|  |
| --- |
|  |
| mViewTM **FINAL REPORT** |
| **Obese neuropathy**  **MICH-05-18VWCLP** |
| CLIENT: University of Michigan  CLIENT: Eva Feldman, MD, PhD  AUTHOR: Aga Truax, PhD  APPROVAL: Brian Ingram, PhD  DATE: March 21, 2019 |
|  |





Metabolon, Inc. • 617 Davis Drive, Suite 400, Morrisville, NC 27560 • (919) 572-1711 www.metabolon.com • Contact: [info@metabolon.com](mailto:info@metabolon.com)

Table of Contents

[**Objective** 3](#_Toc4014488)

[**Experimental Procedures** 3](#_Toc4014489)

[**Results and Biological Interpretation** 3](#_Toc4014490)

[Metabolite Summary and Significantly Altered Biochemicals 3](#_Toc4014491)

[Biological Interpretation 6](#_Toc4014492)

[**Conclusions** 10](#_Toc4014493)

[**Study Parameters** 11](#_Toc4014494)

[Data Quality: Instrument and Process Variability 11](#_Toc4014495)

[**Appendix** 12](#_Toc4014496)

[Metabolon Platform 12](#_Toc4014497)

[Lipidomics Platform 17](#_Toc4014498)

[Statistical Methods and Terminology 18](#_Toc4014499)

# **Objective**

***Purpose of Experiment***

The goal of this study was to characterize biochemical profiles observed in human plasma samples originating from an obese cohort stratified by a diagnosis of neuropathy as well as a cohort of lean control subjects without a clinical manifestation of neuropathy.

# **Experimental Procedures**

***Experimental design***

Global biochemical profiles were determined in human plasma collected from cohorts as below.

|  |  |  |
| --- | --- | --- |
| **Group** | **n** | **Description** |
| Lean | 43 | Lean subjects without neuropathy |
| Obese | 44 | Obese subjects without neuropathy |
| Obese neuropathy | 44 | Obese subjects with neuropathy |

# **Results and Biological Interpretation**

## Metabolite Summary and Significantly Altered Biochemicals

The present dataset comprises a total of 842 compounds of known identity (named biochemicals) detected on the HD4 platform and 983 detected on the CLP platform. Following log transformation and imputation of missing values, if any, with the minimum observed value for each compound, linear model contrasts were used to identify biochemicals that differed significantly between experimental groups. A summary of the numbers of biochemicals that achieved statistical significance (*p*≤0.05), as well as those approaching significance (0.05<*p*<0.10), is shown below.

An estimate of the false discovery rate (*q*-value) is calculated to take into account the multiple comparisons that normally occur in metabolomic-based studies. For example, when analyzing 200 compounds, we would expect to see about 10 compounds meeting the *p*≤0.05 cut-off by random chance. The *q*-value describes the false discovery rate; a low *q*-value (*q*<0.10) is an indication of high confidence in a result. While a higher *q*-value indicates diminished confidence, it does not necessarily rule out the significance of a result. Other lines of evidence may be taken into consideration when determining whether a result merits further scrutiny. Such evidence may include a) significance in another dimension of the study, b) inclusion in a common pathway with a highly significant compound, or c) residing in a similar functional biochemical family with other significant compounds. Refer to APPENDIX for general definitions and further descriptions of false discovery rate and other statistical tests used at Metabolon.





We have also included in the electronic deliverables, a file with data for each biochemical displayed as box plots like that shown in the example figure below.



## Biological Interpretation

Neuropathy, a result of damage to peripheral nerves, often causes weakness, numbness and pain, usually in hands and feet. There are many causes of peripheral neuropathy, including diabetes, chemo, hereditary disorders, inflammatory infection, auto immune diseases, poor nutrition, certain medications etc. Studies have shown that neuropathy is associated with cardiovascular factors including raised triglyceride levels, body mass-index (BMI), smoking and hypertension (PMIDs: 6811067). This study explored the biochemical alterations associated with neuropathy in human plasma samples collected either from lean control subjects or obese subjects with and without neuropathy.

Datasets provided in the mView product can be quite large and contain a great deal of information. To provide an initial focus for further consideration, a few observations are offered below from a cursory view of the data. These are not presented as a comprehensive analysis; the PI, with a much greater knowledge of the experimental system, is encouraged to make a detailed study of the data for additional or alternative interpretations. For the discussion below, please refer to graphical illustrations in the accompanying PowerPoint file, or to the statistical tables and box plots in the accompanying client data table file (MICH-05-18VWCLP CDT). For convenience, biochemicals are highlighted in **bold** **text** in the report when they correspond to data plots and tables shown in figures of the accompanying Graphics file. This report first presents a high-level overview of the data structure and will then highlight biochemical alterations between groups of interest. Sample 400\_206 has qualified as an outlier in Complex Lipid Analysis, hence the client data table (CDT) provides values and box plots with and without the outlier.

**Overview (slide 3-4):** There were 842 named biochemicals from analysis of the HD4 dataset and 983 from the Complex Lipid Panel (CLP) dataset in plasma samples collected from lean controls and obese subjects with and without neuropathy. At a significance level of *p*<0.05, 42 (HD4) and 49 (CLP) differences between group comparisons can be expected from random chance alone. The number of significant differences amongst the obese subjects with and without neuropathy in the HD4 dataset (62 biochemicals) was found to be just slightly above that expected by random chance alone suggesting that the metabolic signatures associated with neuropathy itself amongst obese subjects will be subtle within the dataset. Furthermore, analysis of the CLP dataset indicated that only 11 biochemicals were significantly different between these groups. As expected a far greater number of changes were observed when the obese subjects (with or without neuropathy) were compared to the healthy controls’ ad noted using linear model contrasts incorporating age, BMI and gender.

* **Principle Component Analysis** **(PCA) (slide 6-8)** is a mathematical procedure that reduces the dimensionality of the data (while retaining most of the variation in a dataset) to allow visual assessment of how individual samples compare to one another (additional information is provided in the “Statistical Analysis” section of the Appendix). Briefly, each ‘circle’ in a PCA plot represents a sample – samples with similar metabolomic profiles will group together in the analysis. The first PCA conducted for this study did not utilize the complex lipid metabolites but it did include the lipid classes measured on the HD4 platform. When analyzed by PCA, there was no obvious gross-level differences between subjects with and without neuropathy **(slide 6, right bottom)**. The most significant separation was noted amongst samples collected from lean controls and obese subjects **(slide 6, right top)**. Additional PCA analysis looking at the BMI and gender of the subjects did not reveal any additional association with neuropathy. *Although none of the subjects with normal weight had neuropathy, there has been incidence of no neuropathy amongst subjects that were very obese with BMI over 50 indicating that although subjects in this study with higher BMI have increased incidence of neuropathy there are still very obese subjects without the disease* **(slide 7, left)**. Similar even neuropathy distribution was noted when looking at male and female subjects **(slide 7, right)**. PCA analysis for CLP in similar fashion demonstrated separation based on obesity but not neuropathy itself **(slide 8)**.
* **Hierarchical clustering analysis (HCA) (slide 9)** is a stepwise clustering method that groups metabolically similar samples close to one another and provides a complimentary mode of analysis to the PCA. HCAhuman plasmasamples did not show very clear distinction between obese subjects with and without neuropathy. Consistent with the above results, the top-level split in the dendrogram was based primarily on obesity status **(slide 9, blue arrows).**
* **Random Forest (RF) analysis differentiated HD4 samples collected from healthy controls and obese subjects with and without neuropathy with a predictive accuracy of 68.7% (slide 10).**  RF is a supervised classification technique based on an ensemble of decision trees (see Appendix for greater detail) and can aid in the identification of biomarkers differentiating classification groups. RF analysis also generates an accompanying list of the top metabolites contributing most to the separation of the groups being compared. Here, a predictive accuracy of 68.7% was obtained when the HD4 data of the three groups were compared (as compared to 33% expected by random chance alone). This number however was largely driven by the ability of the RF to discriminate between the lean and obese subjects. The top 30 metabolites contributing to group separation on the HD4 platform primarily included carbohydrates *(e.g. glucose)*, amino acids, nucleotides and peptides. Pathways related to these biochemicals are discussed further below.
* **Obese subjects with and without neuropathy displayed significant changes in glucose metabolism when compared to the lean controls (slide 11-12). Glucose** can be utilized to support a variety of physiological processes, including energy generation, fatty acid synthesis, protein glycosylation, and nucleotide biogenesis. In this study, obese subjects with and without neuropathy had significantly increased levels of **glucose** when compared to lean controls. Further support for statistically significant higher levels of **glucose** comes from non-significantly decreased levels of **1,5-anhydroglucitol (1,5-AG)**, a stable marker of long-term glycemic control. 1,5-AG in is inversely proportional to circulating **glucose** levels as it competes with **glucose** for reabsorption in the kidney. Through glycolysis, glucose is oxidized to form **pyruvate** and **lactate** to fuel energy production in the TCA cycle. Although without statistical significance, levels of **pyruvate** and **lactate** were elevated in samples collected from obese subjects with and without neuropathy vs samples collected from lean controls further confirming perturbations in glucose metabolism amongst obese subjects **(slide 11)**. Glucose catabolites such as **2,3-diphosphoglycerate** and **3-phosphoglycerate** were significantly decreased amongst samples collected from obese subjects with neuropathy vs obese one without neuropathy suggesting that neuropathy might be further related with changes in glucose metabolism and should be further evaluated. The TCA cycle is a key pathway for oxidative generation of cellular energy. Here, levels of **isocitrate** were significantly elevated while **alpha-ketoglutarate** and malate trended towards higher levels in the obese vs lean comparisons **(slide 12)**. The increase in those metabolites might be associated with altered glycolytic input into the TCA cycle. Interestingly, levels of succinate were significantly elevated amongst obese subjects with neuropathy and decreased with those without neuropathy when contrasted against lean controls pointing out another trend perhaps towards differential energetic use amongst obese subjects with and without neuropathy.
* **Obesity and neuropathy related changes in levels of primary and secondary bile acids (slide 13)**. Bile acids, which are derived from cholesterol, are important for the emulsification and digestion of lipids and have been suggested to play a role in glucose metabolism that has been discussed above (PMID: 19346330, 25194176). Bile acid metabolism is also intimately involved with microbial co-metabolism in the intestine. Secondary bile acids are produced by bacterial enzymatic modifications of primary bile acids and facilitate digestion and absorption of dietary fats. Obese subjects with neuropathy had significantly elevated levels of **cholate** when compared to obese subjects without neuropathy. Additional non-significant increase in most of the detected secondary bile acid metabolites were also noted amongst subjects with neuropathy with significant increase in levels of lithocholate sulfate (1) which might reflect changes in microbiome composition. Furthermore, levels of several primary and secondary bile acids are significantly increased in samples collected from obese subjects (regardless of the neuropathy presence) when compared to healthy controls with the examples of **glycochenodeoxycholate** and **glycoursodeoxycholic acid sulfate (1)**, which could reflect altered liver function or gut microbial metabolism and in consequence might have an impact on glucose metabolism.
* **Total complex lipid levels changes are associated with obesity and neuropathy (slide 14).** The complex lipid panel quantitates individual lipid species across different classes. Lipid class concentrations can then be calculated from the sum of all molecular species within a class. While most lipid classes did not display statistically significant differences between obese subjects with and without neuropathy, there a trending decrease in the levels of sphingomyelins was observed indicating some perturbations in lipid metabolism due to neuropathy. Sphingomyelins are membrane phospholipids and the neuropathy related decrease may result from decreased synthesis or increased degradation. While contrasting obese subjects (with and without neuropathy) to lean controls, a significant increase in the levels of **diacylglycerols** and **triacylglycerols** as well as dihydroceramides was noted. Additionally, significant decreases were also noted in the levels of **lysophosphatidylcholines** (derived from breakdown of phosphatidylcholine) **(slide 14)**. Phospholipids in the blood plasma, would be presumed to derive from lipoprotein particles and other membrane fragments. Changes in phospholipid composition might reflect dietary or lipid synthesis differences.
* **Altered systemic fatty acid levels and oxidation with obesity (regardless of neuropathy) (slide 15-16).** High fat dietary intake and obesity is often associated with systemic accumulation of free fatty acids (FFAs) in rodents (PMID: 26318121). Comparison images from Surveyor contrasting obese subject’s vs lean controls indicate significant increase in several FFAs that are possibly due to changes in absorption and metabolism differences amongst obese and lean individuals **(slide 15, right image)**. These changes were coupled with increase in acylcarnitines, along with slight (but non-significant) increases in the ketone bodies 3-hydroxybutyrate (BHBA) and acetoacetate along with carnitine in samples collected from obese subjects with and without neuropathy vs lean controls **(slide 16)**. Conjugation with carnitine is necessary for transport of long-chain fatty acids across the inner mitochondrial membrane, the site at which they may undergo subsequent β-oxidation. Higher **acylcarnitines**, together with changes in BHBA levels, is often indicative of changes in the β-oxidation of fatty acids. Again, this phenotype seems to be associated more with obesity and less with neuropathy itself. Perhaps of some interest, there were a few acylcarnitines that were significantly elevated when from the obese subjects with neuropathy were compared to those without neuropathy.
* **Obese subjects with and without neuropathy had increased plasma levels of various drugs that might impact metabolites discussed in this report and mask lack of significant differences associated with neuropathy (Slide 17).** In this study, the obese subjects displayed increases in the levels of **metformin** (which is often used for treatment of type 2 diabetes (PMID: 23835523) and elevated levels of **warfarin** indicating cardiovascular complications*.* Additional non-significant changes were also present amongst pain medications such as 2-acetominophen sulfate and **2-hydroxyibuprofen** in comparison between obese subjects and lean controls. Levels of **caffeine and cotinine** were also higher amongst obese subjects indicating increased coffee consumption and tabaco use which also might also affect the metabolic signatures observed here.

# **Conclusions**

In conclusion, the analysis of this study comparing human plasma samples collected from lean controls and obese subjects with and without neuropathy identified a number of biochemical pathways that are known to be associated with obesity and few that might be related to neuropathy itself. The differences identified between the obese subjects and lean controls included a decrease in 1,5-AG levels, hyperglycemia, alterations in the TCA cycle, and changes in several lipid metabolites. Additional changes observed amongst primary and secondary bile acid metabolites might be indicative of changes in the microbiome that are obesity and to some degree neuropathy related. Metabolic differences noted on the Complex Lipid Platform associated with obesity and neuropathy included altered free fatty acid levels, changes in fatty acid levels, and altered complex lipid composition, specifically TAGs, DAGs.

In order to see neuropathy associated changes follow up experiments looking at other variables such as the disease severity, disease length etc. might be needed.

# **Study Parameters**

## Data Quality: Instrument and Process Variability

|  |  |  |  |
| --- | --- | --- | --- |
| ***QC Sample*** | ***Measurement*** | ***Median RSD Global*** | ***Median RSD Complex Lipids*** |
| Internal Standards | Instrument Variability | 4 % | NA |
| Endogenous Biochemicals | Total Process Variability | 9 % | 9 % |

Instrument variability was determined by calculating the median relative standard deviation (RSD) for the internal standards that were added to each sample prior to injection into the mass spectrometers. Overall process variability was determined by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in 100% of the MTRX samples, which are technical replicates created from a large pool of extensively characterized human plasma. Values for instrument and process variability as shown in the table above meet Metabolon’s acceptance criteria.

# **Appendix**

## Metabolon Platform

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

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

**QA/QC:** Several types of controls were analyzed in concert with the experimental samples: a pooled matrix sample generated by taking a small volume of each experimental sample (or alternatively, use of a pool of well-characterized human plasma) served as a technical replicate throughout the data set; extracted water samples served as process blanks; and a cocktail of QC standards that were carefully chosen not to interfere with the measurement of endogenous compounds were spiked into every analyzed sample, allowed instrument performance monitoring and aided chromatographic alignment. Tables 1 and 2 describe these QC samples and standards. Instrument variability was 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 was determined by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in 100% of the pooled matrix samples. Experimental samples were 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 customer sample. | Assess the effect of a non-plasma matrix on the Metabolon process and distinguish biological variability from process variability. |
| PRCS | Aliquot of ultra-pure water | Process Blank used to assess the contribution to compound signals from the process. |
| 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 utilized 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 was dried then reconstituted in solvents compatible to each of the four methods. Each reconstitution solvent contained a series of standards at fixed concentrations to ensure injection and chromatographic consistency. One aliquot was analyzed using acidic positive ion conditions, chromatographically optimized for more hydrophilic compounds. In this method, the extract was gradient eluted from a C18 column (Waters UPLC BEH C18-2.1x100 mm, 1.7 µm) using water and methanol, containing 0.05% perfluoropentanoic acid (PFPA) and 0.1% formic acid (FA). Another aliquot was also analyzed using acidic positive ion conditions, however it was chromatographically optimized for more hydrophobic compounds. In this method, the extract was gradient eluted from the same afore mentioned C18 column using methanol, acetonitrile, water, 0.05% PFPA and 0.01% FA and was operated at an overall higher organic content. Another aliquot was analyzed using basic negative ion optimized conditions using a separate dedicated C18 column. The basic extracts were gradient eluted from the column using methanol and water, however with 6.5mM Ammonium Bicarbonate at pH 8. The fourth aliquot was analyzed via negative ionization following elution from a HILIC column (Waters UPLC BEH Amide 2.1x150 mm, 1.7 µm) using a gradient consisting of water and acetonitrile with 10mM Ammonium Formate, pH 10.8. The MS analysis alternated between MS and data-dependent MSn scans using dynamic exclusion. The scan range varied slighted between methods but covered 70-1000 m/z. Raw data files are archived and extracted as described below.

**Bioinformatics:** The informatics system consisted 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 information interpretation and visualization tools for use by data analysts. The hardware and software foundations for these informatics components were the LAN backbone, and a database server running Oracle 10.2.0.1 Enterprise Edition.

**LIMS:**  The purpose of the Metabolon LIMS system was 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 and instrumental analysis and reporting and advanced data analysis. All of the subsequent software systems are grounded in the LIMS data structures. It 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 was 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 were 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 between the experimental data and authentic standards. The MS/MS scores are based on a comparison of the ions present in the experimental spectrum to the ions present in the library 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 3300 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 were carried out to ensure that a high quality data set was made available for statistical analysis and data interpretation. The QC and curation processes were designed to ensure accurate and consistent identification of true chemical entities, and to remove those representing system artifacts, mis-assignments, and background noise. Metabolon data analysts use proprietary visualization and interpretation software to confirm the consistency of peak identification among the various samples. Library matches for each compound were checked for each sample and corrected if necessary.

**Metabolite Quantification and Data Normalization:** Peaks were quantified using area-under-the-curve. For studies spanning multiple days, a data normalization step was performed to correct variation resulting from instrument inter-day tuning differences. Essentially, each compound was corrected in run-day blocks by registering the medians to equal one (1.00) and normalizing each data point proportionately (termed the “block correction”; Figure 2). For studies that did not require more than one day of analysis, no normalization is necessary, other than for purposes of data visualization. In certain instances, biochemical data may have been normalized to an additional factor (e.g., cell counts, total protein as determined by Bradford assay, osmolality, etc.) to account for differences in metabolite levels due to differences in the amount of material present in each sample.

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

## Lipidomics Platform

**Sample preparation**: Lipids were extracted from the bio-fluid in the presence of deuterated internal standards using an automated BUME extraction according to the method of Lofgren et al. (J Lipid Res 2012;53(8):1690-700).

**Data Acquisition**: The extracts were dried under nitrogen and reconstituted in a dichloromethane:methanol solution containing ammonium acetate. The extracts were transferred to vials for infusion-MS analysis, performed on a Shimadzu LC with nano PEEK tubing and the Sciex SelexIon-5500 QTRAP. The samples were analyzed via both positive and negative mode electrospray. The 5500 QTRAP was operated in MRM mode with a total of more than 1,100 MRMs. Individual lipid species were quantified by taking the ratio of the signal intensity of each target compound to that of its assigned internal standard, then multiplying by the concentration of internal standard added to the sample. Lipid class concentrations were calculated from the sum of all molecular species within a class, and fatty acid compositions were determined by calculating the proportion of each class comprised by individual fatty acids.

## 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* , and the degrees of freedom is given by , where , are the sample means, s1, s2, are the sample standard deviations, and n1, n2 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).

1. **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 (x1 – x2) for each subject; test whether the mean difference is zero or not). The test statistic is given by , with *n* – 1 degrees of freedom, where , are the sample means for groups 1 and 2, respectively, sd is the standard deviation of the differences, *n* is the number of *subjects* (so there are 2*n* observations).

1. **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* , and the degrees of freedom is given by

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

where , where the degrees of freedom is *nA* + *nB* – 2.

The contrast from the ANOVA,

where , where the degrees of freedom is given by where the degrees of freedom is *nA* + *nB* + *nC* – 3.

1. **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.







1. **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.

1. **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.

1. **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: , where *nx* and *n*y are the numbers of samples in each group, is the mean vector of the variables from group 1, is the mean vector of variables from group 2 and **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).

1. **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.

1. **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



1. **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.

1. **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.

1. **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\*0.4\*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.