Supplementary reports (tissue): Report S1
Experimental procedure, statistical analyses and data management systems and results and biological interpretation

Objective

Purpose of Experiment

The goal of this study was to identify the biochemical profiles of matched-pairs of head and neck squamous cell carcinoma samples from indolent and metastatic tissues along with benign adjacent tissues.

Experimental Procedures

Experimental design

Metabolon received 57 tumor tissue samples on December 10, 2014. Global metabolic profiles were determined from the experimental groups outlined below.

Group	Group Description	n
Control	Normal adjacent tissue	19
Primary	Primary HNSCC tumor tissue	19
Metastatic	Metastatic tumor tissue	19

Results and Biological Interpretation

Metabolite Summary and Significantly Altered Biochemicals

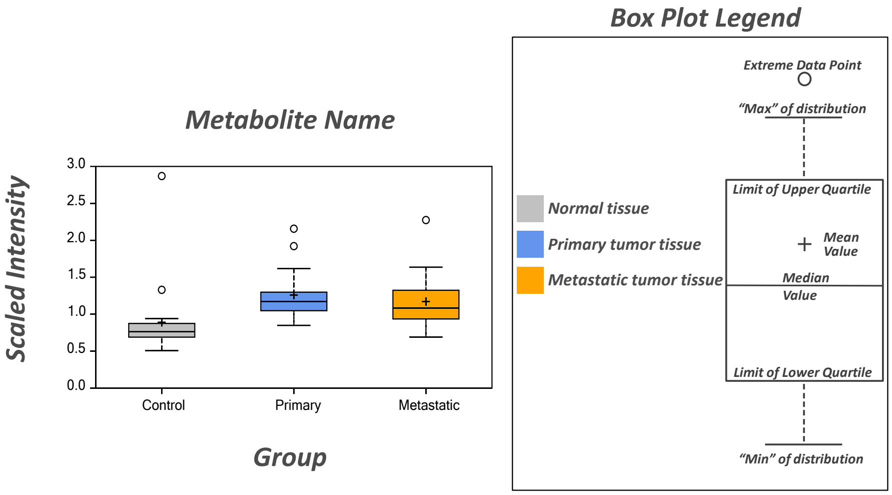
The present dataset comprises a total of 569 compounds of known identity (named biochemicals). Following log transformation and imputation of missing values, if any, with the minimum observed value for each compound, ANOVA contrasts were used to identify biochemicals that differed significantly between experimental groups. A summary of the numbers of biochemicals that achieved statistical significance ($p \leq 0.05$), as well as those approaching significance ($0.05 < p < 0.10$), is shown below. Analysis by two-way ANOVA with repeated measures identified biochemicals exhibiting main effects of the group experimental parameter.

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 $p \leq 0.05$ 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 the Appendix for general definitions and further descriptions of false discovery rate and other statistical tests used at Metabolon.

Statistical Comparisons			
ANOVA Contrasts	**Primary** **Control**	**Metastatic** **Control**	**Metastatic** **Primary**
Total biochemicals ***p*≤0.05**	385	383	63
Biochemicals **(↑↓)**	**301 | 84**	**288 | 95**	**22 | 41**
Total biochemicals **0.05<*p*<0.10**	30	35	42
Biochemicals **(↑↓)**	**16 | 14**	**19 | 16**	**17 | 25**

Statistical Comparisons	
RM ANOVA	**Group Main Effect**
Total biochemicals ***p*≤0.05**	405
Total biochemicals **0.05<*p*<0.10**	24

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

The majority of head and neck cancers are squamous cell carcinomas, which typically originate from the mucosal epithelial lining of the oral cavity. These cancers show a strong association with tobacco use, alcohol consumption, gastrointestinal reflux (GERD) as well as human papillomavirus (HPV) infection and are often aggressive, but respond well to surgical excision and radiation therapy if detected early. The goal of this study was to identify metabolomic differences between squamous cell tumors (either primary or metastatic) and neighboring tissue, with a secondary goal of identifying biomarkers associated with metastatic potential. A related project (MICH-01-15VW) will assess potential biomarkers in saliva in individuals with or without disease.

Datasets provided in the mView product can be quite large and contain a great deal of information. A few observations are offered below as an initial overview of the changes in metabolic profiles in clinical tissue samples; key references are cited by PubMed Identification number (PMID) at certain points throughout the report. For convenience, biochemicals are highlighted in **bold text** in the report when they correspond to plots shown in figures of the accompanying Graphics file. Comparison of global biochemical profiles derived from tumors (either primary or metastatic) or "normal" adjacent tissue revealed several metabolic differences, some of which are highlighted below:

- **Overview of the dataset**: Principal component analysis (PCA) transforms a large number of metabolic variables into a smaller number of orthogonal variables (Component 1, Component 2, etc…) in order to analyze variation between groups and to provide a high-level overview of the dataset. Samples derived from normal adjacent tissue (called Control here) showed good separation from tumor samples; however, Primary and Metastatic tumor samples formed an overlapping population on the PCA. When analyzed by tumor grade, tumor samples again formed overlapping populations. *Less advanced grades may be* *pulling away from more advanced grades on the PCA (slide 6), but low sample numbers in lower grades make it difficult to assess populations.* In the hierarchical clustering analysis (HCA), Control samples tended to cluster to the left major branch of the dendrogram, while tumor samples clustered to the right branch; tumor samples tended to for sub-clusters with their matched tumor pair rather than by tumor status (primary or metastatic).

Random forest analysis (RFA) is a statistical tool utilizing a supervised classification technique based on an ensemble of decision trees (please see Appendix for greater detail) and can aid in the identification of biomarkers differentiating classification groups. RFA showed good efficiency at separating Control, Primary and Metastatic samples, with a predictive accuracy of 67% (random chance would be expected to yield a predictive accuracy of 33% in this analysis). Tumor samples appeared to be the source of error in the analysis: when these samples were mis-segregated, they tended to be classified with the alternate tumor group (for example, Primary were mis-segregated into the Metastatic bin). *The dividing line between metastatic and primary tumor is more of a continuum; in this case, further division based on genetic (Ras, p53 mutational status) or cell biological (e.g., matrix metalloprotease expression) criteria may give better results.* RFA attempting to classify tumors by subtype (Primary or Metastatic) yielded a predictive accuracy of 53% (similar to a random segregation of 50%). *The Top 30 metabolites for predicting Control, Primary and Metastatic treatment groups included biochemicals related to lipid metabolism (caprate, docosatrienoate, CDP-choline, adrenate), urea cycle (pro-hydroxy-pro, putrescine) and inflammation (kynurenine).*

- **Energetics**: Glucose can be utilized to support a variety of physiological processes, including energy generation, fatty acid synthesis, protein glycosylation, and nucleotide biogenesis. Glycolytic metabolites were suggestive of increased use in both Primary and Metastatic (compared to Control): **Glucose** and 3-carbon glycolytic intermediates (2-phosphoglycerate, **3-phosphoglycerate**, phosphoenolpyruvate) were decreased, while **glucose-6-phosphate** and fructose-6-phosphate were increased. *Decreases in the 3-carbon glycolytic intermediates are typically indicative of increasing use, while elevation in 6-carbon intermediates can indicate changes in glucose availability (potentially reflecting increasing glucose import).* In Metastatic (compared to Primary), a non-significant increase in glucose (and decrease in 3-phosphoglycerate and phosphoenolpyruvate) could indicate increasing glucose availability (potentially due to upregulated GLUT transporters) with increased glycolytic use. *Interestingly, increased expression of the glycolytic enzyme fructose-bisphosphate aldolase A (ALDOA) is correlated with metastasis (and poor prognosis) in lung squamous cell cancers (PMID: 24465716).* The glycolytic end-products **pyruvate** and **lactate** were both increased, consistent with increased glycolytic use. *Acetylcarnitine can be used as a surrogate marker for acetyl CoA; both Primary and Metastatic (compared to Control) showed decreased acetylcarnitine, with a further non-significant decrease in Metastatic (compared to Primary) suggestive of increasing energy demand. Alternately, decreasing acetylcarnitine could indicate Warburg metabolism, where conversion of pyruvate to lactate takes precedence over entry into the TCA cycle to support oxidative metabolism (decreased* *citrate with elevated isocitrate, potentially reflecting increasing pool size due to decreased substrate input, may indicate declining entry of pyruvate into the TCA cycle).*

Glycogen metabolism: Glycogen synthesis proceeds by the reaction of glucose-1-phosphate (G1P) with UTP, producing UDP-glucose for glycogen chain incorporation; glycogenolysis produces glycogen fragments (e.g., maltopentaose, maltotetraose) which can then be converted to G1P for use in glycolysis. Glycogen metabolites **maltohexaose**, **maltopentaose**, **maltotetraose**, **maltotriose**, and **maltose** were all decreased (Primary and Metastatic vs Control), while UDP-glucose was increased. *This pattern could suggest increasing glycogen synthesis; several tumor types show increased glycogen storage in response to hypoxic condition (PMID: 23177934). Alternately, UDP-glucose may be elevated to support glycosylation, with decreased glycogen metabolites indicative of increased use to support glycolysis. A histological assessment of glycogen deposition could shed further light on glycogen use in squamous cell tumors.*

BCAA catabolism: Metabolites derived from leucine, isoleucine or valine catabolism can enter gluconeogenesis or the TCA cycle for energy production; **leucine**, **isoleucine** and **valine** were all increased (Primary and Metastatic vs Control). Catabolic metabolites of leucine (**4-methyl-2-oxopentanoate**, beta-hydroxyisovalerate, and alpha-hydroxyisovaleroyl carnitine), isoleucine (**3-methyl-2-oxovalerate**, tiglyl carnitine) and valine (**3-methyl-2-oxobutyrate**, 3-hydroxyisobutyrate) were all decreased, suggestive of changing use. *Increases in 3-methylhistidine, which is derived from actin/myosin turnover, could suggest an increase in muscle catabolism to support energy demand.* Metastatic (compared to Primary) showed further decreases in several catabolic intermediates, including beta-hydroxyisovaleroylcarnitine, alpha-hydroxyisovalerate, tiglyl carnitine, 2-hydroxy-3-methlvalerate, and alpha-hydroxyisocaproate, which may be indicative of increasing use of BCAAs for energetics. *Increases in branched-chain aminotransferase (BCAT) expression, which catalyzes the first step of BCAA catabolism, have been identified in nasopharyngeal carcinoma, which have been associated with increased metastatic potential (PMID: 23758864). Changes in BCAA catabolites could indicate increasing use for energetics.*

Lipid metabolism: Fatty acids (FAs) are a critical source of energy for mitochondrial oxidation and cellular ATP generation. **Long-chain fatty acids** were increased as a class (Primary and Metastatic vs Control), which could reflect changes in plasma membrane architecture (potentially associated with changes in cell signaling) or changes in beta-oxidative use. Long-chain FAs must be conjugated to carnitine for transport across the mitochondrial membrane prior to oxidation; increases in higher chain length acylcarnitines could reflecting changing beta-oxidative use (Primary and Metastatic vs Control). Metastatic (compared to Control or Primary) showed a decrease in the ketone body **3-hydryoxybutyrate (BHBA)**, and an increase in medium-chain fatty acids (**caprylate**, **caprate**, **5-dodecenoate**) and a subset of long-chain fatty acids (myristate, myristoleate, oleate, arachidate), which could indicate decreased use of beta-oxidation (potentially with increased reliance on glycolysis and/or amino acid catabolism for energy generation).

TCA cycle: Changes in the pattern of TCA metabolites could indicate changing function in Primary and Metastatic (compared to Control); interestingly, **glutamine** showed trends toward decrease in Metastatic (compared to Control or Primary), which could suggest increased use for glutaminolysis (increases in **alpha-ketoglutarate** could reflect glutamine entry into the TCA cycle). *A previous metabolomics study has suggested glycolysis as the primary energy source in squamous cell cancers of the head and neck (PMID: 21692052); increasing glutamine input into the TCA cycle in metastatic tumors may be indicative of increased energy demand. Interestingly, one recent study (PMID: 24316975) has shown that glutamine can be converted into 2-hydroxyglutarate (which was also elevated in this project, Primary and Metastatic vs Control) in cancers with myc activation, which was associated with poor prognosis. PET imaging using labelled glutamine substrates (PMID: 22095958) could be tested as one marker of metastatic risk.*

- **Inflammation-associated metabolites:** Changes in the ratio of n3:n6 polyunsaturated fatty acids (PUFAs) can be one readout of inflammation; PUFAs were increased regardless of C=C placement in Primary and Metastatic (compared to Control), which could indicate changes in plasma membrane structure/organization (potentially resulting from changes in cell signaling). Eicosanoids are enzymatically derived from PUFAs; **prostaglandin F2alpha** (**PGF2a**) was elevated in both Primary and Metastatic (compared to Control), with a non-significant increase in **prostaglandin E2** (**PGE2**) and **6-keto prostaglandin F1alpha** (**6-keto PGF1a**). Endocannabinoids are typically considered to be anti-inflammatory (but are induced in inflammatory states), promoting their effects through interaction with cannabinoid receptors and, in some cases, G-protein coupled receptors (GPCRs). N-stearoyltaurine, oleic ethanolamide and palmitoyl ethanolamide were all elevated in Primary and Metastatic (compared to Control). *Few significant differences in eicosanoids or endocannabinoids were identified in Metastatic (compared to Primary), which could suggest similar levels of inflammation.*

  Changes in **tryptophan** metabolites can also indicate inflammatory states: indoleamine 2,3-dioxygenase (IDO), which catalyzes the conversion of tryptophan to kynurenine, is activated by pro-inflammatory cytokines (e.g., IFN-γ, TNF-α). **Kynurenine** and its degradative metabolite **kynurenate** were both elevated (Primary and Metastatic vs Control), consistent with increased inflammation. *Kynurenine plays an anti-inflammatory role as a "natural brake" on the immune response; non-significant decrease in kynurenine and kynurenate (Metastatic vs Primary) could indicate changing inflammation (but not necessarily a decrease).* Finally, **histamine** showed a trend toward decrease (Metastatic vs Primary), with increases in the histamine degradation product **1-methylimidazoleacetate**. *Decreased histamine may suggest changes in mast cell number or function are associated with metastatic potential (mast cells have also been linked to changes in tumor vasculature).*

- **Redox Homeostasis:** Increases in methionine sulfone and methionine sulfoxide (oxidative products of methionine), S-methylcysteine (an oxidative product of cysteine) and cysteine-glutathione disulfide (an oxidative product of glutathione) could be indicative of oxidative stress (Primary and Metastatic vs Control). **Oxidized (GSSG)** and **reduced (GSH) glutathione** and gamma-glutamyl amino acids were also increased, consistent with elevated oxidative stress. *Metastatic compared to Primary showed increased GSSG, suggestive of elevated oxidative stress.* **Methionine** metabolites appeared elevated as a class, which could suggest high demand for glutathione synthesis. *Increases in norophthalmate and non-significant increase in ophthalmate, tripeptide analogues of GSH produced by glutathione synthetase in which cysteine has been replaced by alanine or 2-aminobutyrate, respectively, are consistent with increased glutathione demand.* **Carnosine** and **anserine**, two dipeptide derivatives of histidine with anti-oxidant function, were also decreased (Primary and Metastatic vs Control), consistent with increased use to support redox homeostasis. *Further decrease in carnosine and anserine in Metastatic (compared to Primary) could support increasing oxidative stress in metastatic cancer cells.*

Other changes of potential interest:

- <u>Heme</u>: Heme was decreased in Primary and Metastatic tumors (compared to Control tissue), which could suggest relative decreases in vascularization (compared to adjacent tissue). Poor vascularization in many tumors result in necrotic cores with relatively poor oxygenation, which can lead to shifts away from oxidative metabolism for energy generation (which may be seen here as increases in lactate production). *Similarly, primary and secondary bile acids showed trends toward decrease (Primary and Metastatic vs Control); these products are typically re-absorbed in the intestine and are returned to the liver via the bloodstream.*

- <u>Choline metabolism</u>: Abnormal choline metabolism has previously been associated with metastatic tumors (reviewed in PMID: 22089420). Choline showed a trend toward increase (Metastatic vs Primary), with decreased GPC-containing lysolipids (and non-significant decrease in glycerophosphorylcholine, GPC) potentially indicating changes in phospholipase D (PLD) or PC-phospholipase C (PC-PLC) activity in metastatic tumors. *Mutations in PLD isoforms have been identified in a number of cancers and correlates with invasive potential in breast cancer cells.*

- <u>Cotinine metabolites</u>: Cotinine, the active metabolite of nicotine, was detected in approximately half of samples in each group. Further sub-grouping by tumor origin (HPV-related, smoking-induced, etc.) could reveal metabolic differences associated with tumor subtypes.

Conclusions

In conclusion, the results from this global metabolomic study comparing control adjacent, primary and metastatic tumor samples differed in a number of metabolic readouts, including changes in metabolites related to energetics, inflammation and markers of oxidative stress. In the principal component analysis (PCA), tumor samples (Primary and Metastatic) formed an overlapping population that was well-separated from Control samples. Similarly, in the hierarchical clustering analysis (HCA), normal and tumor samples clustered into different branches of the dendrogram, with sub-clustering of tumor samples from the same subject. Energetics metabolites suggested increased use of glycolysis for energetics in tumor samples (both primary and metastatic), with decreased fatty acid beta-oxidation in Metastatic (compared to Primary) tumors suggestive of increased reliance on glycolysis, glutaminolysis and/or protein catabolism for energetics. Inflammation-related metabolites were elevated in tumor samples, with subtle trends toward decrease in several markers of inflammation in metastatic samples (compared to primary tumor). Finally, patterns of metabolites suggest increasing oxidative stress in tumor samples, with a further increase in metastatic samples. Further studies correlating biochemicals with genetic mutations, cell biological criteria, and response to treatment could further link changes in particular metabolites with cancer progression, metastasis and response to therapy.

Data Quality: Instrument and Process Variability

QC Sample	*Measurement*	*Median RSD*
Internal Standards	Instrument Variability	6 %
Endogenous Biochemicals	Total Process Variability	8 %

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 Client Matrix samples, which are technical replicates of pooled client samples. Values for instrument and process variability meet Metabolon's acceptance criteria as shown in the table above.

Appendix

Metabolon Platform

Sample Accessioning: Following receipt, samples were inventoried and immediately stored at -80°C. 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 -80°C until processed.

Sample Preparation: Samples were prepared using the automated MicroLab STAR® system from Hamilton Company. A recovery standard was 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: one for analysis by UPLC-MS/MS with positive ion mode electrospray ionization, one for analysis by UPLC-MS/MS with negative ion mode electrospray ionization, one for analysis by UPLC-MS/MS polar platform (negative ionization), one for analysis by GC-MS, and one sample was reserved for backup. Samples were placed briefly on a TurboVap® (Zymark) to remove the organic solvent. For LC, the samples were stored overnight under nitrogen before preparation for analysis. For GC, each sample was dried under vacuum overnight 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.
DS	Derivatization Standard	Assess variability of derivatization for GC-MS samples.
IS	Internal Standard	Assess variability and performance of instrument.

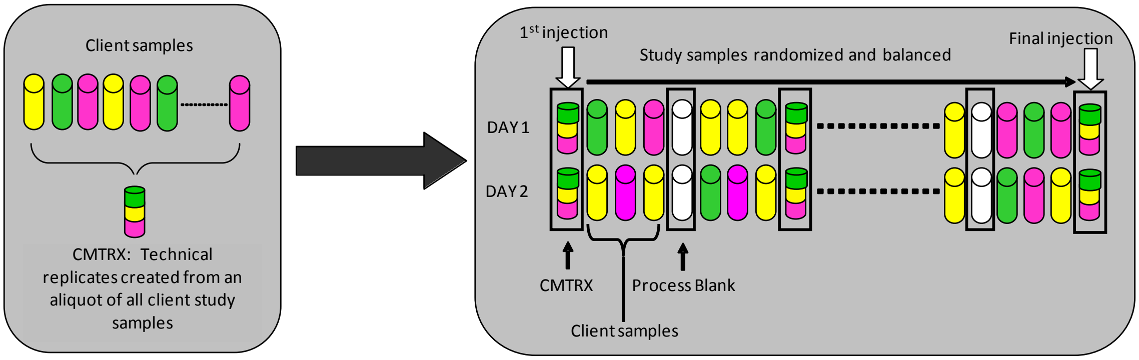

Client samples
CMTRX: Technical replicates created from an aliquot of all client study samples
1st injection
Study samples randomized and balanced
Final injection
DAY 1
DAY 2
CMTRX
Process Blank
Client samples

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):
The LC/MS portion of the platform was based on 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 acidic or basic LC-compatible solvents, each of which contained 8 or more

injection standards at fixed concentrations to ensure injection and chromatographic consistency. One aliquot was analyzed using acidic positive ion optimized conditions and the other using basic negative ion optimized conditions in two independent injections using separate dedicated columns (Waters UPLC BEH C18-2.1x100 mm, 1.7 µm). Extracts reconstituted in acidic conditions were gradient eluted from a C18 column using water and methanol containing 0.1% formic acid. The basic extracts were similarly eluted from C18 using methanol and water, however with 6.5mM Ammonium Bicarbonate. The third 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. The MS analysis alternated between MS and data-dependent MS^2 scans using dynamic exclusion, and the scan range was from 80-1000 *m/z*. Raw data files are archived and extracted as described below.

Gas Chromatography-Mass Spectroscopy (GC-MS): The samples destined for analysis by GC-MS were dried under vacuum for a minimum of 18 h prior to being derivatized under dried nitrogen using bistrimethyl-silyltrifluoroacetamide. Derivatized samples were separated on a 5% diphenyl / 95% dimethyl polysiloxane fused silica column (20 m x 0.18 mm ID; 0.18 um film thickness) with helium as carrier gas and a temperature ramp from 60° to 340°C in a 17.5 min period. Samples were analyzed on a Thermo-Finnigan Trace DSQ fast-scanning single-quadrupole mass spectrometer using electron impact ionization (EI) and operated at unit mass resolving power. The scan range was from 50–750 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 +/- 0.005 amu, 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 distribution to both the LC-MS and GC-MS 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.

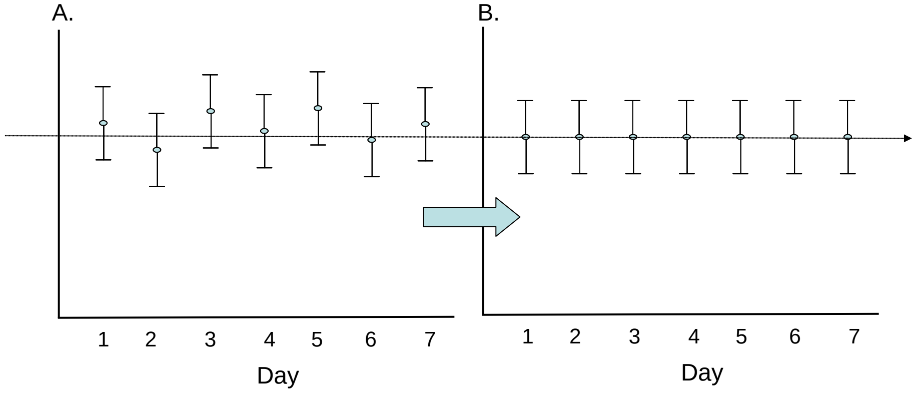


Figure 2: Visualization of data normalization steps for a multiday platform run.

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 $t = (\bar{x}_1 - \bar{x}_2)/n$, 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 2*n* 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

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

where $s_{AB}^2 = \left((n_A - 1)s_A^2 + (n_B - 1)s_B^2\right)/(n_A + n_B - 2)$, where the degrees of freedom is $n_A + n_B - 2$.
The contrast from the ANOVA,

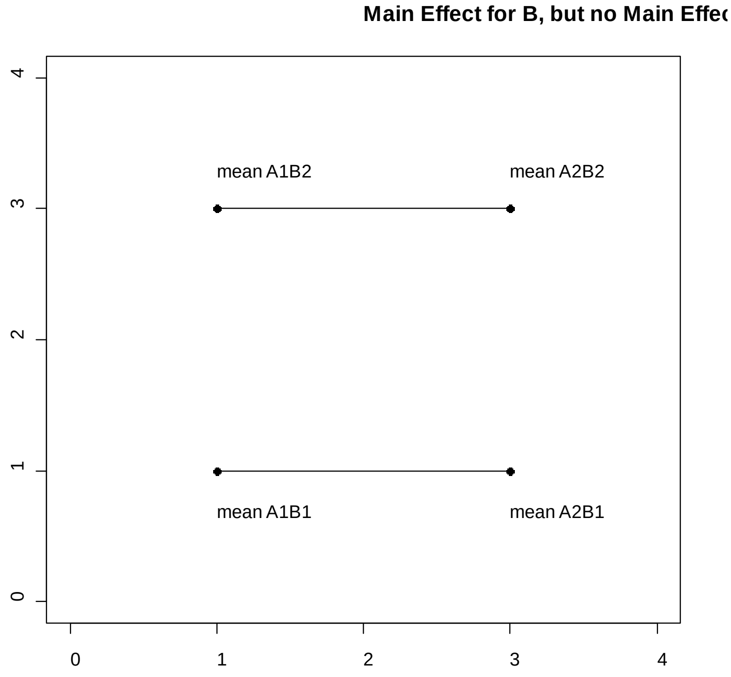
$$t = (\bar{x}_A - \bar{x}_B)/\sqrt{s^2(1/n_A + /n_B)}$$

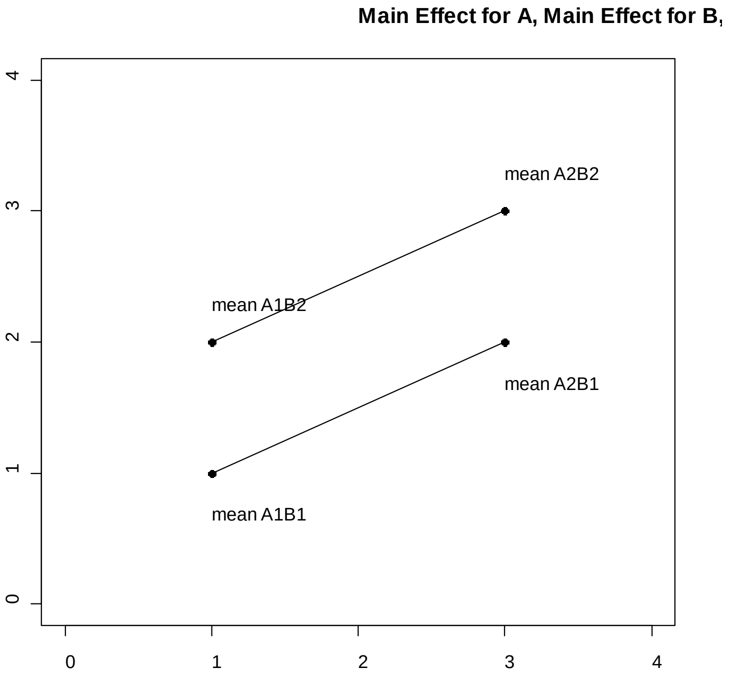
where $s^2 = \left((n_A - 1)s_A^2 + (n_B - 1)s_B^2 + (n_C - 1)s_C^2\right)/(n_A + n_B + n_C - 3)$, where the degrees of freedom is given by where the degrees of freedom is $n_A + n_B + n_C - 3$.

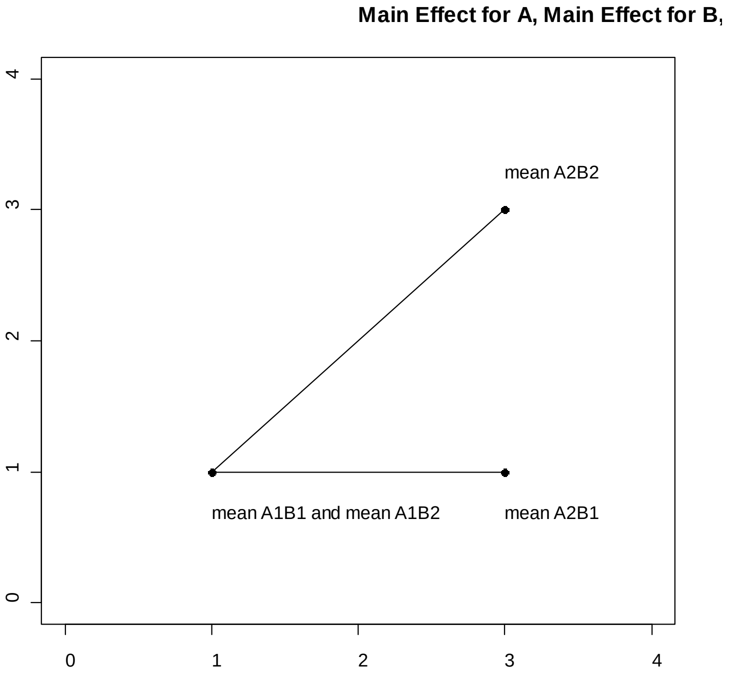
4. Two-way ANOVA

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 A, Main Effect for B,


Main Effect for A, Main Effect for B,

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 T^2 test**

   The Hotelling's T^2 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: $t^2 = \left(\frac{n_x\, n_y}{n_x+n_y}\right) * (\overline{\boldsymbol{x}} - \overline{\boldsymbol{y}})^T\, \boldsymbol{S}^{-1}\, (\overline{\boldsymbol{x}} - \overline{\boldsymbol{y}})$, where n_x and n_y are the numbers of samples in each group, $\overline{\boldsymbol{x}}$ is the mean vector of the variables from group 1, $\overline{\boldsymbol{y}}$ 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).

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*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.

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.

Supplementary reports (saliva): Report S1

Experimental procedure, statistical analyses and data management systems and results and biological interpretation

Objective

Purpose of Experiment

The goal of this study was to characterize the biochemical profiles of saliva from patients with head and neck squamous cell carcinoma in order to identify biomarkers of disease.

Experimental Procedures

Experimental design

Metabolon received 60 saliva samples on February 19, 2015. Global metabolic profiles were determined from the experimental groups outlined below.

Group	Group Description	n
Control	Saliva from healthy controls	13
Disease	Saliva from patients with head and neck squamous cell carcinoma	47

Results and Biological Interpretation

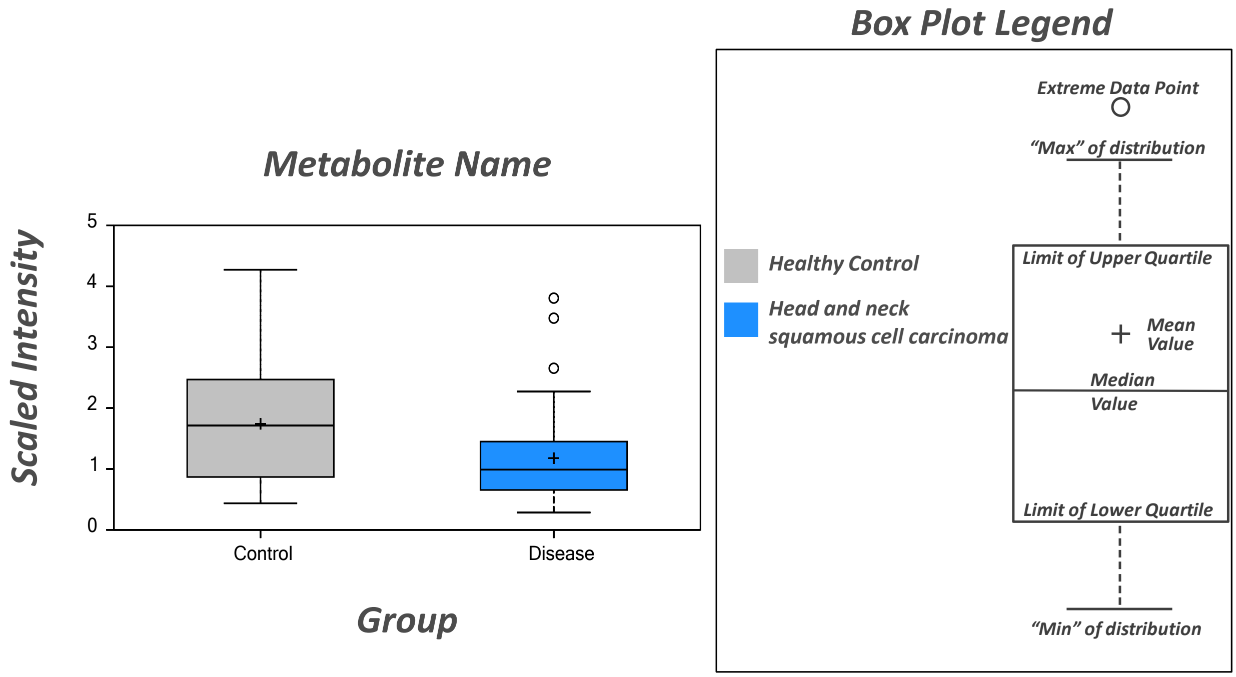
Metabolite Summary and Significantly Altered Biochemicals

The present dataset comprises a total of 481 compounds of known identity (named biochemicals). Following log transformation and imputation of missing values, if any, with the minimum observed value for each compound, Welch's two-sample *t*-test was used to identify biochemicals that differed significantly between experimental groups. A summary of the numbers of biochemicals that achieved statistical significance ($p \leq 0.05$), 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 $p \leq 0.05$ 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 the Appendix for general definitions and further descriptions of false discovery rate and other statistical tests used at Metabolon.

Statistical Comparisons		
Welch's Two-Sample *t*-Test	**Disease** **Control**	**Disease** **Control (No 379-399)**
Total biochemicals *p*≤0.05	175	234
Biochemicals (↑↓)	**126 | 49**	**120 | 114**
Total biochemicals 0.05<*p*<0.10	40	40
Biochemicals (↑↓)	**22 | 18**	**26 | 14**

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

The majority of head and neck cancers are squamous cell carcinomas, which typically originate from the mucosal epithelial lining of the oral cavity. These cancers show a strong association with tobacco use and alcohol consumption, gastrointestinal reflux (GERD) as well as human papillomavirus (HPV) infection. These cancers are often aggressive, but respond well to surgical excision and radiation therapy if detected early. The goal of this study is to identify salivary biomarkers associated with head and neck squamous cell carcinomas (which will also be compared to metabolic changes identified in the related project, MICH-03-14VW, which assessed metabolites in primary and metastatic tumors compared to normal adjacent tissue).

Datasets provided in the mView product can be quite large and contain a great deal of information. A few observations are offered below as an initial overview of the changes in metabolic profiles in saliva samples; key references are cited by PubMed Identification number (PMID) at certain points throughout the report. For convenience, biochemicals are highlighted in **bold text** in the report when they correspond to plots shown in figures of the accompanying Graphics file. Comparison of global biochemical profiles derived from the saliva of Control or tumor-bearing individuals (here called Disease) revealed several metabolic differences, some of which are highlighted below:

- **Overview of the dataset**: Principal component analysis (PCA) transforms a large number of metabolic variables into a smaller number of orthogonal variables (Component 1, Component 2, etc…) in order to analyze variation between groups and to provide a high-level overview of the dataset. Control and Disease samples tended to form partially overlapping populations, with Disease samples showing a wider spread across the PCA. *Increased spread could reflect position of the tumor within the oral cavity (nearness to sampling site), tumor stage or metastatic potential.* In the hierarchical clustering analysis (HCA), Control and Disease samples tended to cluster by disease status, though three control samples (1027, 1030, and 1031) formed a separate cluster. *These three samples appear to have a different numbering scheme compared to other control samples (which have a "D" in front of the number); it is possible that differential clustering may reflect an alternate collection site or protocol. These three samples also separated from the Control group in the PCA (highlighted with a red dashed line). A statistical analysis without these samples is also included in the Client Data Table; however, this report will focus on differences in the complete data set.*

  Random forest analysis (RFA) is a statistical tool utilizing a supervised classification technique based on an ensemble of decision trees (please see Appendix for greater detail) and can aid in the identification of biomarkers differentiating classification groups. RFA was very effective at separating Control from Disease samples, with a predictive accuracy of 100% (a 50% predictive accuracy would be expected by random chance). Interestingly, the three samples that showed separation from the Control population in the PCA and HCA were classified correctly. *The Top 30 metabolites for predicting treatment groups included biochemicals related to carbohydrate (3-phosphoglycerate, an isobar of fructose/glucose 1,6-diphosphate or myo-inositol 1,4 or 1,3-diphosphate, glucose-6-phosphate, and maltose) nucleotide (2',3'-cGMP, 3'-AMP, allantoin), and lipid (maleate, caproate, heptanoate, carnitine) metabolism.*

- **Energetics**: Glucose can be utilized to support a variety of physiological processes, including energy generation, fatty acid synthesis, protein glycosylation, and nucleotide biogenesis. While **glucose** levels were similar, glycolytic metabolites were increased as a class (though **lactate** was not significantly changed), potentially reflecting increasing glycolytic use in the tumor. The pentose phosphate metabolite **6-phosphogluconate** was also elevated, with decreased pentose products (which may reflect increased proliferative demand for nucleotides). Metabolites in the TCA cycle could indicate changing function: increased **citrate** could reflect increased glycolytic or beta-oxidative input, while elevated **alpha-ketoglutarate** could indicate increased glutaminolysis (**glutamine** levels were non-significantly decreased). *Lactate levels were increased in the related tumor samples (MICH-03-14VW), but was not increased in Disease saliva (compared to Control). One study has suggested that our bacterial flora can convert lactate into the short chain fatty acid (SCFA) butyrate (PMID: 15466518). The Metabolon discovery platform does not detect butyrate (though a surrogate molecule, butyrylcarnitine, was elevated); however valerate and several medium chain fatty acids (MCFAs) did show an increase. One attractive (and somewhat speculative) explanation for static levels of lactate might be conversion into SCFAs (and lower chain-length MCFAs) by resident microflora. A corollary would then be that cancers might change neighboring microbial communities by altering available energy sources. It is uncertain how this might affect mucosal biology, but given studies showing effects of certain bacteria on immune responses in the gut, there may be an effect on immune tumor surveillance.*

  <u>Glycogen metabolites</u>: Glycogen metabolites (maltopentaose, maltotetraose, and maltose) were higher in Disease (compared to Control), potentially suggesting increased glycogen mobilization to support tumor metabolism, though maltopentaose and maltotetraose experienced poor fill (these metabolites were detected in Disease with greater frequently than Control). *It is possible that metabolic changes associated with tumor progression underlies the appearance of glycogen metabolites in saliva. Further studies assessing stage/grade or molecular status could identify these metabolites as predictive biomarkers for disease.*

  <u>BCAA catabolism</u>: Metabolites derived from leucine, isoleucine or valine catabolism can enter gluconeogenesis or the TCA cycle for energy production; **leucine**, **isoleucine** and **valine** showed non-significant trends toward decrease (Disease vs Control), while increases in catabolic products of leucine (**isovalerate**, isovalerylcarnitine), isoleucine (**2-methylbutyrylcarnitine**, tiglyl carnitine), and valine (**3-methyl-2-oxobutyrate**, isobutyrylcarnitine) could indicate increased cellular use of BCAAs for energetics. Increases in propionylcarnitine (a surrogate reporter for propionyl CoA, one end-product of isoleucine

catabolism as well as the oxidation of odd-chain fatty acids) and 2-methylcitrate/homocitrate (both of which are formed by condensation of propionyl CoA with TCA intermediates) could also support increasing use of BCAAs for energetics. *The opposite pattern of metabolites was observed in the tumor matrix; this pattern in saliva could reflect changes in both BCAA use and secretion.*

Lipid metabolism: Fatty acids (FAs) are a critical source of energy for mitochondrial oxidation and cellular ATP generation. Short and medium-chain fatty acids were increased, while long-chain fatty acids showed non-significant trends toward decrease (potentially reflecting increased use for beta-oxidation). Long-chain FAs must be conjugated to carnitine for transport across the mitochondrial membrane prior to oxidation; acylcarnitine conjugates (**hydroxybutyrylcarnitine**, **hexanoylcarnitine**, **octanoylcarnitine**) were increased, with increases in **carnitine** and the ketone body **3-hydroxybutyrate** (**BHBA**) suggestive of increased beta-oxidative use.

Comparison of energetics in saliva and tumor matrices: Changes in salivary metabolites include contributions from both tumor and non-tumor populations, though several signatures of tumor metabolism were seen. Increases in glycolytic and BCAA catabolic metabolites are consistent with changes in tumor metabolism observed in MICH-03-14VW, while increased in beta-oxidative metabolites may reflect non-tumor (or stromal) metabolism. *Increased markers of beta-oxidative use may indicate a metabolic shift in tumor-adjacent stroma in response to high tumor glucose demand (or changes in SCFA/MCFA availability).* Glycogen metabolites were increased in saliva but decreased in tumor, which could suggest mobilization of glycogen reserves to support highly glycolytic tumor metabolism. Finally, TCA metabolites may show a hybrid effect: increasing citrate may indicate increasing beta-oxidative input in "normal" cells, while changes in alpha-ketoglutarate may reflect elevated BCAA catabolism in the tumor.

- **Inflammation-associated metabolites:** Changes in **tryptophan** metabolites can also indicate inflammatory states: indoleamine 2,3-dioxygenase (IDO), which catalyzes the conversion of tryptophan to kynurenine, is activated by pro-inflammatory cytokines (e.g., IFN-γ, TNF-α). **Kynurenine** was elevated, while its degradation product **kynurenate** was decreased, suggestive of increasing inflammation (Disease vs Control). *While frequently used as a biomarker of inflammation, kynurenine functions as an endogenous "brake" on immune activation and has been suggested as one mediator of tumor-induced immune suppression, with effects on invasive tumor growth (PMID: 21993754).* **Histamine** and **N-acetylhistamine** were non-significantly decreased, with a significant increase in the degradatory metabolite **1-methylimidazoleacetate**. *This signature was also apparent in tumor cells from MICH-03-14VW and could reflect changes in mast cell biology or inflammation.*

- **Redox Homeostasis:** Gamma-glutamyl AAs are generated by gamma-glutamyl transpeptidase (GGT), which transfers the gamma-glutamyl moiety of reduced glutathione (GSH) to an amino acid acceptor, modulating the intra- and extracellular exchange of GSH. Increases in several gamma-glutamyl AAs (e.g., gamma-glutamylalanine, gammaglutamylglutamate) could indicate increasing oxidative stress (Disease vs Control), while elevated **5-oxoproline** may reflect increased gamma-glutamyl amino acid exchange to replenish glutathione. **Oxidized glutathione** (**GSSG**) was increased in Disease (compared to Control), as were cysteine-glutathione disulfide (a product of glutathione oxidation), cys-gly, oxidized, and **ophthalmate** (a tripeptide analogue of glutathione also produced by glutathione synthetase that can be used as a marker of glutathione demand). *Note that many of these markers experienced poor fill in both Control and Disease groups, though they were detected with greater frequency in Disease samples (consistent with increasing oxidative stress seen in tumor samples, MICH-03-14VW).* A subset of **methionine** metabolites were also elevated, including **cysteine**, **hypotaurine** and **taurine** (potentially in support of glutathione production).

Other observations of interest:

- Heme: Heme levels were increased in the saliva of Disease (compared to Control); interestingly, this metabolite was below the threshold of detection in Control samples (while Disease showed 28% fill). *The presence of heme could indicate blood in the saliva, potentially derived from tumors in the oral cavity.*

- Nicotine: While nicotine levels were decreased in Disease (compared to Control), nicotine metabolites (cotinine, hydroxycotinine, and cotinine N-oxide) were not significantly altered. Since nicotine metabolism primarily occurs in the liver (PMID: 19184645), it is possible that lower nicotine reflects less recent use. Interestingly, levulinate, an additive that is used to increase nicotine binding to receptors, was increased in Disease (compared to Control). *Glycols (pentaethylene glycol, hexaethylene glycol, heptaethylene glycol and octaethylene glycol) were also increased; these could derive from medications (where PEGylation can improve solubility and decrease renal clearance) or from nicotine delivery (e-cigarettes can use polyethylene glycol as a solvent).*

- Drugs: The opioids oxycodone and its metabolites noroxycodone and oxymorphone were also detected in a subset of Disease samples (Disease vs Control). The presence of these drugs may indicate more advanced disease in the patients from which these samples were derived. *Individuals with detected oxycodone did not show good overlap with those where heme was detected.*

Conclusions

In conclusion, the results from this global metabolomic study comparing saliva samples from Control or Tumor-bearing subjects differed in a number of metabolic readouts, including changes in metabolites related to energetics, redox homeostasis, and inflammation. In the principal components analysis (PCA), Control and Disease formed overlapping populations, though Disease showed wider sample spread across the PCA. Interesting, three control samples separated from the rest of the Control population; statistical analysis is provided with and without these samples. Increases in glycolytic and TCA metabolites could reflect increased use of glycolysis and glutaminolysis in tumors; interestingly, markers of lipid metabolism pointed to increased beta-oxidative use (potentially reflecting changes in normal tissue or tumor stroma). Trends in the dataset also pointed to increasing oxidative stress and inflammation. Finally, heme was detected in a subset of Disease samples, which could indicate more advanced tumor stage or progression in these samples. Overall, changes in metabolites in saliva and tissue matrices (from the related project MICH-03-14VW) showed good correlation, suggesting altered biochemicals may be useful prognostic biomarkers of disease.

Study Parameters

Data Quality: Instrument and Process Variability

QC Sample	*Measurement*	*Median RSD*
Internal Standards	Instrument Variability	4 %
Endogenous Biochemicals	Total Process Variability	6 %

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 Client Matrix samples, which are technical replicates of pooled client samples. Values for instrument and process variability meet Metabolon’s acceptance criteria as shown in the table above.

Appendix

Metabolon Platform

Sample Accessioning: Following receipt, samples were inventoried and immediately stored at -80^{o}C. 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 -80^{o}C until processed.

Sample Preparation: Samples were prepared using the automated MicroLab STAR® system from Hamilton Company. A recovery standard was 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: one for analysis by UPLC-MS/MS with positive ion mode electrospray ionization, one for analysis by UPLC-MS/MS with negative ion mode electrospray ionization, one for analysis by UPLC-MS/MS polar platform (negative ionization), one for analysis by GC-MS, and one sample was reserved for backup. Samples were placed briefly on a TurboVap® (Zymark) to remove the organic solvent. For LC, the samples were stored overnight under nitrogen before preparation for analysis. For GC, each sample was dried under vacuum overnight 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.
DS	Derivatization Standard	Assess variability of derivatization for GC-MS samples.
IS	Internal Standard	Assess variability and performance of instrument.

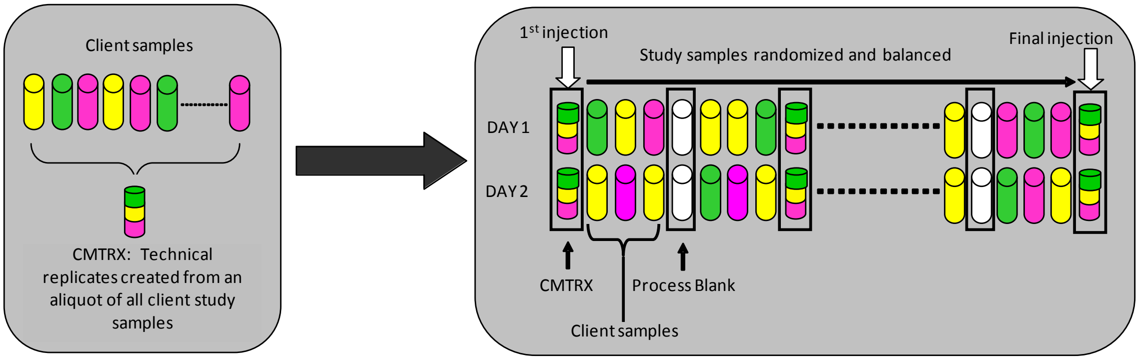

Client samples
CMTRX: Technical replicates created from an aliquot of all client study samples
1st injection
Study samples randomized and balanced
Final injection
DAY 1
DAY 2
CMTRX
Process Blank
Client samples

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): The LC/MS portion of the platform was based on 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 acidic or basic LC-compatible solvents, each of which contained 8 or more injection standards at fixed concentrations to ensure injection and chromatographic consistency. One aliquot was analyzed using acidic positive ion optimized conditions and the other using basic negative ion optimized conditions in two independent injections using separate dedicated columns (Waters UPLC BEH C18-2.1x100 mm, 1.7 µm). Extracts reconstituted in acidic conditions were gradient eluted from a C18 column using water and methanol containing 0.1% formic acid. The basic extracts were similarly eluted from C18 using methanol and water, however with 6.5mM Ammonium Bicarbonate. The third 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. The MS analysis alternated between MS and data-dependent MS^2 scans using dynamic exclusion, and the scan range was from 80-1000 *m/z*. Raw data files are archived and extracted as described below.

Gas Chromatography-Mass Spectroscopy (GC-MS): The samples destined for analysis by GC-MS were dried under vacuum for a minimum of 18 h prior to being derivatized under dried nitrogen using bistrimethyl-silyltrifluoroacetamide. Derivatized samples were separated on a 5% diphenyl / 95% dimethyl polysiloxane fused silica column (20 m x 0.18 mm ID; 0.18 um film thickness) with helium as carrier gas and a temperature ramp from 60° to 340°C in a 17.5 min period. Samples were analyzed on a Thermo-Finnigan Trace DSQ fast-scanning single-quadrupole mass spectrometer using electron impact ionization (EI) and operated at unit mass resolving power. The scan range was from 50–750 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 +/- 0.005 amu, 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 distribution to both the LC-MS and GC-MS 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.

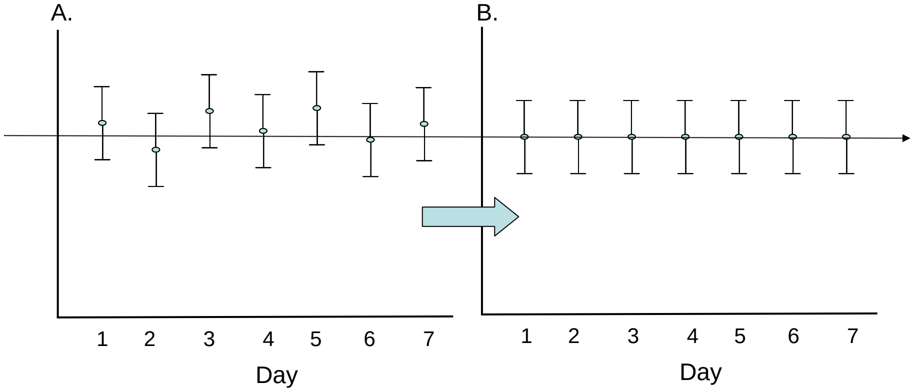


Figure 2: Visualization of data normalization steps for a multiday platform run.

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 $t = (\bar{x}_1 - \bar{x}_2)/n$, 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 2*n* 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

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

where $s_{AB}^2 = \left((n_A - 1)s_A^2 + (n_B - 1)s_B^2\right)/(n_A + n_B - 2)$, where the degrees of freedom is $n_A + n_B - 2$.
The contrast from the ANOVA,

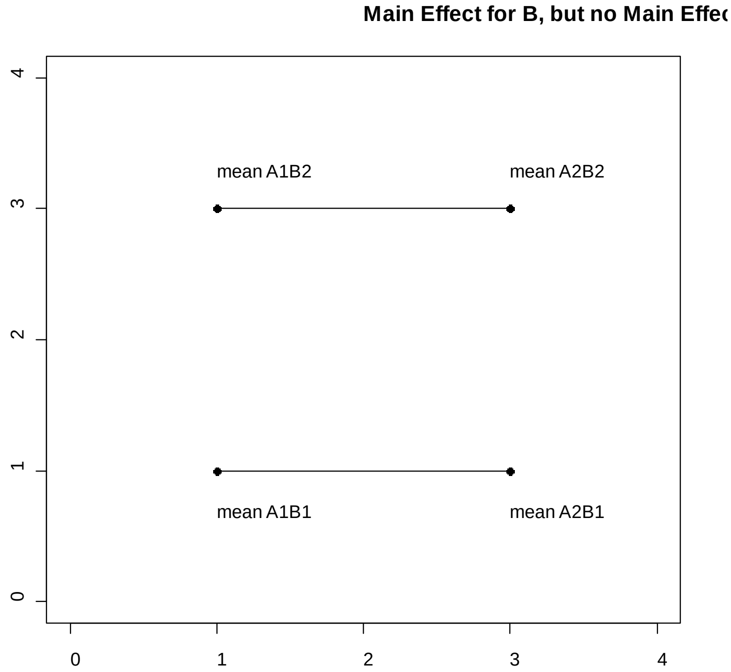
$$t = (\bar{x}_A - \bar{x}_B)/\sqrt{s^2(1/n_A + /n_B)}$$

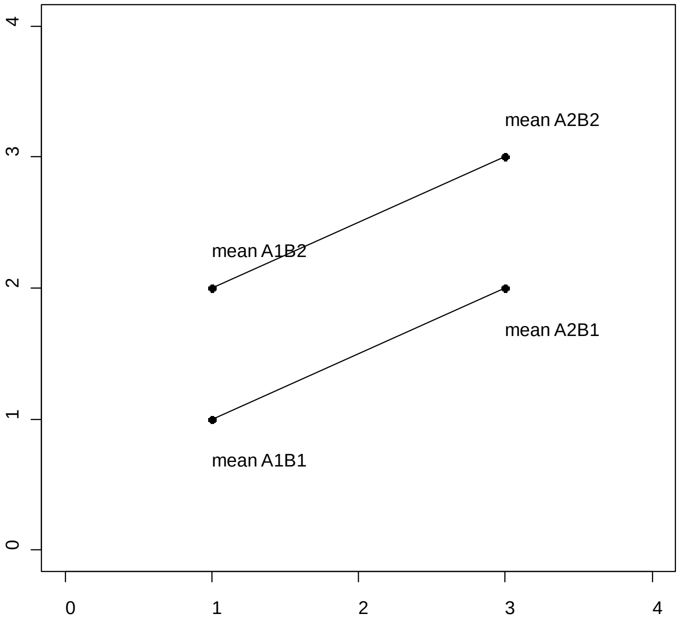
where $s^2 = \left((n_A - 1)s_A^2 + (n_B - 1)s_B^2 + (n_C - 1)s_C^2\right)/(n_A + n_B + n_C - 3)$, where the degrees of freedom is given by where the degrees of freedom is $n_A + n_B + n_C - 3$.

4. Two-way ANOVA

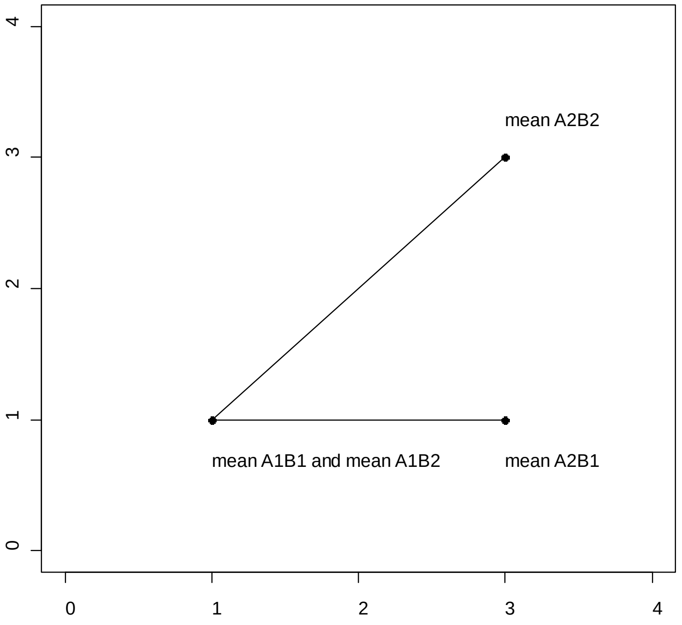
ANOVA stands for analysis of variance. For ANOVA, it is assumed that all populations have the same variances. For a two-way ANOVA, three statistical tests are typically performed: the main effect of each factor and the interaction. Suppose we have two factors A and B, where A represent the genotype and 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 A, Main Effect for B,



Main Effect for A, Main Effect for B,

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 T^2 test**

   The Hotelling's T^2 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: $t^2 \ = \left(\frac{n_x\, n_y}{n_x+n_y}\right) * (\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})^T\, \boldsymbol{S}^{-1}\, (\bar{\boldsymbol{x}} - \bar{\boldsymbol{y}})$, where n_x and n_y are the numbers of samples in each group, $\bar{\boldsymbol{x}}$ is the mean vector of the variables from group 1, $\bar{\boldsymbol{y}}$ 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).

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*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.

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.