
Additional file 6 – Detailed Metabolomics Methodology

Metabolon HD4 Platform

Sample Accessioning: Following receipt, samples are inventoried and immediately stored at -80°C. Each sample received is accessioned into the Metabolon LIMS system and assigned a unique identifier that is associated with the original source identifier only. This identifier is used by the LIMS system to track all samples (and all derived aliquots), sample handling, tasks, results, *etc.* All portions of any sample are automatically assigned their own unique identifiers by the LIMS when a new task is created; the relationship of these samples is also tracked. All samples are maintained at -80°C until processed.

Sample Preparation: Samples are prepared using the automated MicroLab STAR® system from Hamilton Company. Several recovery standards are added prior to the first step in the extraction process for QC purposes. Samples are extracted with methanol under vigorous shaking for 2 min (Glen Mills GenoGrinder 2000) to precipitate protein and dissociate small molecules bound to protein or trapped in the precipitated protein matrix, followed by centrifugation to recover chemically diverse metabolites. The resulting extract is divided into five fractions: two for analysis by two separate reverse phase (RP)/UPLC-MS/MS methods using positive ion mode electrospray ionization (ESI), one for analysis by RP/UPLC-MS/MS using negative ion mode ESI, one for analysis by HILIC/UPLC-MS/MS using negative ion mode ESI, and one reserved for backup. Samples are placed briefly on a TurboVap® (Zymark) to remove the organic solvent. The sample extracts are stored overnight under nitrogen before preparation for analysis.

QA/QC: Several types of quality control samples are analyzed in concert with the experimental samples. These include: 1) technical replicate samples derived from a pool of well-characterized human plasma (MTRX) or, alternatively, generated by combining a small portion of each (non-plasma) experimental sample (CMTRX), spaced evenly among experimental samples; 2) extracted water samples (process blanks) and solvent blanks; and 3) a cocktail of QC standards, carefully chosen not to interfere with the measurement of endogenous compounds, spiked into every analyzed sample, allowing instrument performance monitoring and aiding with chromatographic alignment. Tables 1 and 2 describe these QC samples and standards. Instrument variability is determined by calculating the median relative standard deviation (RSD) for the standards that were added to each sample prior to injection into the mass spectrometers. Overall process variability is determined by calculating the median RSD for all endogenous metabolites (*i.e.*, non-instrument standards) present in each of the pooled MTRX (or CMTRX) technical replicate samples. Experimental samples are randomized across the platform run, with QC samples spaced evenly among the injections, as outlined in Figure 1.

Table 1: Description of Metabolon QC Samples

| Type | Description | Purpose |
|-------|--|--|
| MTRX | Large pool of human plasma maintained by Metabolon that has been characterized extensively | Assure that all aspects of the Metabolon process are operating within specifications |
| CMTRX | Pool created by taking a small aliquot from every client sample | Assure that all aspects of the Metabolon process for non-plasma matrices are operating within specifications |
| PRCS | Aliquot of ultra-pure water that provides a baseline reference signal | Process Blank used to assess process background contribution to compound signals |
| SOLV | Aliquot of solvents used in extraction | Solvent Blank used to segregate contamination sources in the extraction |

Table 2: Metabolon QC Standards

| Type | Description | Purpose |
|------|-------------------|--|
| RS | Recovery Standard | Assess variability and verify performance of extraction and instrumentation. |
| IS | Internal Standard | Assess variability and performance of instrument. |

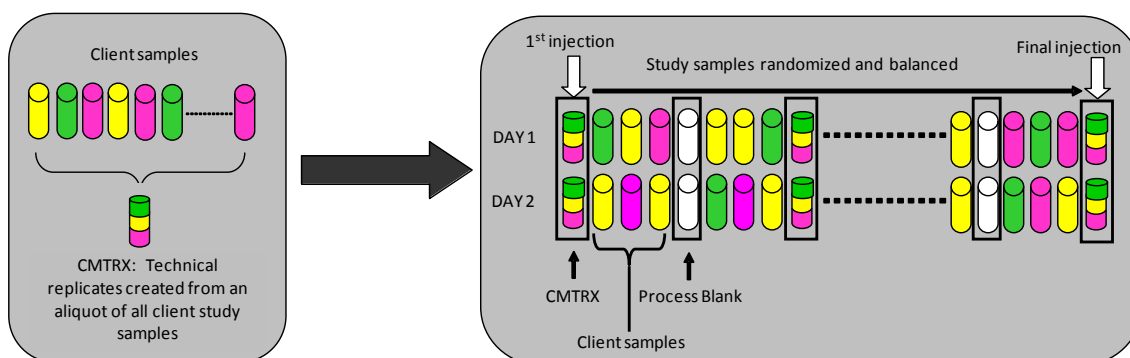


Figure 1. Preparation of client-specific technical replicates. A small aliquot of each client sample (colored cylinders) is pooled to create a CMTRX technical replicate sample (multi-colored cylinder), which is then injected periodically throughout the platform run. Variability among consistently detected biochemicals can be used to calculate an estimate of overall process and platform variability.

Ultrahigh Performance Liquid Chromatography-Tandem Mass Spectroscopy (UPLC-MS/MS): All methods utilize a Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer operated at 35,000 mass resolution. The sample extract is dried then reconstituted in solvents compatible to each of the four methods. Each reconstitution solvent contains a series of standards at fixed concentrations to ensure injection and chromatographic consistency. One aliquot is

analyzed using acidic positive ion conditions, chromatographically optimized for more hydrophilic compounds. In this method, the extract is gradient-eluted from a C18 column (Waters UPLC BEH C18-2.1x100 mm, 1.7 μ m) using water and methanol, containing 0.05% perfluoropentanoic acid (PFPA) and 0.1% formic acid (FA). A second aliquot is also analyzed using acidic positive ion conditions, but is chromatographically optimized for more hydrophobic compounds. In this method, the extract is gradient eluted from the aforementioned C18 column using methanol, acetonitrile, water, 0.05% PFPA and 0.01% FA, and is operated at an overall higher organic content. A third aliquot is analyzed using basic negative ion optimized conditions using a separate dedicated C18 column. The basic extracts are gradient-eluted from the column using methanol and water, however with 6.5mM Ammonium Bicarbonate at pH 8. The fourth aliquot is analyzed via negative ionization following elution from a HILIC column (Waters UPLC BEH Amide 2.1x150 mm, 1.7 μ m) using a gradient consisting of water and acetonitrile with 10mM Ammonium Formate, pH 10.8. The MS analysis alternates between MS and data-dependent MSⁿ scans using dynamic exclusion. The scan range varies slightly between methods, but covers approximately 70-1000 m/z. Raw data files are archived and extracted as described below.

Bioinformatics: The informatics system consists of four major components, the Laboratory Information Management System (LIMS), the data extraction and peak-identification software, data processing tools for QC and compound identification, and a collection of statistical, visualization, and interpretation tools for use by data analysts. The hardware and software foundations for these informatics components are the LAN backbone and database servers running Oracle 10.2.0.1 Enterprise Edition.

LIMS: The purpose of the Metabolon LIMS system is to enable fully auditable laboratory automation through a secure, easy to use, and highly specialized system. The scope of the Metabolon LIMS system encompasses sample accessioning, sample preparation, instrumental analysis and reporting, and advanced data analysis. All the subsequent software systems are grounded in the LIMS data structures, which has been modified to leverage and interface with the in-house information extraction and data visualization systems, as well as third party instrumentation and data analysis software.

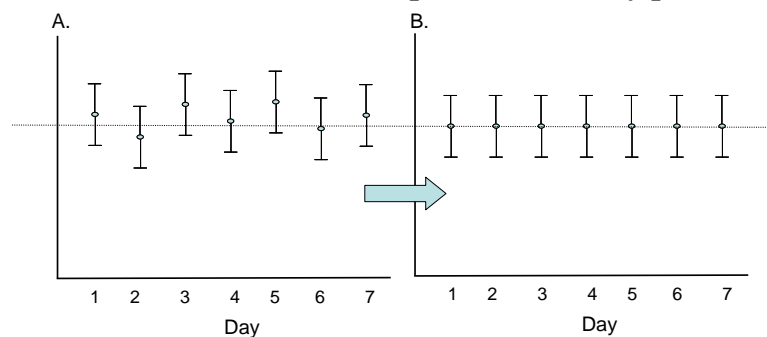
Data Extraction and Compound Identification: Raw data are extracted, peak-identified, and QC processed using Metabolon's hardware and software. These systems are built on a web-service platform utilizing Microsoft's .NET technologies, which run on high-performance application servers and fiber-channel storage arrays in clusters to provide active failover and load-balancing. Compounds are identified by comparison to library entries of purified standards or recurrent unknown entities. Metabolon maintains a library based on authenticated standards that contains the retention time/index (RI), mass to charge ratio (m/z), and chromatographic data (including MS/MS spectral data) on all molecules present in the library. Furthermore, biochemical identifications are based on three criteria: retention index within a narrow RI window of the proposed identification, accurate mass match to the library +/- 10 ppm, and the MS/MS forward and reverse scores. MS/MS scores are based on a comparison of the ions present in the experimental spectrum to ions present in the library entry spectrum. While there may be similarities between these molecules based on one of these factors, the use of all three data points can be utilized to distinguish and differentiate biochemicals. More than 4500 commercially available purified standard compounds have been acquired and registered into LIMS for analysis on all platforms for determination of their analytical characteristics. Additional mass spectral

entries have been created for structurally unnamed biochemicals, which have been identified by virtue of their recurrent nature (both chromatographic and mass spectral). These compounds have the potential to be identified by future acquisition of a matching purified standard or by classical structural analysis.

Curation: A variety of curation procedures are performed to ensure that a high quality data set is made available for statistical analysis and data interpretation. The QC and curation processes are designed to ensure accurate and consistent identification of true chemical entities, and to remove those representing system artifacts, mis-assignments, redundancy, and background noise. Metabolon data analysts use internally-developed visualization and interpretation software to confirm the consistency of peak identification among the various samples. Library matches for each compound are checked for each sample and corrected if necessary.

Metabolite Quantification and Block Correction: Peaks are quantified as area-under-the-curve detector ion counts. For studies spanning multiple days, a data adjustment step is performed to correct block variation resulting from instrument inter-day tuning differences, while preserving intra-day variance. Essentially, each compound is corrected in balanced run-day blocks by registering the daily medians to equal one (1.00), and adjusting each data point proportionately (termed the “block correction”; Figure 2). For studies that do not require more than one day of analysis, no adjustment of raw data is necessary, other than scaling for purposes of data visualization.

Figure 2: Visualization of data correction steps for a multiday platform run.



Normalization: In certain instances, biochemical data may be normalized to an additional factor, such as cell counts, total protein (Bradford assay), osmolality, *etc.*, to account for systematic variation in metabolite levels due to differences in the amount of starting material in each sample.

Statistical Methods and Terminology

Statistical Calculations: For many studies, two types of statistical analysis are usually performed: (1) significance tests and (2) classification analysis. Standard statistical analyses are performed in ArrayStudio on log transformed data. For those analyses not standard in ArrayStudio, the programs R (<http://cran.r-project.org/>) or JMP are used. Below are examples of frequently employed significance tests and classification methods followed by a discussion of p- and q-value significance thresholds.

1. Welch's two-sample t-test

Welch's two-sample t -test is used to test whether two unknown means are different from two independent populations.

This version of the two-sample t -test allows for unequal variances (variance is the square of the standard deviation) and has an *approximate* t -distribution with degrees of freedom estimated using Satterthwaite's approximation. The test statistic is given by $t = (\bar{x}_1 - \bar{x}_2) / \sqrt{s_1^2/n_1 + s_2^2/n_2}$, and the degrees of freedom is given by $\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2 / \left(\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 - 1} \right)$, where \bar{x}_1, \bar{x}_2 are the sample means, s_1, s_2 , are the sample standard deviations,

and n_1, n_2 are the samples sizes from groups 1 and 2, respectively. We typically use a two-sided test (tests whether the means are different) as opposed to a one-sided test (tests whether one mean is greater than the other).

2. Matched pairs t-test

The matched pairs t -test is used to test whether two unknown means are different from paired observations taken on the same subjects.

The matched pairs t -test is equivalent to the one-sample t -test performed on the differences of the observations taken on each subject (i.e., calculate $(x_1 - x_2)$ for each subject; test whether the mean difference is zero or not). The test statistic is given by $t = (\bar{x}_1 - \bar{x}_2) / s_d / \sqrt{n}$, with $n - 1$ degrees of freedom, where \bar{x}_1, \bar{x}_2 are the sample means for groups 1 and 2, respectively, s_d is the standard deviation of the differences, n is the number of *subjects* (so there are $2n$ observations).

3. One-way ANOVA

ANOVA stands for analysis of variance. For ANOVA, it is assumed that all populations have the same variances. One-way ANOVA is used to test whether at least two unknown means are all equal or whether at least one pair of means is different. For the case of two means, ANOVA gives the same result as a two-sided t -test with a pooled estimate of the variance.

An ANOVA uses an F-test which has two parameters – the numerator degrees of freedom and the denominator degrees of freedom. The degrees of freedom in the numerator are equal to $g - 1$, where g is the number of groups. If n is the total number of observations ($n_1 + n_2$), then, the denominator degrees of freedom is equal to $n - g$. The F-statistic is the ratio of the between-groups variance to the within-groups variance, hence the higher the F-statistic the more evidence we have that the means are different.

Often within ANOVA, one performs linear contrasts for specific comparisons of interest. For example, suppose we have three groups A, B, C, then examples of some contrasts are A vs. B, the average of A and B vs. C, etc. For single-degree of freedom contrasts, these give the same result as a two-sided t -test with the pooled estimate of the variance from the ANOVA and degrees of freedom $n - g$. Below, we show the three formulas for A vs. B from a three group design as shown above. The numerator is same in each case, but the denominator differs by the estimates of the variances, and the degrees of freedom are different for each (if the theoretical assumptions hold, then the contrast has the most power, as it has the largest degrees of freedom).

Welch's two-sample t -test

By $t = (\bar{x}_A - \bar{x}_B) / \sqrt{s_A^2/n_A + s_B^2/n_B}$, and the degrees of freedom is given by

$$\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2 / \left(\frac{\left(\frac{s_A^2}{n_A}\right)^2}{n_A-1} + \frac{\left(\frac{s_B^2}{n_B}\right)^2}{n_B-1}\right)$$

Two-sample t -test with pooled estimate of variance from A and B

$$t = (\bar{x}_A - \bar{x}_B) / \sqrt{s_{AB}^2 (1/n_A + 1/n_B)}$$

where $s_{AB}^2 = ((n_A - 1)s_A^2 + (n_B - 1)s_B^2) / (n_A + n_B - 2)$, where the degrees of freedom is $n_A + n_B - 2$.

The contrast from the ANOVA,

$$t = (\bar{x}_A - \bar{x}_B) / \sqrt{s^2 (1/n_A + 1/n_B)}$$

where $s^2 = ((n_A - 1)s_A^2 + (n_B - 1)s_B^2 + (n_C - 1)s_C^2) / (n_A + n_B + n_C - 3)$, where the degrees of freedom is given by where the degrees of freedom is $n_A + n_B + n_C - 3$.

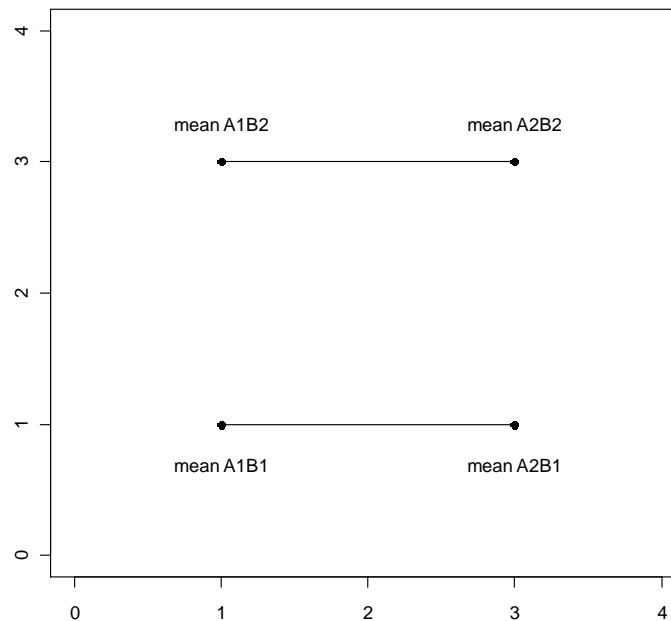
4. Two-way ANOVA

ANOVA stands for analysis of variance. For ANOVA, it is assumed that all populations have the same variances. For a two-way ANOVA, three statistical tests are typically performed: the main effect of each factor and the interaction. Suppose we have two factors A and B, where A represent the genotype and B represent the diet in a mouse study. Suppose each of these factors has two levels (A: wild type, knock out; B: standard diet, high fat diet). For this example, there are 4 combinations ("treatments"): A1B1, A1B2, A2B1, A2B2. The overall ANOVA F-test gives the p-value for testing whether all four of these

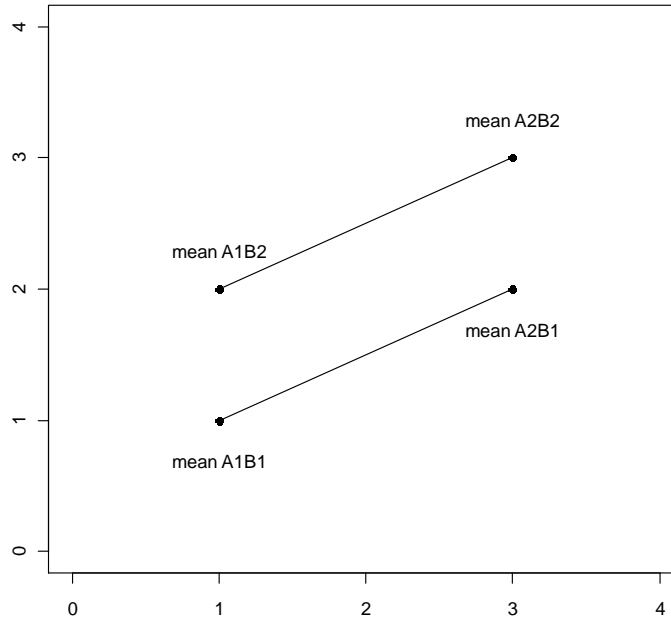
means are equal or whether at least one pair is different. However, we are also interested in the effect of the genotype and diet. A main effect is a contrast that tests one factor across the levels of the other factor. Hence the A main effect compares $(A1B1 + A1B2)/2$ vs. $(A2B1 + A2B2)/2$, and the B-main effect compares $(A1B1 + A2B2)/2$ vs. $(A1B2 + A2B2)/2$. The interaction is a contrast that tests whether the mean difference for one factor depends on the level of the other factor, which is $(A1B2 + A2B1)/2$ vs. $(A1B1 + A2B2)/2$.

Some sample plots follow. For the first plot, there is a B main effect, but no A main effect and no interaction, as the effect of B does not depend on the level of A. For the second plot, notice how the mean difference for B is the same at each level of A and the difference in A is the same for each level of B, hence there is no statistical interaction. The final plot also has main effects for A and B, but here also has an interaction: we see the effect of B depends on the level of A (0 for A1 but 2 for A2), i.e., the effect of the diet depends on the genotype. We also see here the interpretation of the main effects depends on whether there is an interaction or not.

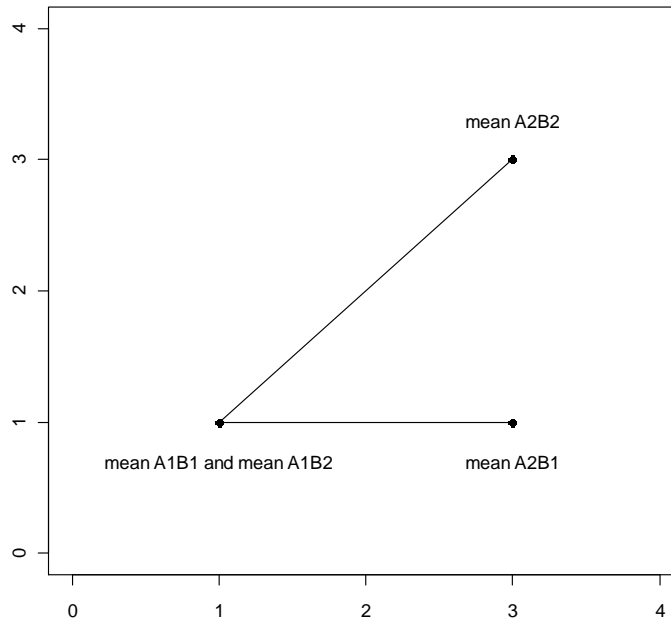
Main Effect for B, but no Main Effect for A, no Interaction



Main Effect for A, Main Effect for B, No Interaction



Main Effect for A, Main Effect for B, with Interaction



5. Two-way Repeated Measures ANOVA

This is typically an ANOVA where one factor is applied to each subject and the second factor is a time point. See two-way ANOVA as many of the details are similar except that the model takes into account the repeated measures, i.e., the treatments are given to the same subject over time. The two main effects and the interaction are assessed, with particular interest to the interaction, as this shows where the time profiles are parallel or not for the treatments (parallel mean no interaction).

One additional note, the standard analysis assumes a condition referred to as compound symmetry, which assumes the correlation between each pair of levels of the repeated-measures factor is the same. Thus, for the case of time, it assumes the correlation is the same between time points 1 and 2, 1 and 3, and 2 and 3.

6. Correlation

Correlation measures the strength and direction of a *linear* association between two variables. The statistical test for correlation tests whether the true correlation is zero or not.

The square of the correlation is the percentage of the total variation explained by a linear relationship between the two variables. Thus, with large sample sizes there may be a sample correlation of 0.1 that is statistically significant. This means we have high confidence that the true correlation is zero, however, only $100*(0.1*0.1)\% = 1\%$ of the variation of one variable is explained by a linear relationship with the other variable, so while there is an association, it has little predictive ability.

7. Hotelling's T² test

The Hotelling's T² test is a multivariate generalization of the *t*-test, but here we are testing whether the mean vectors are different or not (the vector consists of multiple metabolites).

The Hotelling statistic is: $t^2 = \left(\frac{n_x n_y}{n_x + n_y} \right) * (\bar{\mathbf{x}} - \bar{\mathbf{y}})^T \mathbf{S}^{-1} (\bar{\mathbf{x}} - \bar{\mathbf{y}})$, where n_x and n_y are the numbers of samples in each group, $\bar{\mathbf{x}}$ is the mean vector of the variables from group 1, $\bar{\mathbf{y}}$ is the mean vector of variables from group 2 and \mathbf{S} is the pooled estimate of the variance-covariance matrix of the variables. This analysis assumes the underlying variance-covariance matrix is the same for each group. Notice that in the case of uncorrelated variables, this is simply a weighted average of the squared mean differences with weights inversely proportional to the sample variances (i.e., the metabolites less variable within a group are given higher weights).

8. p-values

For statistical significance testing, p-values are given. The lower the p-value, the more evidence we have that the null hypothesis (typically that two population means are equal)

is not true. If “statistical significance” is declared for p-values less than 0.05, then 5% of the time we incorrectly conclude the means are different, when actually they are the same.

The p-value is the probability that the test statistic is at least as extreme as observed in this experiment given that the null hypothesis is true. Hence, the more extreme the statistic, the lower the p-value and the more evidence the data gives against the null hypothesis.

9. q-values

The level of 0.05 is the false positive rate when there is one test. However, for a large number of tests we need to account for false positives. There are different methods to correct for multiple testing. The oldest methods are family-wise error rate adjustments (Bonferroni, Tukey, etc.), but these tend to be extremely conservative for a very large number of tests. With gene arrays, using the False Discovery Rate (FDR) is more common. The family-wise error rate adjustments give one a high degree of confidence that there are zero false discoveries. However, with FDR methods, one can allow for a small number of false discoveries. The FDR for a given set of compounds can be estimated using the q-value (see Storey J and Tibshirani R. (2003) Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA 100: 9440-9445; PMID: 12883005).

In order to interpret the q-value, the data must first be sorted by the p-value then choose the cutoff for significance (typically $p < 0.05$). The q-value gives the false discovery rate for the selected list (i.e., an estimate of the proportion of false discoveries for the list of compounds whose p-value is below the cutoff for significance). For Table 1 below, if the whole list is declared significant, then the false discovery rate is approximately 10%. If everything from Compound 079 and above is declared significant, then the false discovery rate is approximately 2.5%.

Table 1: Example of q-value interpretation

| Compound | p-value | q-value |
|--------------|---------|---------|
| Compound 103 | 0.0002 | 0.0122 |
| Compound 212 | 0.0004 | 0.0122 |
| Compound 076 | 0.0004 | 0.0122 |
| Compound 002 | 0.0005 | 0.0122 |
| Compound 168 | 0.0006 | 0.0122 |
| Compound 079 | 0.0016 | 0.0258 |
| Compound 113 | 0.0052 | 0.0631 |
| Compound 050 | 0.0053 | 0.0631 |
| Compound 098 | 0.0061 | 0.0647 |
| Compound 267 | 0.0098 | 0.0939 |

10. Random Forest

Random forest is a supervised classification technique based on an ensemble of decision trees (see Breiman L. (2001) Random Forests. Machine Learning. 45: 5-32; <http://link.springer.com/article/10.1023%2FA%3A1010933404324>). For a given decision tree, a random subset of the data with identifying true class information is selected to

build the tree (“bootstrap sample” or “training set”), and then the remaining data, the “out-of-bag” (OOB) variables, are passed down the tree to obtain a class prediction for each sample. This process is repeated thousands of times to produce the forest. The final classification of each sample is determined by computing the class prediction frequency (“votes”) for the OOB variables over the whole forest. For example, suppose the random forest consists of 50,000 trees and that 25,000 trees had a prediction for sample 1. Of these 25,000, suppose 15,000 trees classified the sample as belonging to Group A and the remaining 10,000 classified it as belonging to Group B. Then the votes are 0.6 for Group A and 0.4 for Group B, and hence the final classification is Group A. This method is unbiased since the prediction for each sample is based on trees built from a subset of samples that do not include that sample. When the full forest is grown, the class predictions are compared to the true classes, generating the “OOB error rate” as a measure of prediction accuracy. Thus, the prediction accuracy is an unbiased estimate of how well one can predict sample class in a new data set. Random forest has several advantages – it makes no parametric assumptions, variable selection is not needed, it does not overfit, it is invariant to transformation, and it is fairly easy to implement with R.

To determine which variables (biochemicals) make the largest contribution to the classification, a “variable importance” measure is computed. We use the “Mean Decrease Accuracy” (MDA) as this metric. The MDA is determined by randomly permuting a variable, running the observed values through the trees, and then reassessing the prediction accuracy. If a variable is not important, then this procedure will have little change in the accuracy of the class prediction (permuting random noise will give random noise). By contrast, if a variable is important to the classification, the prediction accuracy will drop after such a permutation, which we record as the MDA. Thus, the random forest analysis provides an “importance” rank ordering of biochemicals; we typically output the top 30 biochemicals in the list as potentially worthy of further investigation.

11. Hierarchical Clustering

Hierarchical clustering is an unsupervised method for clustering the data, and can show large-scale differences. There are several types of hierarchical clustering and many distance metrics that can be used. A common method is complete clustering using the Euclidean distance, where each sample is a vector with all of the metabolite values. The differences seen in the cluster may be unrelated to the treatment groups or study design.

12. Principal Components Analysis (PCA)

Principal components analysis is an unsupervised analysis that reduces the dimension of the data. Each principal component is a linear combination of every metabolite and the principal components are uncorrelated. The number of principal components is equal to the number of observations.

The first principal component is computed by determining the coefficients of the metabolites that maximizes the variance of the linear combination. The second component finds the coefficients that maximize the variance with the condition that the second component is orthogonal to the first. The third component is orthogonal to the first two components and so on. The total variance is defined as the sum of the variances of the predicted values of each component (the variance is the square of the standard deviation), and for each component, the proportion of the total variance is computed. For example, if the standard deviation of the predicted values of the first principal component is 0.4 and the total variance = 1, then $100 \cdot 0.4 \cdot 0.4 / 1 = 16\%$ of the total variance is explained by the first component. Since this is an unsupervised method, the main components may be unrelated to the treatment groups, and the “separation” does not give an estimate of the true predictive ability.

13. Z-scores

An intensity measurement for a metabolite by itself does not tell much. If for example a patient contains a blood glucose level of 300, this could be very good news if most people have blood glucose levels around 300, but less so if most people have levels around 100. In other words a measurement is meaningful only relative to the means of the sample or the population. This can be achieved by transforming the measurements into Z-scores which are expressed as standard deviations from the mean.

The Z-score, also called the standard score or normal score, is a dimensionless quantity derived by subtracting the control population mean from an individual raw score and then dividing the difference by the control population standard deviation. The Z-score indicates how many standard deviations an observation is above or below the mean of the control group. The Z-score is negative when the raw score is below the mean, positive when above. Since knowing the true mean and standard deviation of a control population is often unrealistic, the mean and standard deviation of the control population may be estimated using a random control sample.

Z-score =
$$\frac{x - \mu}{\sigma}$$

where: x is a raw score to be standardized, μ is the mean of the control population, σ is the standard deviation of the control population

Subtracting the mean *centers* the distribution, and dividing by the standard deviation *standardizes* the distribution. The interesting properties of Z-scores are that they have a zero mean (effect of “centering”) and a variance and standard deviation of 1 (effect of “standardizing”). This is because all distributions expressed in Z-scores have the same mean (0) and the same variance (1), so we can use Z-scores to compare observations coming from different distributions. When a distribution is normal most of the Z-scores (more than 99%) lay between the values of -3 and +3.