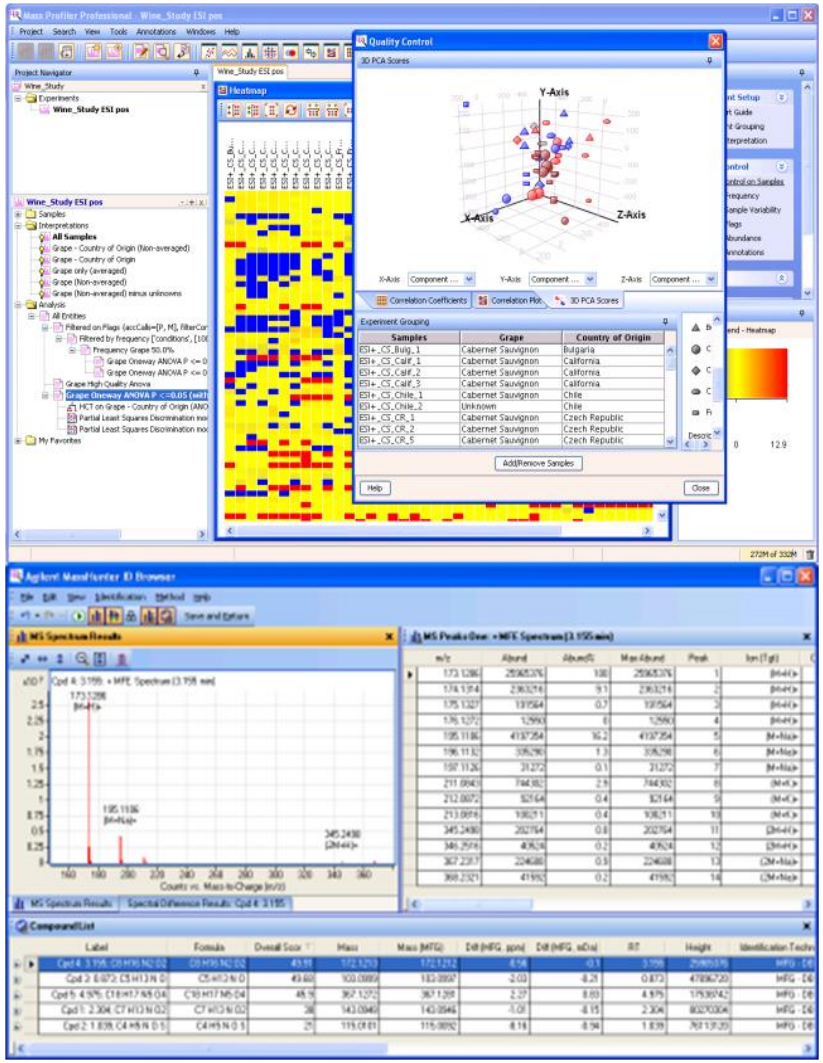




Mass Profiler Professional (MPP) 13:

Statistical Analysis and Visualization Software

- Designed for Mass Spectrometry data from multiple platforms
- **Can Import, Store, and Visualize**
  - Agilent LC/MS Q(TOF), and QQQ
  - Agilent GC/MS Quad, QQQ, and QTOF
  - Agilent ICP/MS and NMR (Craft)
  - Generic file format import
- **Extensive statistical analyses tools**
  - ANOVA, Clustering, PCA, Fold-change, Volcano plots
  - Correlation Analysis, including multi-"omic" correlation!
- ID Browser for compound identification
- Integrated Biology "Omics"
- Pathway Architect for biological contextualization
- NEW! KEGG Pathways
- New Meta Data

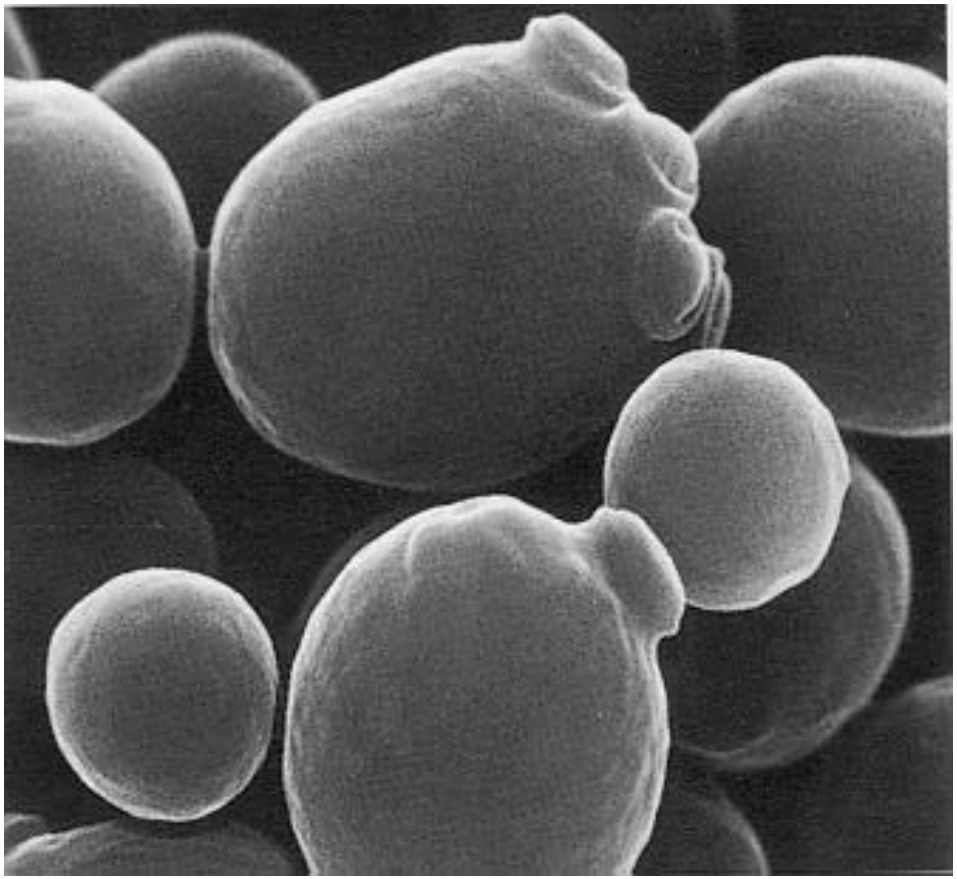


Agilent 6530 QTOF and 1260 series HPLC is a Robust Choice for Metabolomics

- High femtogram-level sensitivity
- Better than 1-ppm MS mass accuracy
- Better than 3-ppm MS/MS mass accuracy;
- Mass resolution (resolving power) of 20,000 -- not dependent on spectral acquisition rate
- Fast data acquisition (= 10 MS/MS spectra/sec) compatible with UHPLC liquid chromatography
- Broad mass range from m/z 25 to 20,000.

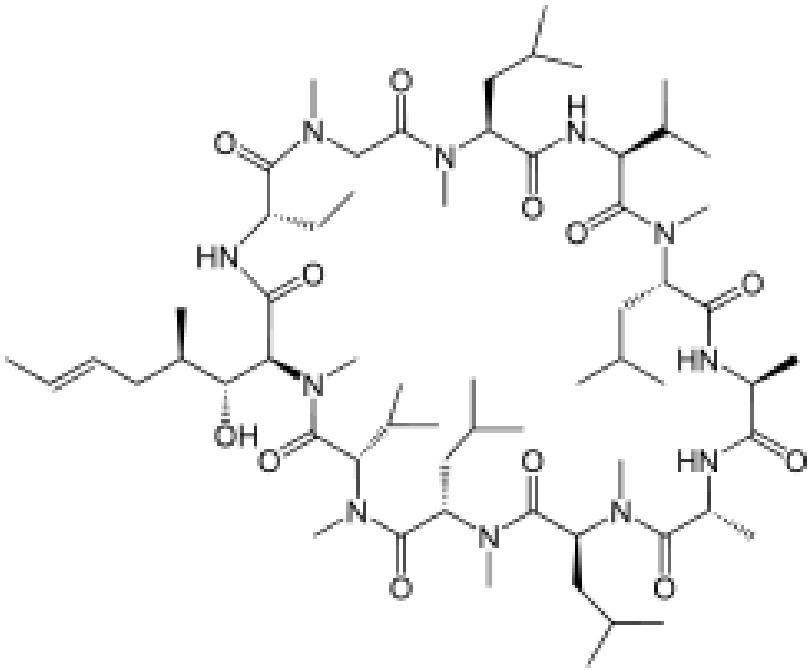
Bakers Yeast is an Ideal Model Organism For Studying Pathways

- *Saccharomyces cerevisiae* is an extensively used model organism.
- Biochemistry and pathways extensively studied.
- Fully sequenced genome.
- Ideal for “Multi-Omics” studies with the goal of facilitating research for other organisms.

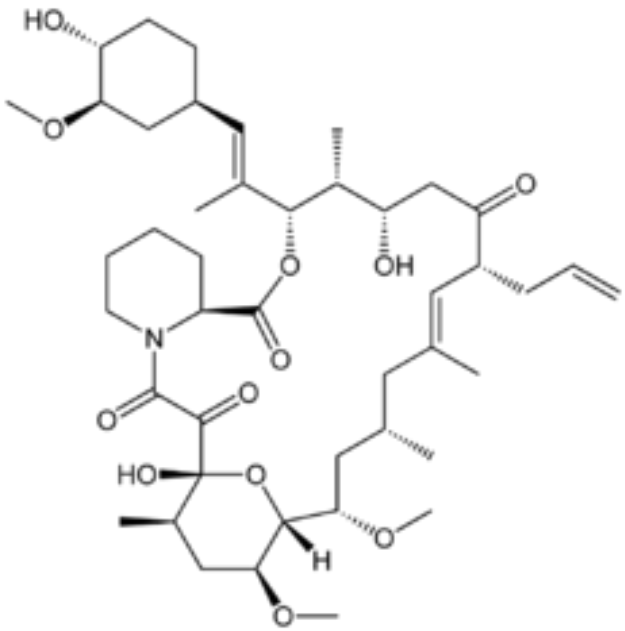


Calcinuerin Inhibitors Were be Used to Study Pathways Related to Immunosuppression

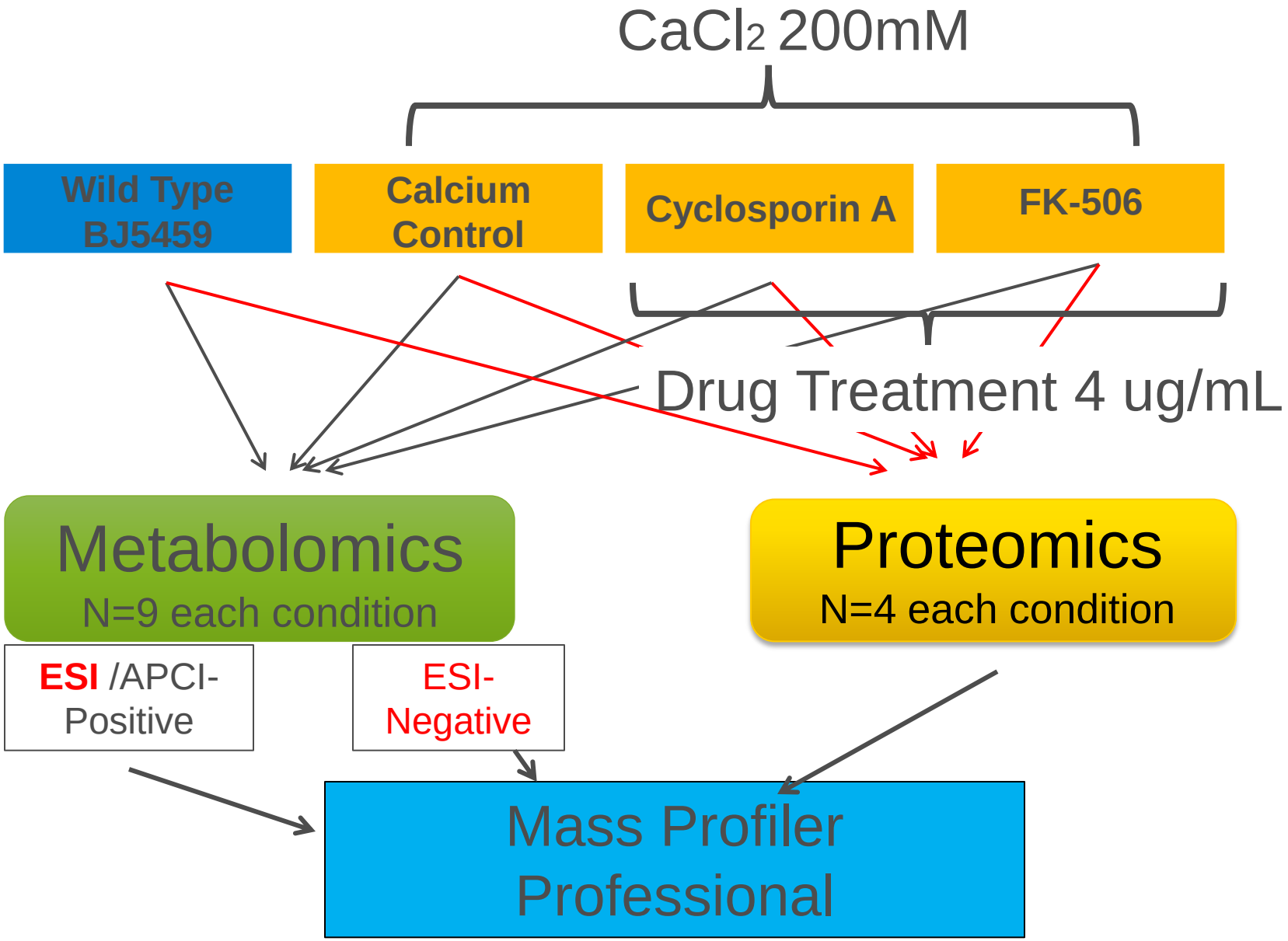
Cyclosporin A



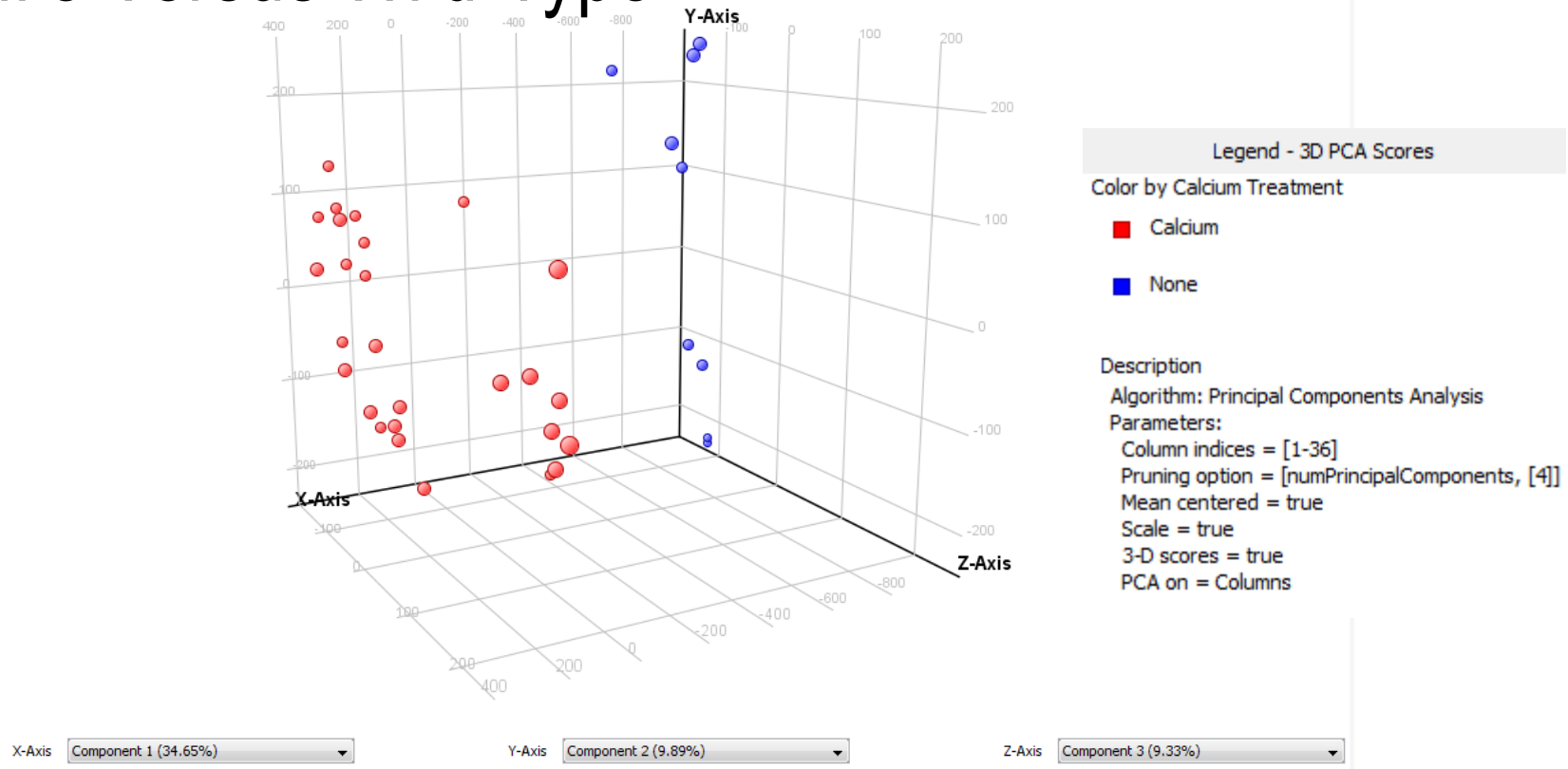
FK-506



Goal: Determine additional metabolites, proteins and pathways affected by the drug treatment


$CaCl_2$ 200mM
Wild Type BJ5459
Calcium Control
Cyclosporin A
FK-506
Drug Treatment 4 ug/mL
Metabolomics
N=9 each condition
Proteomics
N=4 each condition
ESI /APCI-Positive
ESI-Negative
Mass Profiler Professional

Calcium Vector Introduces Changes to the Metabolic Profile Versus Wild-Type



204 Compounds had a P-Value of Less Than 0.1 when comparing Calcium Treated Groups and Wild Type.

An ANOVA Analysis Was Used to Determine Compounds Responding to Drug Treatment

Result Summary

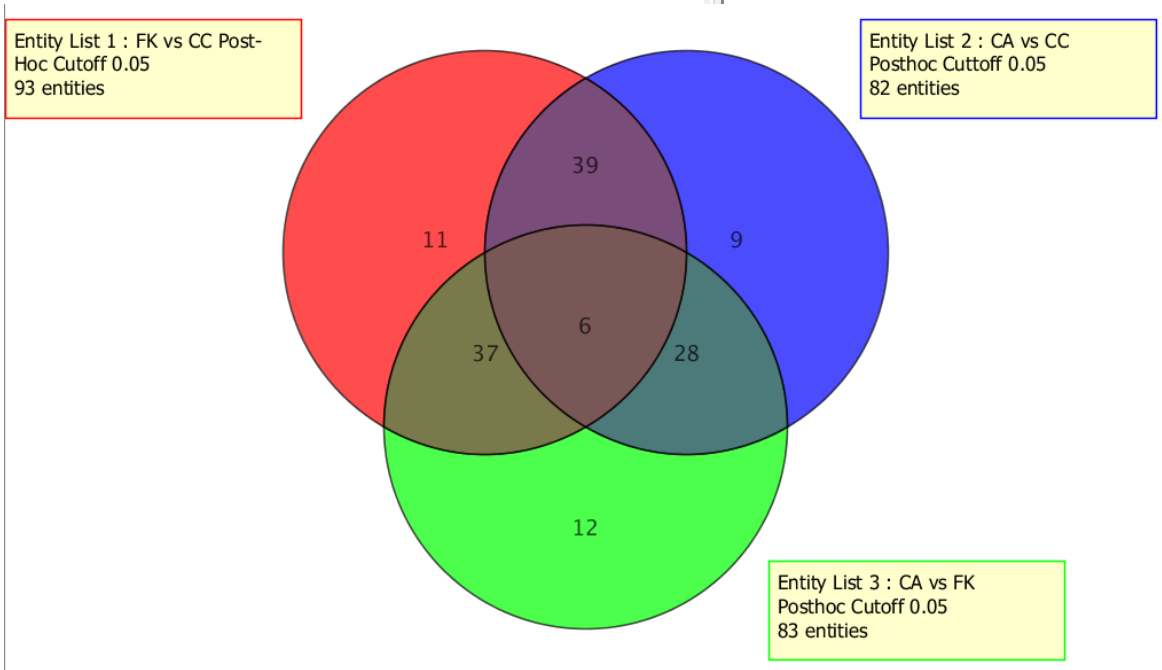
Group Name	[FK Drug]	[Calcium Control]	[Cyclosporin A]
[FK Drug]	142	93	83
[Calcium Control]	49	142	82
[Cyclosporin A]	59	60	142

Blue = Significant Difference in Post-hoc Comparison

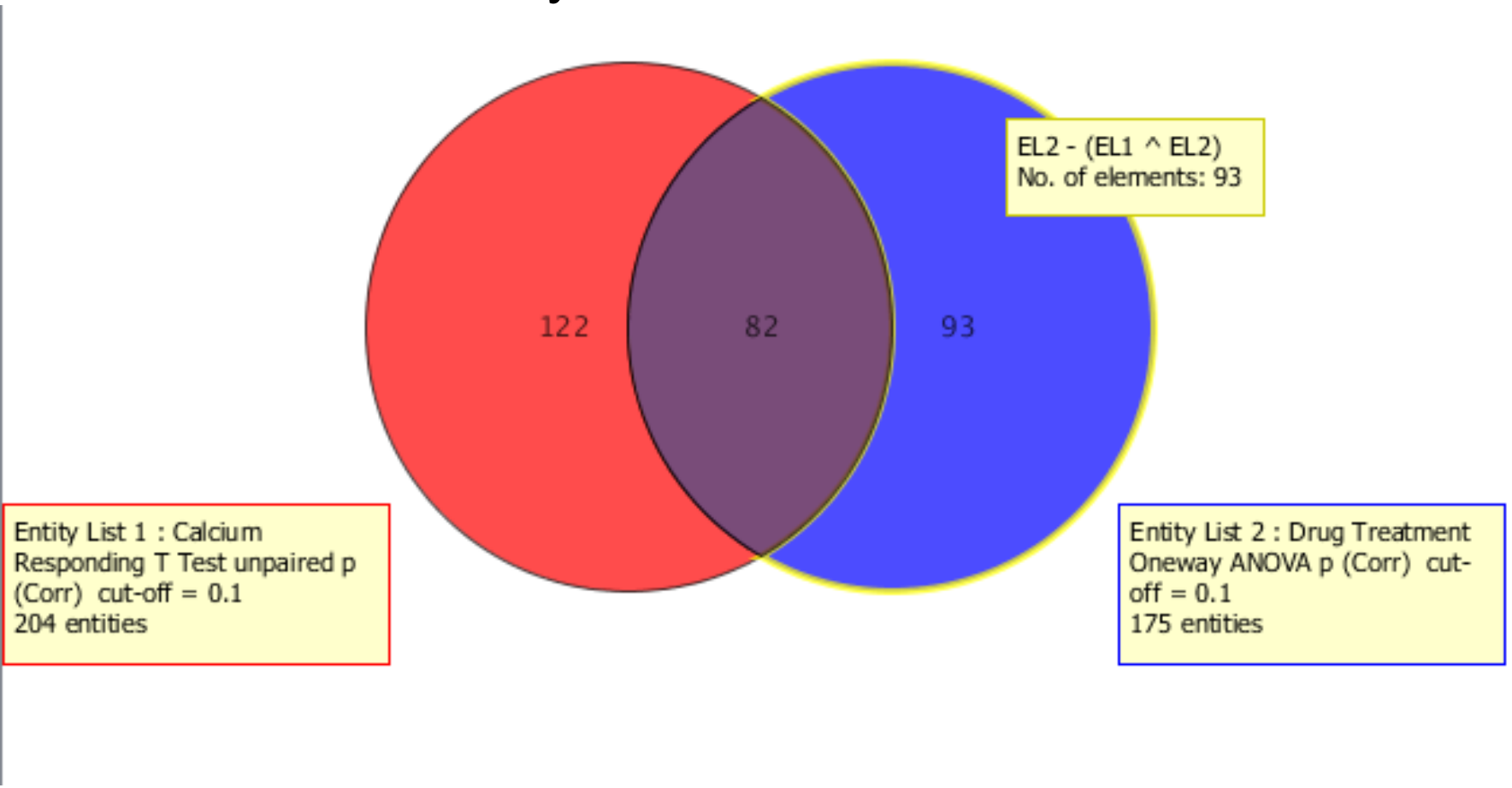
Orange = No Significant Difference in Post-hoc Comparison

Significance Cut-Off= 0.05

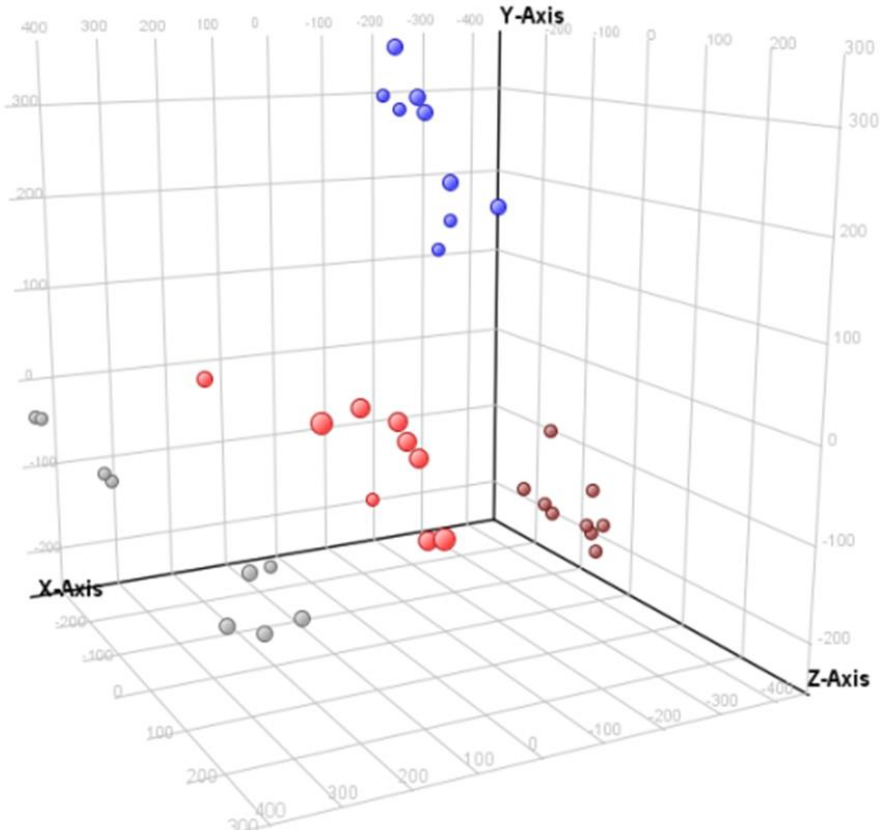
142 Compounds Pass



Venn Diagrams Were Used to Determine 93 Metabolites that Uniquely Respond to Drug Treatment in Positive Polarity



After Removing Calcium Effect We Can See The Effect of Drug Treatments



Responds to Drug Treatment

X-Axis Component 1 (24.75%)

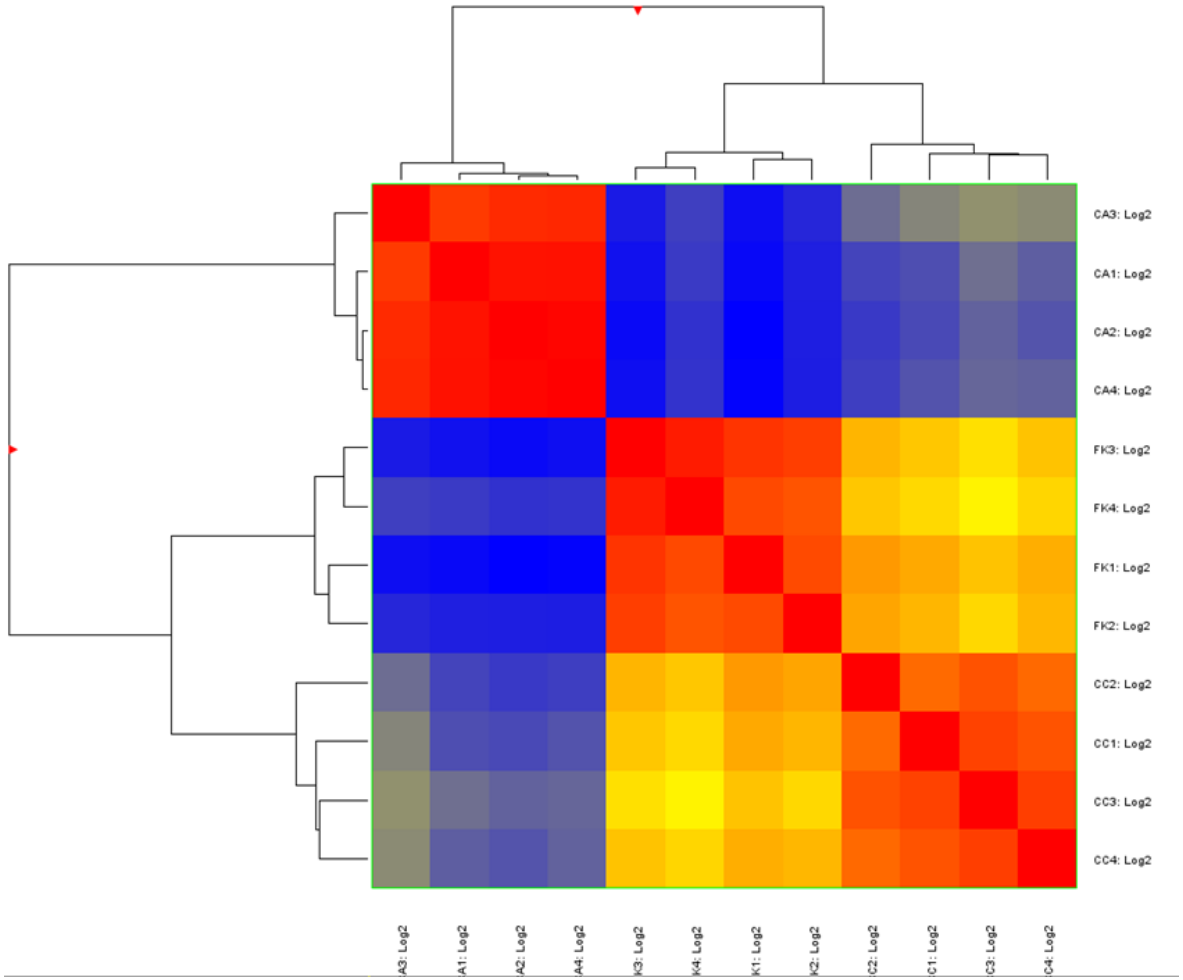
Y-Axis Component 2 (16.48%)

Z-Axis Component 3 (12.73%)

Agilent 6550 and Nano-Flow Chip Cube Bring Enhance Sensitivity for Targeted or Shotgun Proteomics

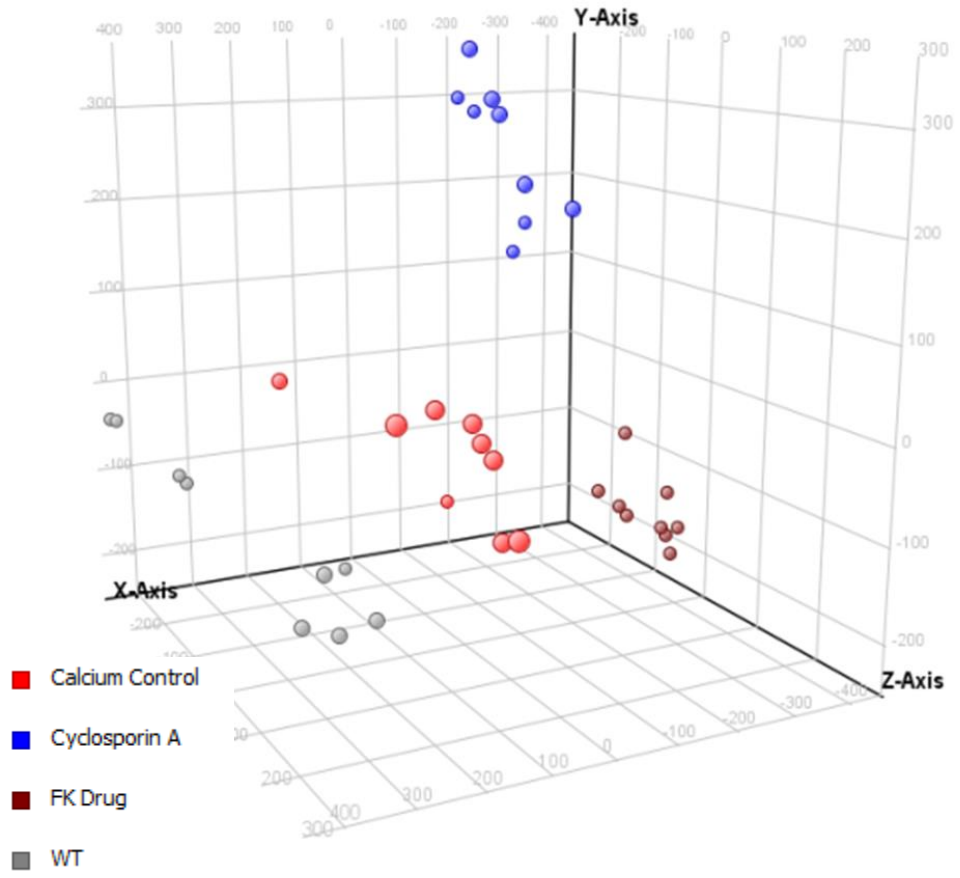
- High attogram to low femtogram sensitivity
- Sub ppm mass accuracy (MS)
- Scan Speeds up to 50 Spectra/s while maintaining 40k resolving power
- 5 orders of magnitude dynamic range
- Low injection volumes and nano-flow for enhanced sensitivity for proteomics applications

New Sample To Sample Correlation Tool Reveals Relationships Between Samples



- Within Groups Proteomics samples correlate r>.90
- FK drug and Cyclosporin A treatment have negative correlations r<-0.2

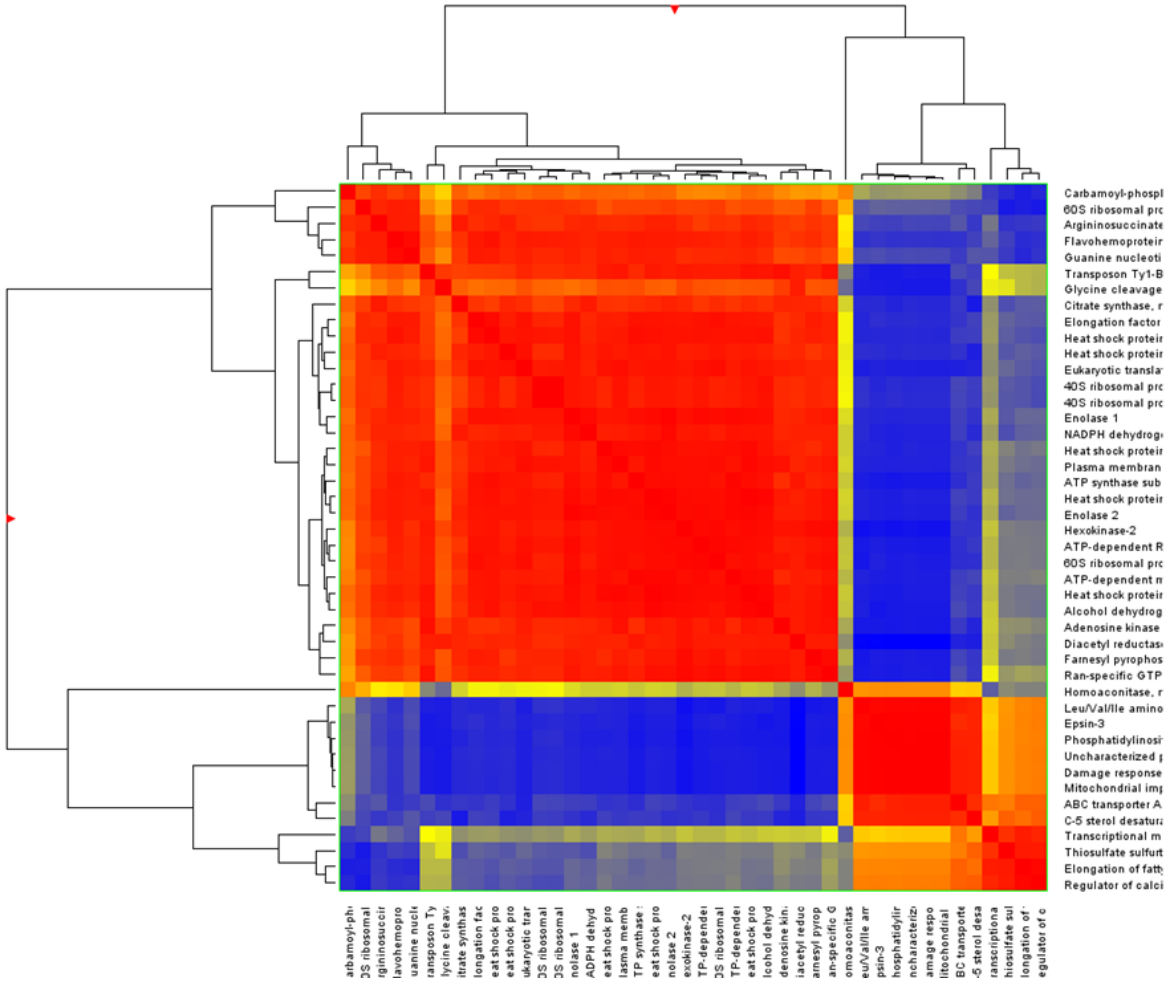
Drug Treatments and Controls Separate Well Using PCA



- Subtracted out the compounds that responded primarily to calcium treatment.
- Focus only on compounds responding to drug treatment.
- Can use correlation-covariance plot or loadings plots to rank compounds important to separation.
- Need to further contextualize the results!

X-Axis Component 1 (24.75%) Y-Axis Component 2 (16.48%) Z-Axis Component 3 (12.73%)

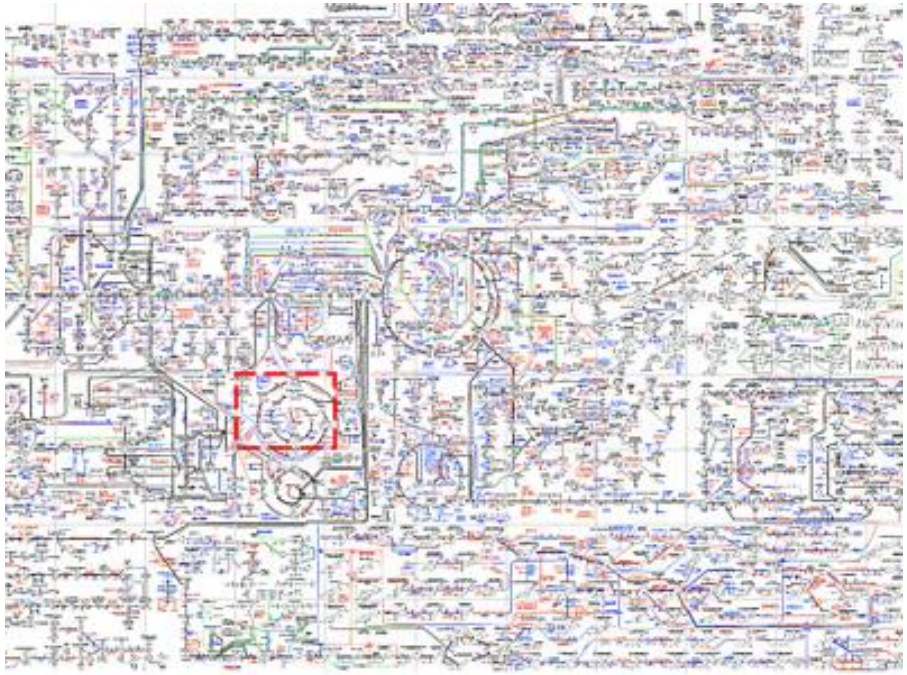
Protein-Protein Correlation Can Be Filtered to Reveal the Most Significant Correlations



- To reduce data complexity correlation map filtered to Fold Change 1.5 and p<0.01
- Provides opportunity to interrogate pathways that co-regulate!

Pathway Architect 13:
Canonical Pathway Data Mapping and Visualization

Central Carbon Metabolism



Browse, filter, and search

Analyze one or two types of –omic data

Supports biological pathways from publicly available databases

- WikiPathways
- BioCyc

Supported formats

- BioPAX 3 – Pathway Commons, Reactome, NCI Nature Pathway
- GPML – PathVisio –custom drawing
- KEGG

Export compound list from pathways

Easy Mining of Complex Pathways for Biological Understanding

Agilent-BridgeDB: Enhanced Metabolite and Protein Mapping

Metabolites Identifiers – more coverage

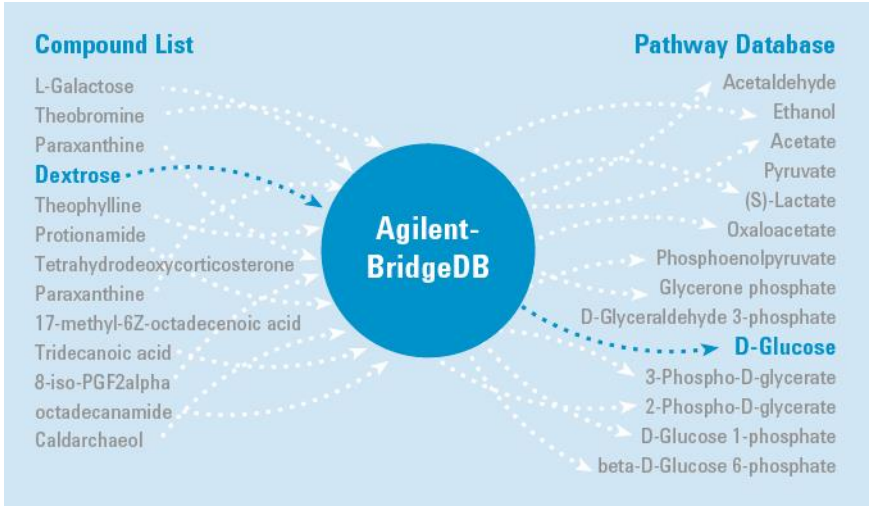
- KEGG
- MetaCyc
- PubChem
- LMP
- HMDB
- ChEBI
- CAS

Proteins Identifiers:

- Swiss-Prot
- UniProt
- UniProt/TrEMBL

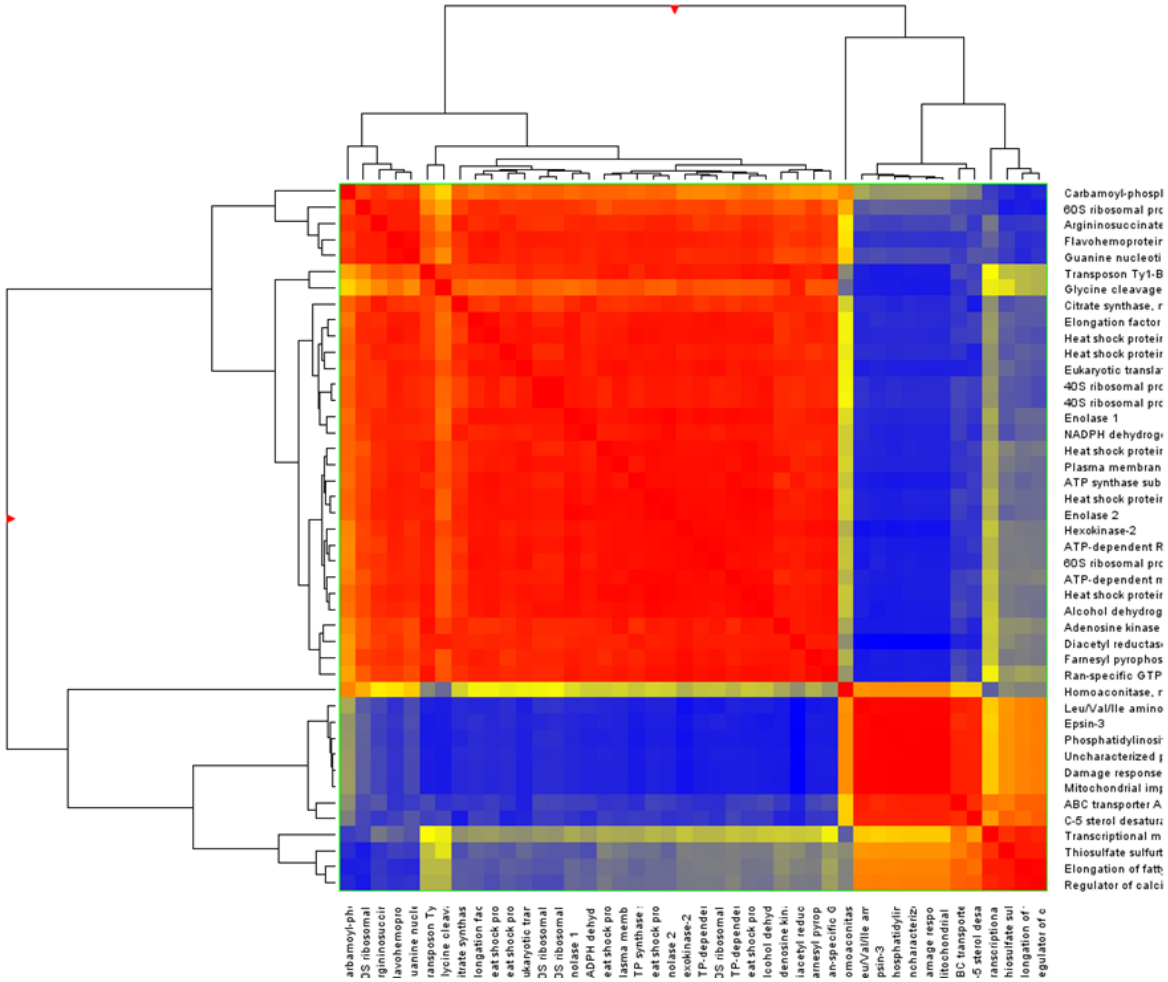
Genes Identifiers :

- Entrez Gene, GenBank, Ensembl
- EC Number, RefSeq, UniGene, HUGO
- HGNC, EMBL



Resolve the Mapping Problem Between Databases

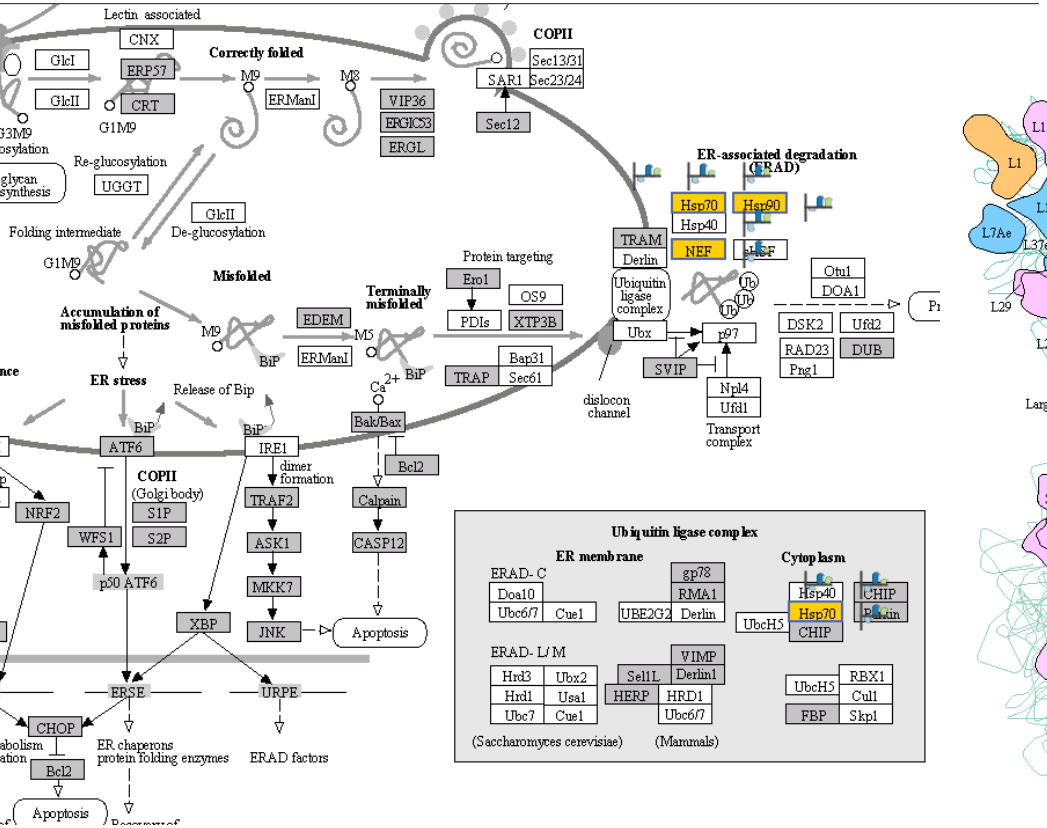
Protein-Protein Correlation Can Be Filtered to Reveal the Most Significant Correlations

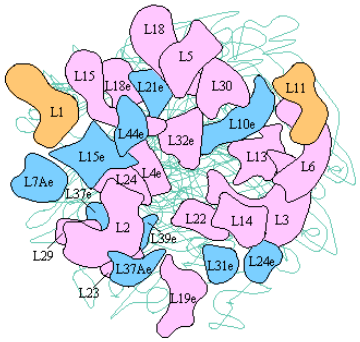


- To reduce data complexity correlation map filtered to Fold Change 1.5 and p<0.01
- Provides opportunity to interrogate pathways that co-regulate!

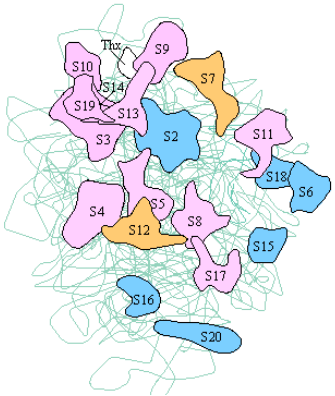
Searching Most Significant Correlations with KEGG Pathways in MPP 13 Reveal Changes in Protein Metabolism with Cyclosporin A treatment

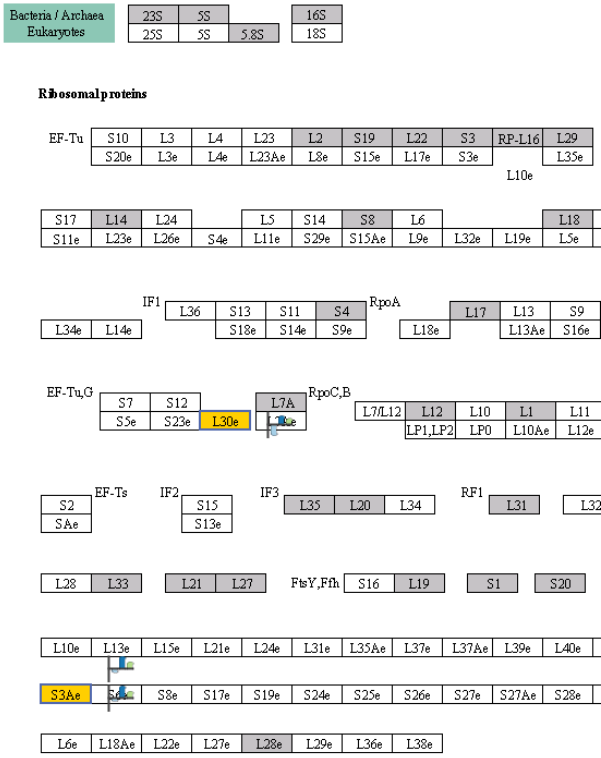
[Calcium, Cyclosporin A]
[Calcium, FK Drug]
[Calcium, None]

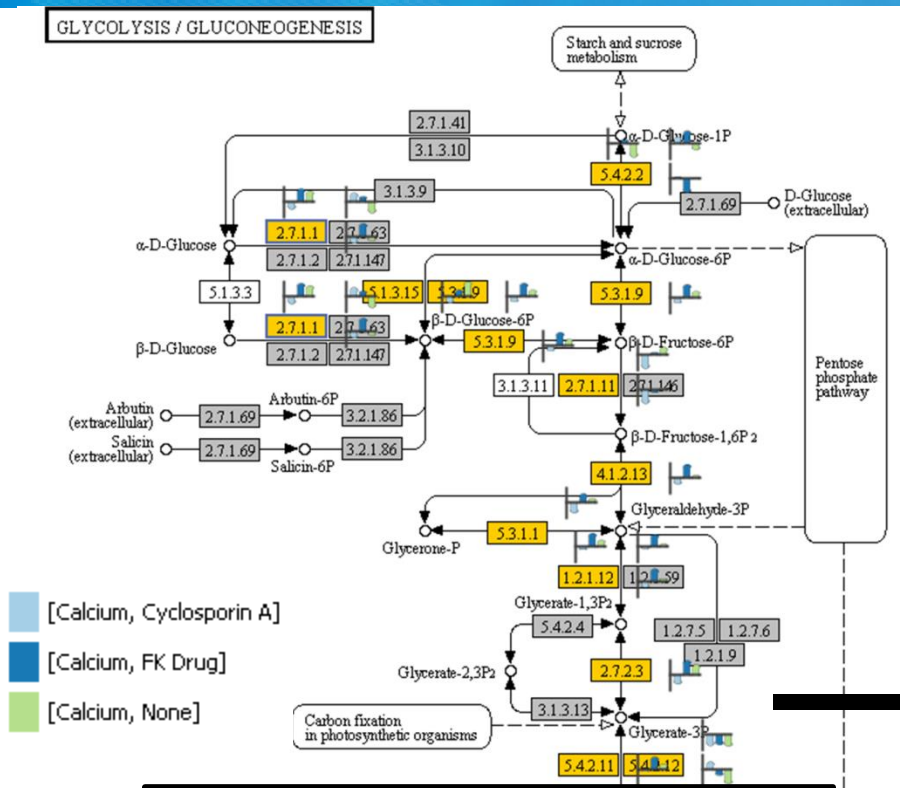


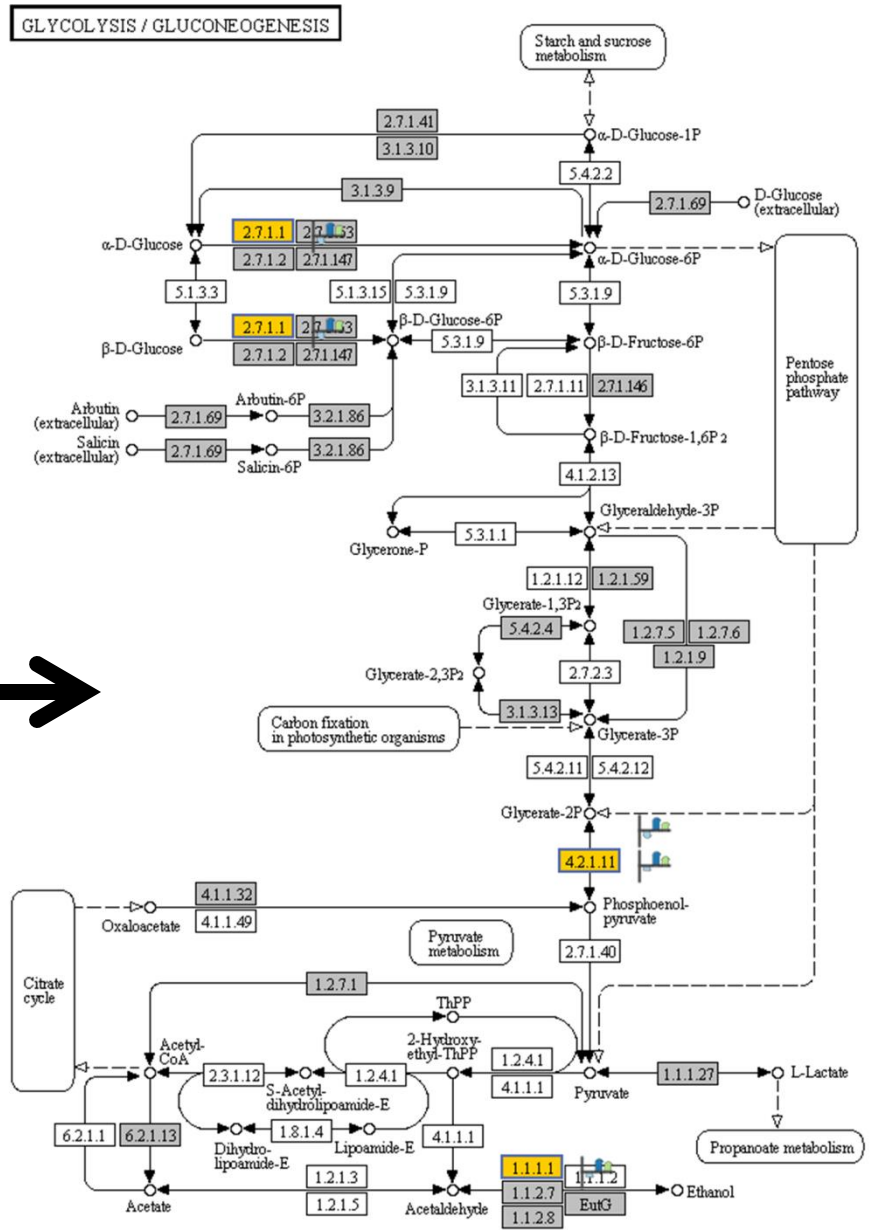


Large subunit (Haloarcula marismortui)





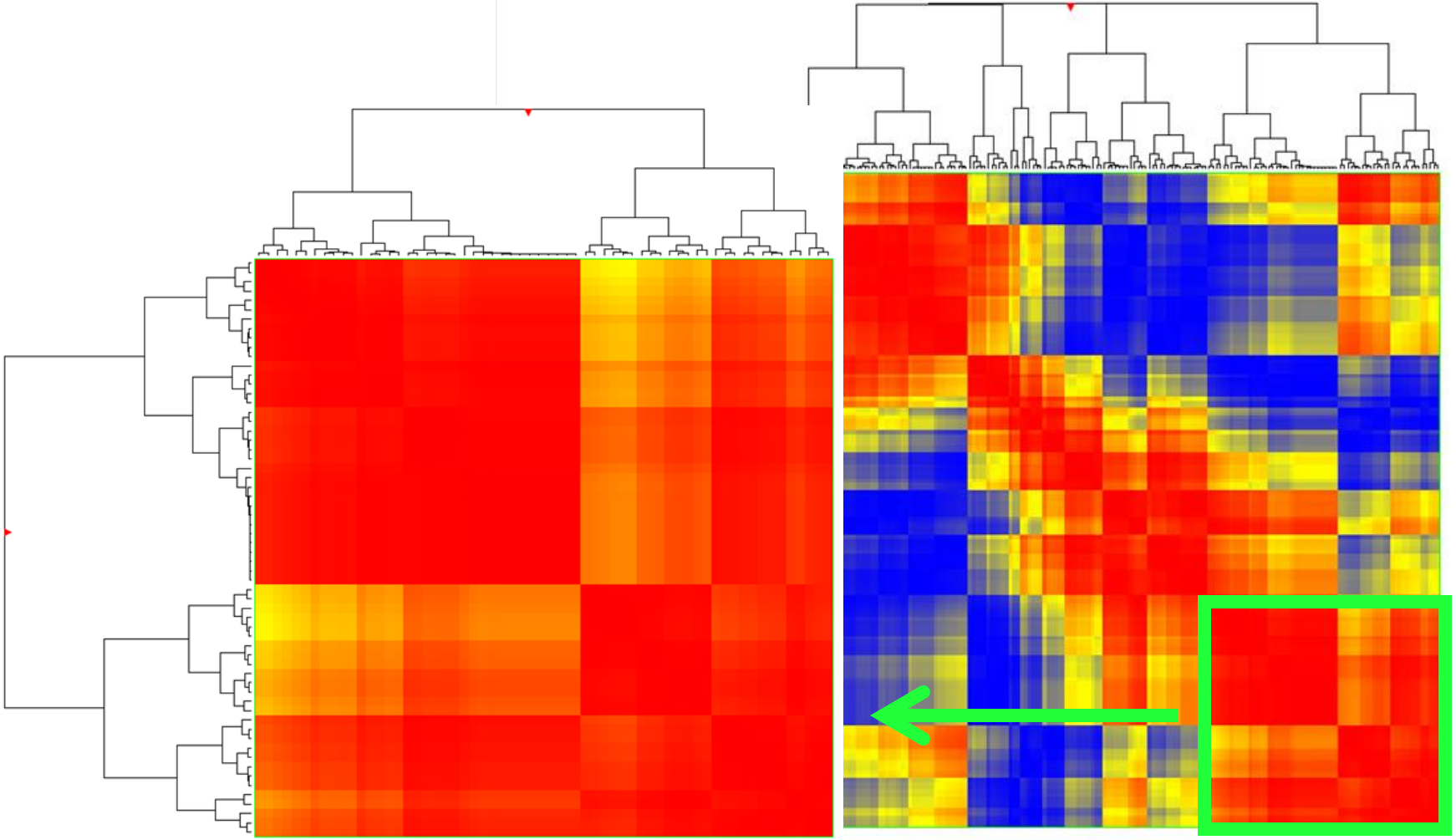




With correlation tools we can easily ask questions about co-regulation and thus make results interpretation easier.

Using correlation can easily ask which pathway components co-regulate

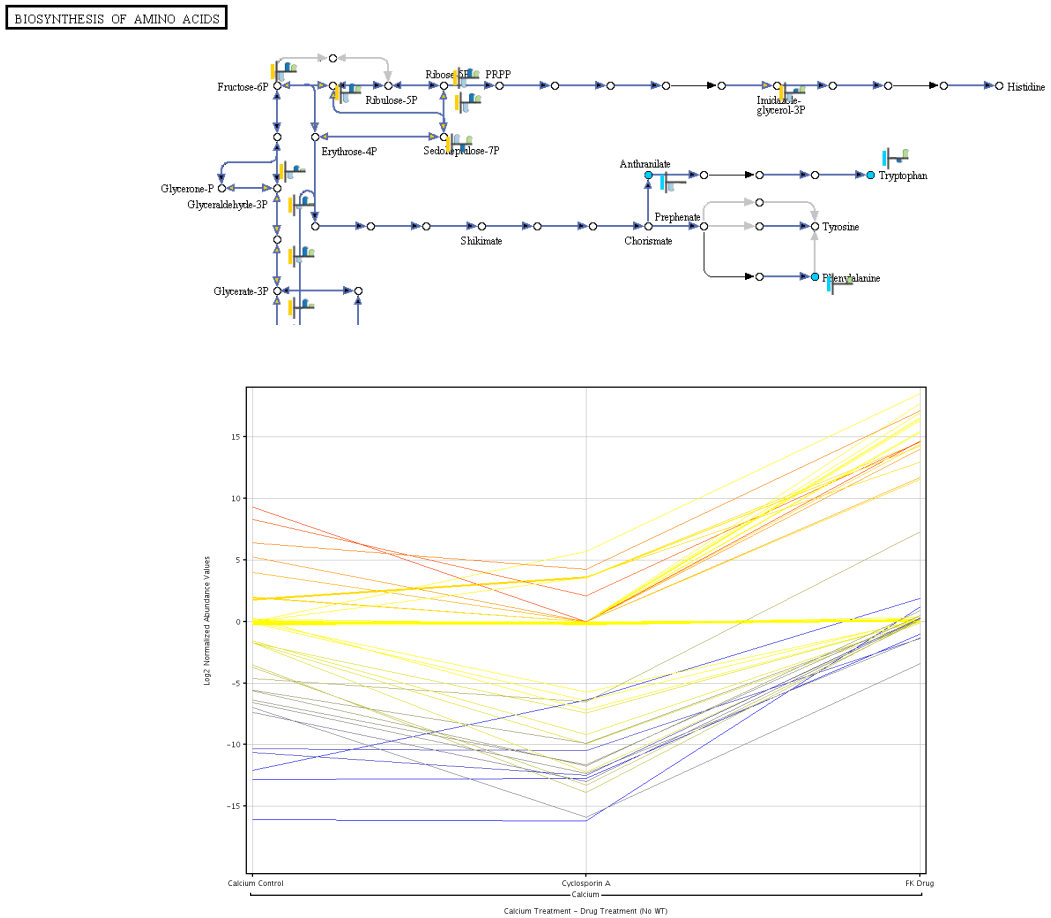
Metabolite-Metabolite Correlation Adds More Depth of Understanding to the FK-Drug Treatment



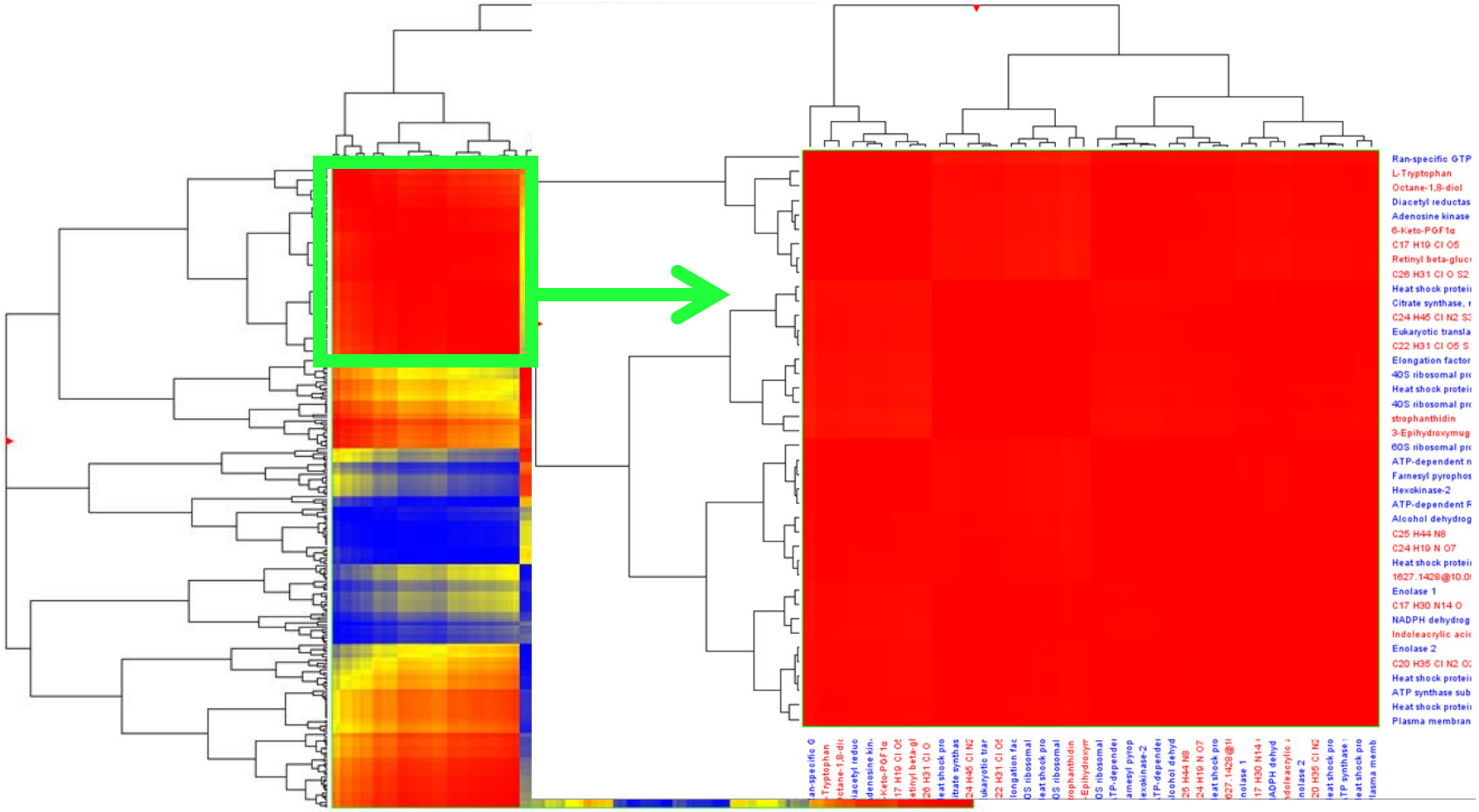
Metabolite Correlations Finds Shifts in Amino Acid Metabolism for FK-Drug Treatment

- Aromatic Amino Acid, Ser-Thr-Gly Amino Acid Biosynthesis Pathways are up-regulated with in FK-Drug

- Purine and NAD Biosynthesis also up-regulated



Multi-”omic” Correlation Can Be Used To Ask About Patterns Of Covariance Between Proteins And Metabolites



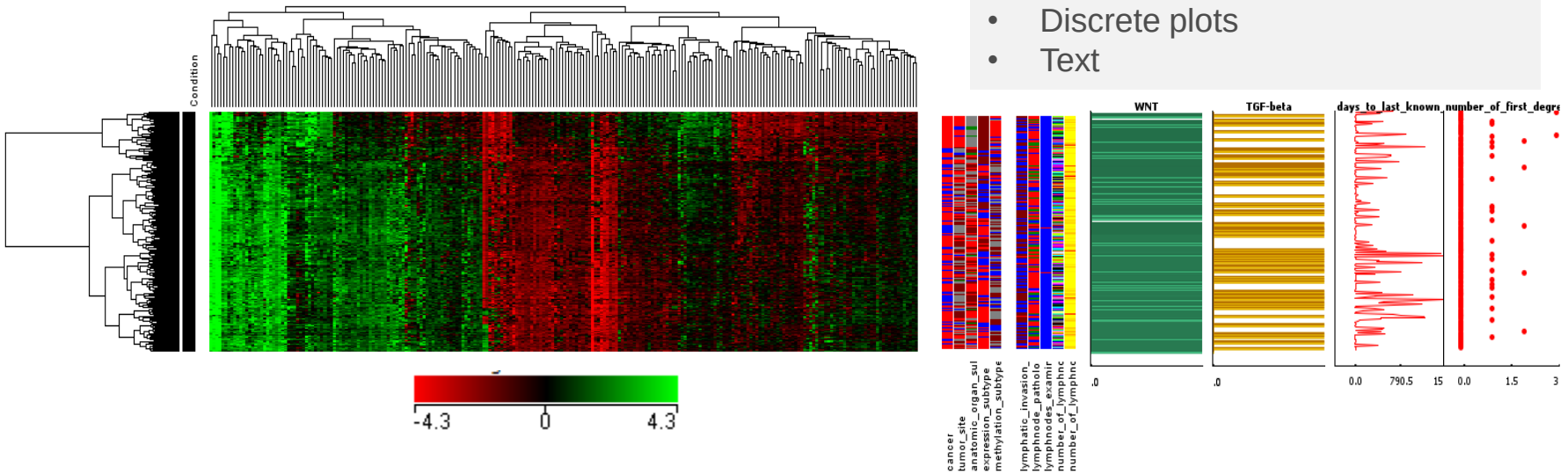
What’s New in MPP 13.0 and Profinder

MPP 13 Supports Meta Data

Analytical Results to Meta Data

Metadata can be displayed as

- Heat maps
- Colored strips
- Graphical plots
- Discrete plots
- Text

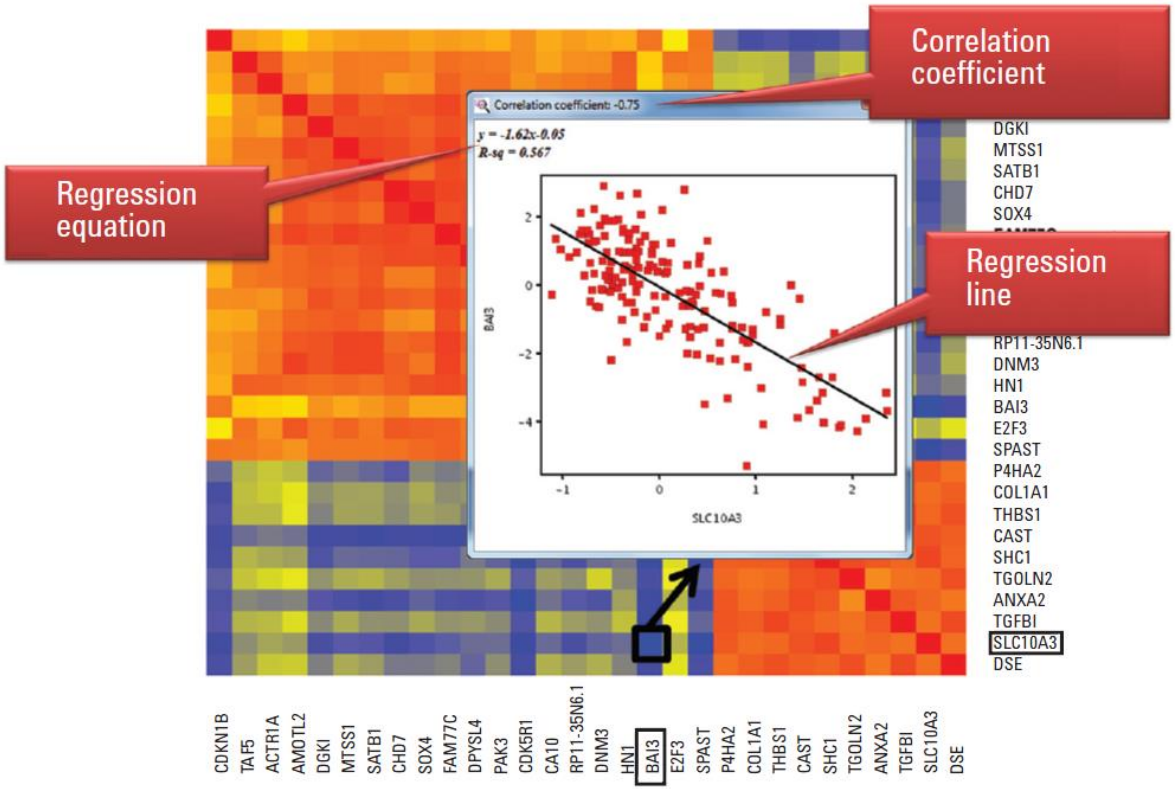


Visualize the significant relationships!

- Maps observable sample information to analytical results
- Provides flexible visualization of metadata (e.g. time points, growth conditions, patient data etc.) to facilitate interpretation

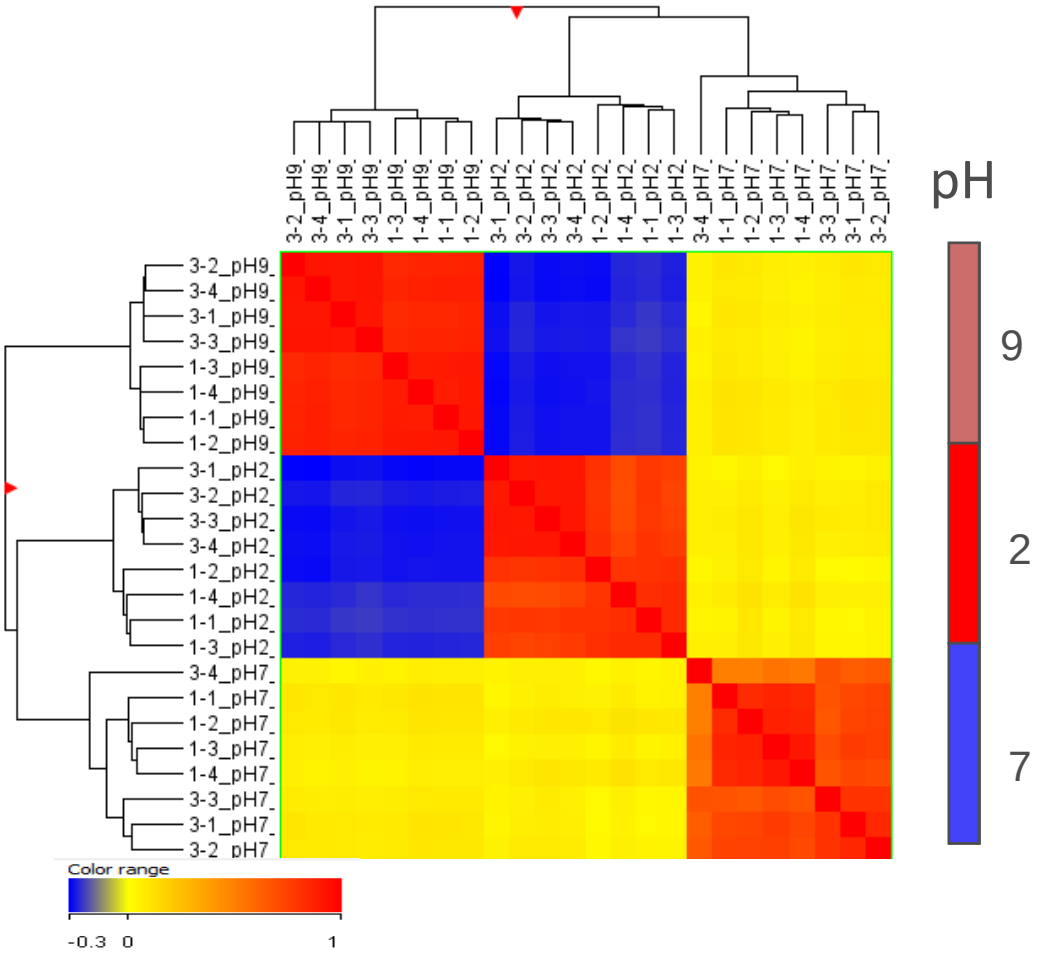
Multi-Omic Correlation Analysis

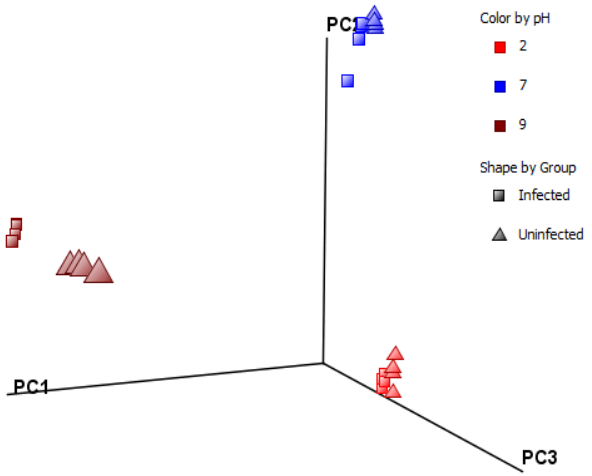
MPP 13 adds support for Correlation Analysis. Researchers can view relationships between entities (compounds or proteins) or samples. Clicking on a cell of a heat map, they can quickly view the specific parameters of the correlation



MPP Correlation Analysis

Sample to Sample





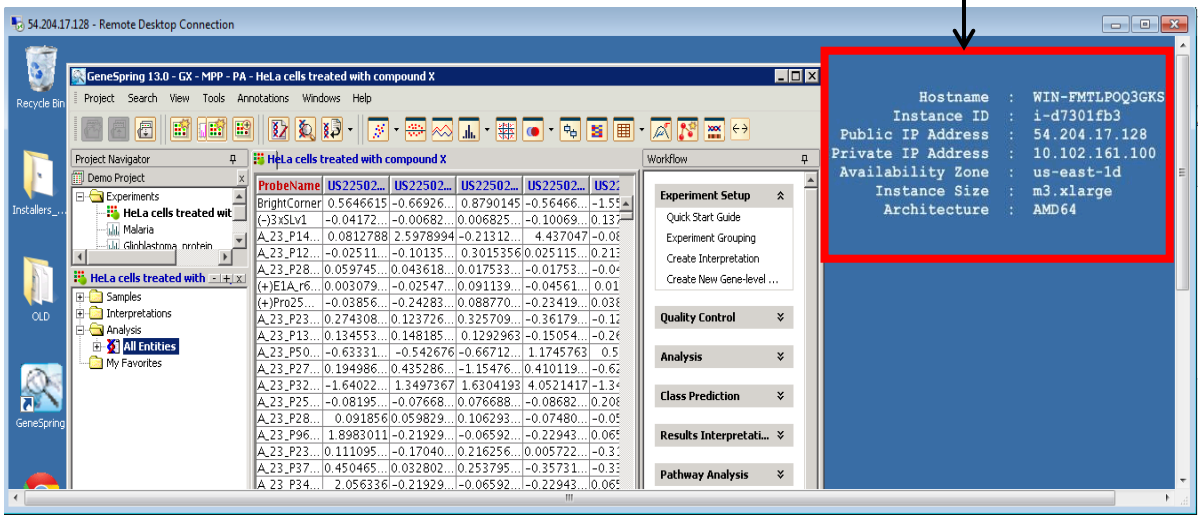
Sample correlation Heatmap indicates sample relationships both *within a particular pH and between samples of different pH*

MPP Has Increased performance and New Cloud Deployment!

GeneSpring MPP was internally and beta tested using a cloud configuration including Amazon Web Services (AWS)

Server specifications (e.g. CPU), many of which are configurable on request

An upcoming technical overview will discuss how to configure a cloud or virtual machine deployment and advantages like flexibility and collaboration

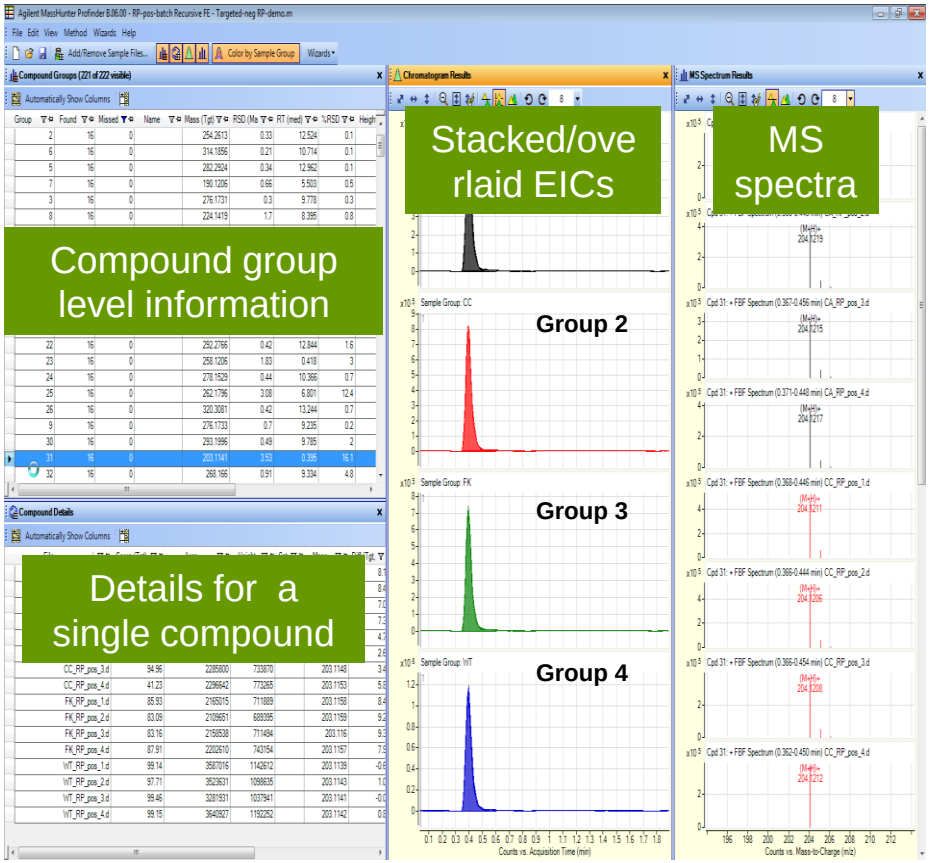


GeneSpring MPP running on Amazon Web Services (AWS)

MassHunter Profinder Software

MassHunter Profinder is a productivity tool for processing multiple samples in metabolomics, proteomics, intact protein analyses

- **Fast Compound Finding**
  - Untargeted using MFE
  - Targeted using Find by Formula
- **Visualize, review, and edit results by compound across many samples**
- **Higher quality results based on cross-sample processing**
- **Minimizes false positive and negative results**
- **Batch Processing**



Profinder B.06.00 SP1
Is Faster and Now Support for Intact Proteins

Profinder B.06.00 SP1 has added the Large Molecular Feature Extraction (LMFE) algorithm, which enables profiling of intact proteins using multiply charged mass spec data. Now small molecule, peptide, and intact proteins can be processed in the same program!