

mSelect™ **FINAL REPORT**

Longitudinal Analysis of *Mycobacterium tuberculosis* Grown Under Hypoxic and Reaerated Conditions in a Fermentor Incubation Model (Sauton's Media)

CLIENT: Max Planck Institute for Infection Biology, Professor
Stefan Kaufmann

PROJECT CODE: MPIB-0203-09SLBL

AUTHOR: Adam D. Kennedy, PhD

APPROVAL: Robert Mohny, PhD

DATE: 12/27/2010



METABOLON

Metabolon, Inc. • 617 Davis Drive, Suite 400, Durham, NC 27713 • (919) 572-1711 www.metabolon.com •
busdev@metabolon.com

Table of Contents

Table of Contents.....	2
Objective	3
Experimental Design	3
Summary of Procedure	4
Data Display	4
Biological Summary of Data.....	5
Path Forward.....	16
Supplemental Information.....	16
Methods.....	18
<i>Sample Preparation</i>	18
<i>Data Collection and Normalization</i>	18
<i>Process Evaluation and Compound Summary</i>	19
<i>Data Analysis</i>	19
Appendix A: Metabolon Platform.....	20
Appendix B: Statistical Terminology	24

Objective

The purpose of this study was to biochemically profile *Mycobacterium tuberculosis* cultured under hypoxic and reaerated conditions in a fermentor model. A secondary objective was to identify biochemicals that were differentially released into the culture media and/or consumed from the media.

Experimental Design

Mycobacterium tuberculosis infects nearly one third of the world's population and is responsible for nearly 3 million deaths per year. In sub-Saharan Africa, many HIV+ individuals are also infected with *M. tuberculosis* due to depressed immune systems as well as the ease of transmission of the *M. tuberculosis* pathogen. *M. tuberculosis* infects the lungs and has the ability to grow and persist in oxygen-rich or hypoxic environments. In order to evade detection by the human immune system, the bacteria cluster to form granulomas in the lungs. The environment of the granuloma is depleted of nutrients including oxygen resulting in a hypoxic environment.

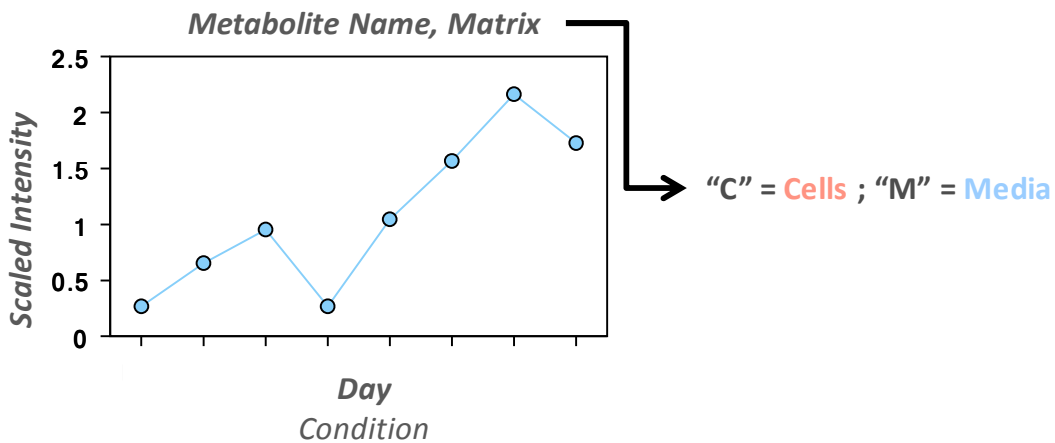
M. tuberculosis was studied under hypoxic and reaerated conditions in a fermentor reactor. A cell pellet sample and spent culture media were harvested on days 0, 1, 2, 3, 5, 7, 8, and 9. The sample on day 1 represented the first sample collected under hypoxic conditions, and reaeration was begun after the day 7 sample was harvested. Samples from days 8 and 9 represent reaerated samples. Global metabolic profiling was performed on all samples in order to study the microorganism at a biochemical level.

Summary of Procedure

A detailed description of the process for sample preparation and data acquisition is in the “Methods” section and Appendix A. In brief, samples were extracted and split into equal parts for analysis on the GC/MS and LC/MS/MS platforms. Proprietary software was used to match ions to an in-house library of standards for metabolite identification and for metabolite quantitation by peak area integration. The QC metrics for this study data, a summary of the total number of biochemicals detected and the number declared significant are reported in the “Methods Section”. All datasets were provided in electronic form to the client.

Data Display

Data obtained in this study were visualized using the program Array Studio from OmicSoft. An interpretive guide for this is shown below.



Biological Summary of Data

As *M. tuberculosis* enters a hypoxic environment, it is logical to presume that the amount of aerobic respiration would decrease. The carbon source for Sauton's media is glycerol which, after several metabolic steps, enters the glycolytic pathway at 3-phosphoglycerate. The amount of 3-phosphoglycerate in cells increased after the onset of hypoxia but then rapidly decreased from days 3 to 5 to undetectable levels (**Figure 1A**). With the exception of glucose, the hexoses decreased during hypoxia and none of the intermediates increased upon re-aeration of the culture. The levels of the glycolytic intermediates did not change from day 0 to day 9 suggesting that the hexose species may have been utilized for cellular processes other than energy production (e.g., cell wall synthesis, protein modification). Additionally, glucose and 1,3-dihydroxyacetone were detected in the spent media (**Figure 1B**). These biochemicals could accumulate in the spent media due to release from the cell wall (glucose) or export of biochemicals from the cells due to regulatory processes (glucose, 1,3-dihydroxyacetone).

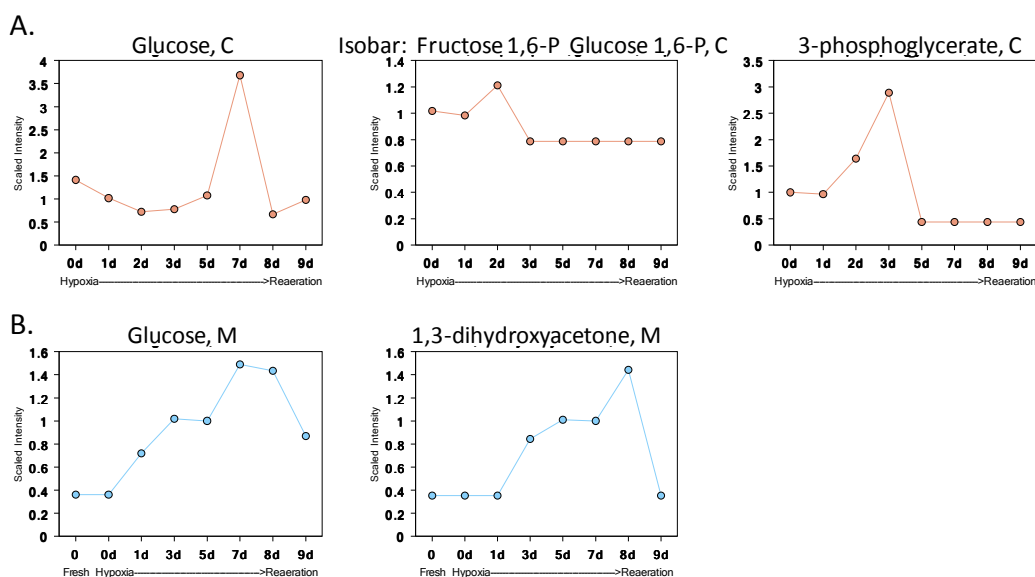


Figure 1. Time-dependent changes in glycolytic intermediates in (A) cells and (B) media.

To alleviate the loss of carbon atoms, *M. tuberculosis* utilizes the glyoxylate cycle which replenishes the intermediates of the citric acid cycle when the primary carbon source in the growth substrate is fatty acids. The glyoxylate cycle converts acetyl-CoA molecules derived from pyruvate, amino acids, and the β -oxidation of fatty acids into intermediates of the citric acid cycle.

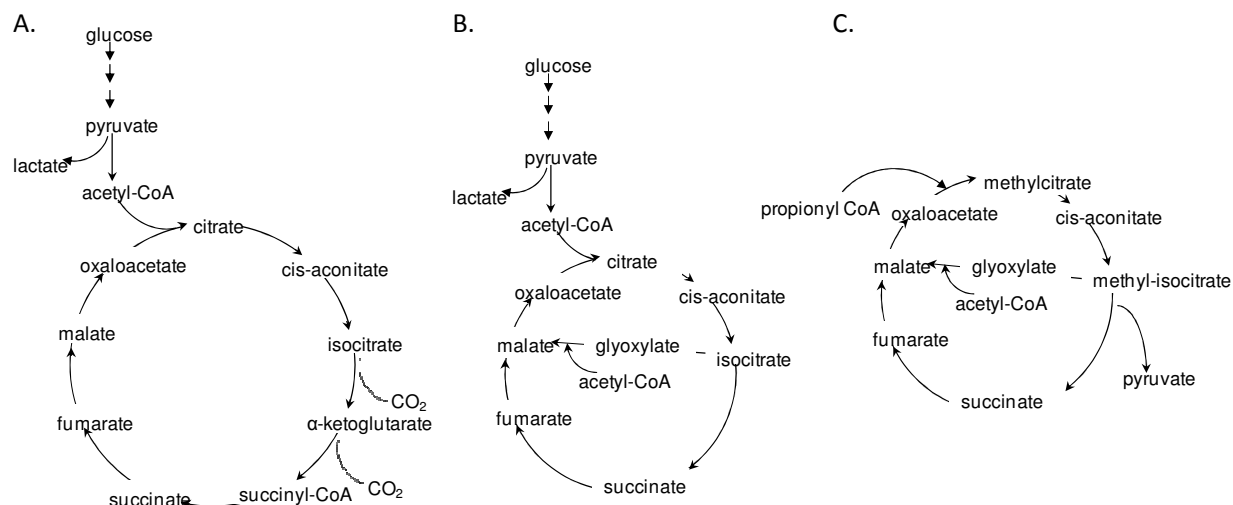
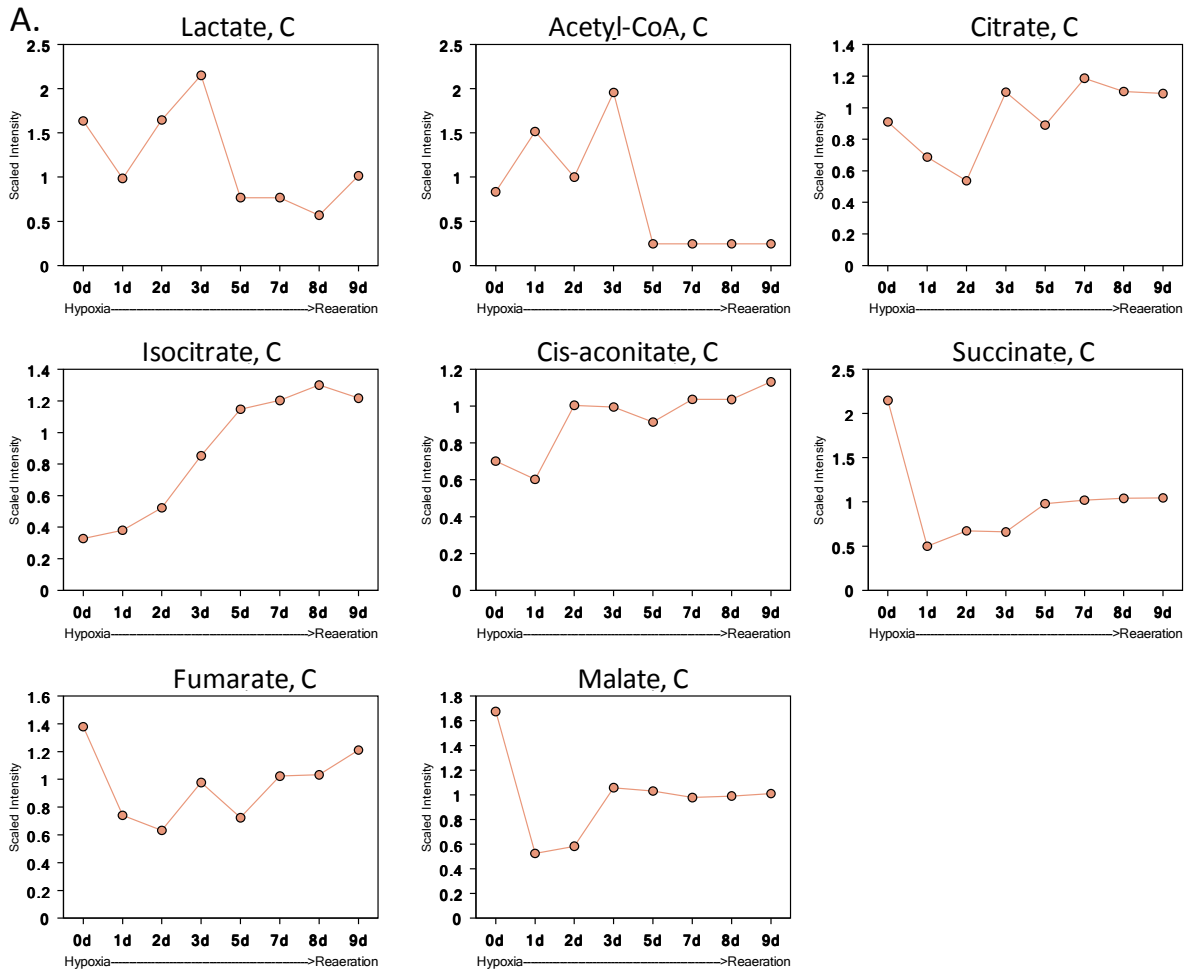


Figure 2. Three pathways for carbon metabolism in *M. tuberculosis*. (A) The TCA cycle, (B) The glyoxylate shunt, (C) The methylcitrate cycle.

M. tuberculosis grown in the fermentor model with Sauton's media in this bioprocessing run showed nearly the same trends as the past experiment (see MPIB-0202-09VSBL), except for succinate and the detection of fumarate (**Figure 3A**). The overall profile for succinate was similar although succinate did reach higher relative levels in MPIB-0202-09VSBL. Malate levels remained relatively constant after day 1 whereas acetyl CoA levels showed a general decrease from days 1 through day 9. The decrease in the amount of acetyl CoA would be consistent with lower levels of β -oxidation of fatty acids. This is discussed later in the report. Although unlikely due to the glycerol in the media as the carbon source, another possible explanation for the reduced malate levels is that malate may have been used in gluconeogenic reactions to produce pyruvate. In order to establish and maintain infection, *M. tuberculosis* relies on gluconeogenic carbon flow from the TCA cycle/glyoxylate shunt.

Many of the metabolites of the glyoxylate shunt also were detected in the spent cell culture media (**Figure 3B**). Several of these metabolites showed increased levels throughout the time course. These intermediates may be present in the culture media for the following reasons. The first is that these biochemicals may be released from dead cells or the live cells may be actively secreting these compounds. *M. tuberculosis* has a thick cell envelope and the disintegration of the envelope and the release of these biochemicals from dead cells is unlikely. Alternatively, *M. tuberculosis* may have active transporters that act to secrete these biochemicals when they reach significant levels inside cells. A third possibility is that the enzyme machinery expressed by *M. tuberculosis* may be present in the media and the glyoxylate cycle may be continuing extracellularly. For example malate levels largely increased in the media but remained at static levels in the cells post day 3. It is conceivable that the bacterium may be actively secreting these biochemicals because the formation of granuloma-like clumps of bacteria would cause differential survival activity. The clumps of bacteria may produce protein factors responsible for the production and secretion of biochemicals that could be utilized by smaller clumps or single bacteria in the culture.



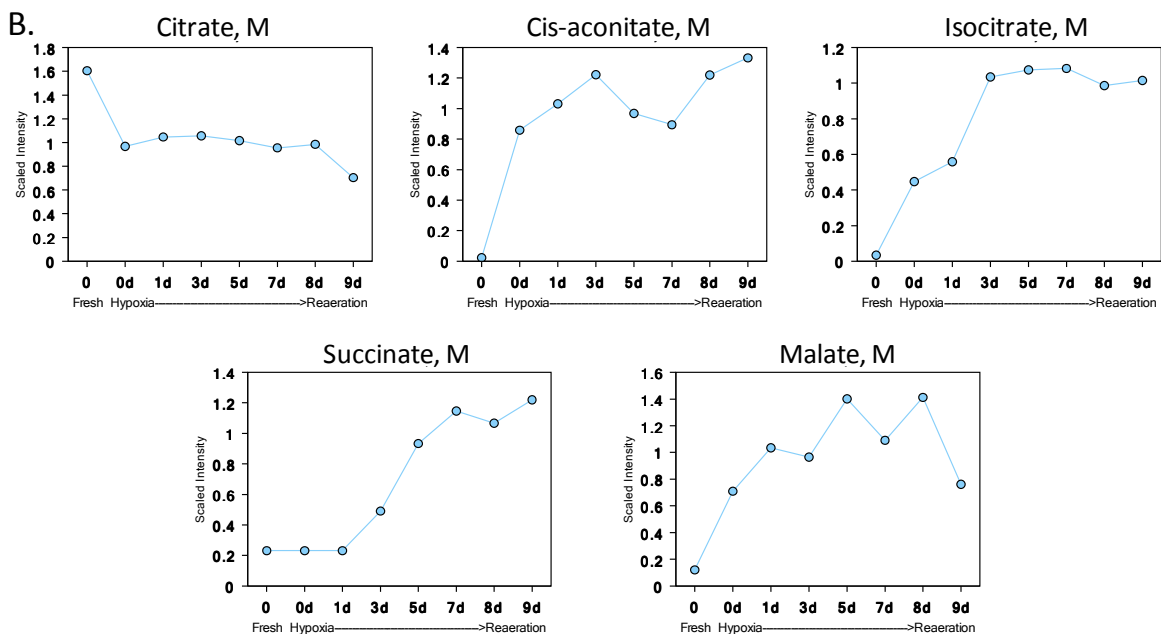


Figure 3. Time-dependent changes in glyoxylate cycle intermediates in (A) cells and (B) cell culture media.

As the bacterium forms granulomas, the cell membrane and envelope will reorganize, possibly resulting in the liberation of fatty acids. The decrease in the amounts of medium-chain and long-chain fatty acids from days 3 to 5 suggests that fatty acids may have been an energy source during this time period of the incubation (**Figure 4A** and **4B**). The accumulation of glycerol within cells can be determined as a sign of β -oxidation of fatty acids. However, glycerol was the carbon source for Sauton's media making it difficult to determine if β -oxidation was occurring. The relative abundance of the medium-chain fatty acids increased during the onset of hypoxia and then rapidly decreased from days 3 to 5 and then showed a slight increase during reaeration of the culture (days 8 to 9) (**Figure 4**). Additionally, long-chain unsaturated fatty acids reached maximal abundance at day 5 and then decreased sharply on day 7 (**Figure 4C**). This suggests that if lipids were a source of energy during hypoxia that saturated fatty acids were utilized first and then unsaturated fatty acids were utilized next. A possible source of the medium- and long-chain fatty acids is from the cellular membrane and envelope of *M. tuberculosis*. Alternatively, *M. tuberculosis* may be synthesizing fatty acids during hypoxia. Fatty acids may be needed for the reorganization of the cell membrane and envelope as the bacterium forms granulomas necessitating the production of long-chain fatty acids. The profiles of the medium-chain fatty acids were different than the profiles from MPIB-0202-09VSBL, but the same medium-chain fatty acids were detected in both experiments.

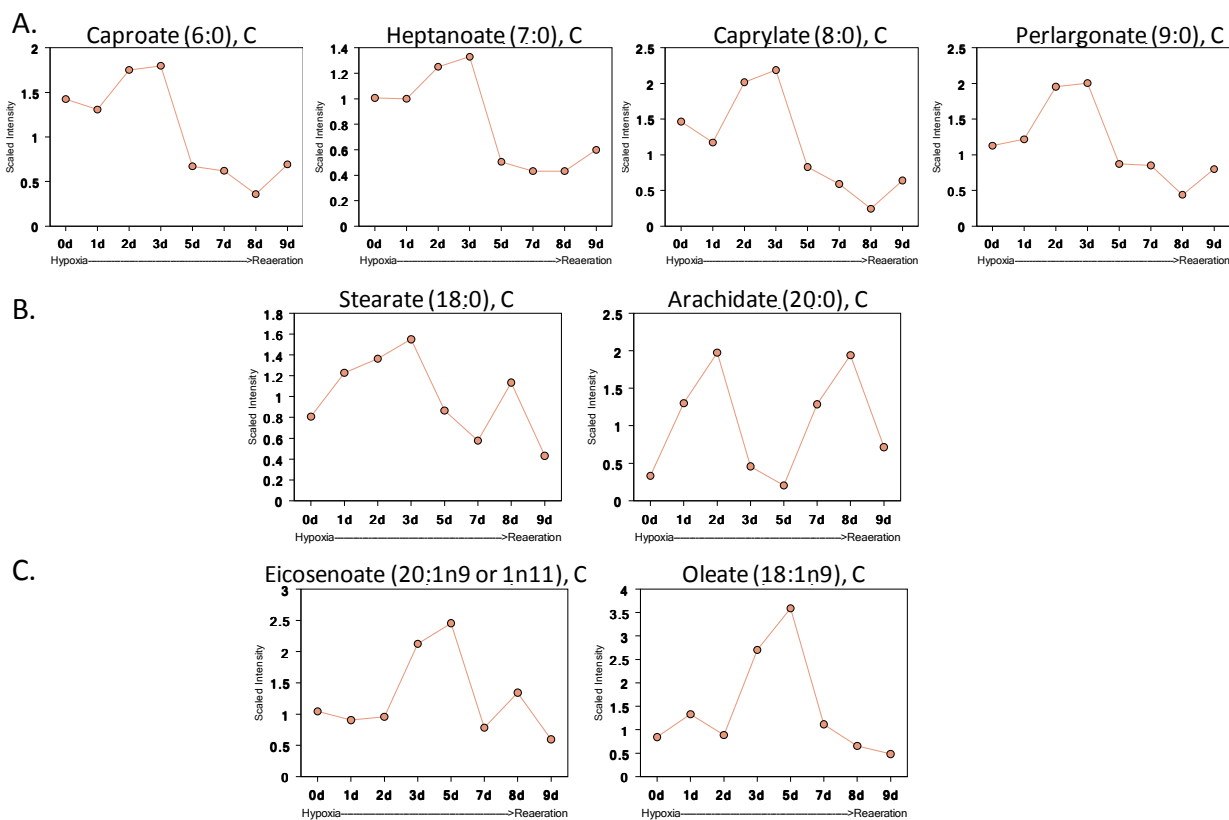


Figure 4. Time-dependent changes in intracellular levels of saturated and unsaturated fatty acids. (A) Medium-chain saturated fatty acids, (B) Long-chain saturated fatty acids, (C) Long-chain unsaturated fatty acids.

β -oxidation of fatty acids results in the production of acetyl-CoA, however metabolism of odd chain-length fatty acids or branch-chain fatty acids results in the formation of propionyl-CoA. Propionyl-CoA, and its metabolites, can be toxic if present in large amounts and *M. tuberculosis* contains the metabolic machinery to metabolize propionyl-CoA (**Figure 2C**). The methylcitrate cycle metabolizes propionyl-CoA in order to prevent the accumulation of this compound. Valproic acid and 2-ethylhexanoic acid (structural isomers) are two branched-chain fatty acids and β -oxidation of these compounds would result in the generation of propionyl-CoA. These compounds showed the same relative profile as other medium-chain fatty acids in that there was a slight increase during the initial days of hypoxia followed by a sharp decrease from days 3 to 8 further suggesting that fatty acids, and in this case a branched-chain fatty acid, was being utilized as a source of energy (**Figure 5**). After reaeration of the culture, there was a small increase in this compound(s) on day 9.

Isobar: valproic acid, 2-ethylhexanoic acid

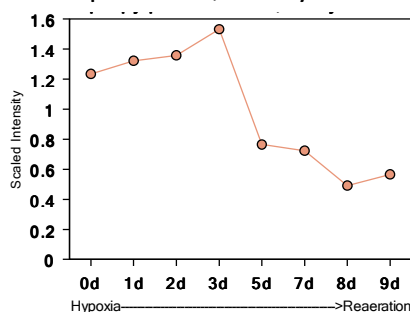


Figure 5. Time-dependent changes in branched-chain fatty acid(s).

In addition to the alterations in the lipid profile, trehalose accumulated in the culture media reaching a maximum at day 7 and did not change in cells after reaeration of the culture (**Figure 6**). Trehalose is composed of two alpha-linked glucose molecules and it is a principle component of glycolipids in the *M. tuberculosis* cell envelope. Together with the alterations in the lipid profile, the liberation of trehalose may result from a reorganization of the cellular envelope as the bacteria are forming granuloma-like cellular clumps. Concomitantly with the increase in trehalose in the media, was an increase in glucose. Inasmuch as glycerol is the primary carbon source in the media, glucose could accumulate in the media from gluconeogenesis or from the liberation of carbohydrates from the cell envelope. As phosphoenolpyruvate, an intermediate of the gluconeogenic pathway, was not detected in the study, the more likely explanation for the accumulation of glucose in the media is from the liberation of trehalose from the bacterial cell envelope.

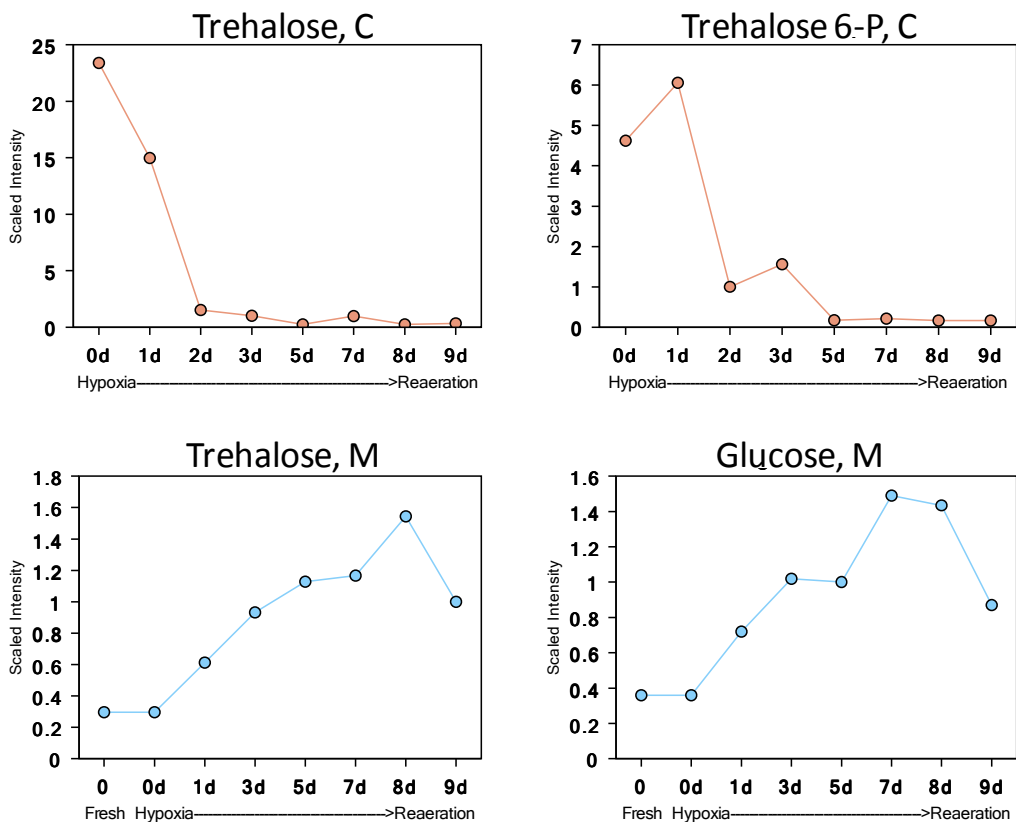


Figure 6. Time-dependent alterations in trehalose in cells and media.

Although actinomycetes such as *M. tuberculosis* do not generate the redox balance compound glutathione, they are capable of producing mycothiol (MSH) which plays a role analogous to that of glutathione. It has been suggested that MSH has a critical function in mycobacterial survival, and the compound may play a role in dormancy. Although MSH was not detected in these samples, mycothione (MSSM), the oxidized form of MSH, was detected (**Figure 7**). The levels of mycothione in the cells were sharply reduced during hypoxia and did not increase upon aeration of the culture. Consistently, the MSH precursor myo-inositol was concomitantly decreased during hypoxia (**Figure 7**). This suggests that mycothione levels may have decreased due to the decrease in precursors, or the level of mycothiol may have increased during hypoxia and the mycothiol formed complexes with other molecules.

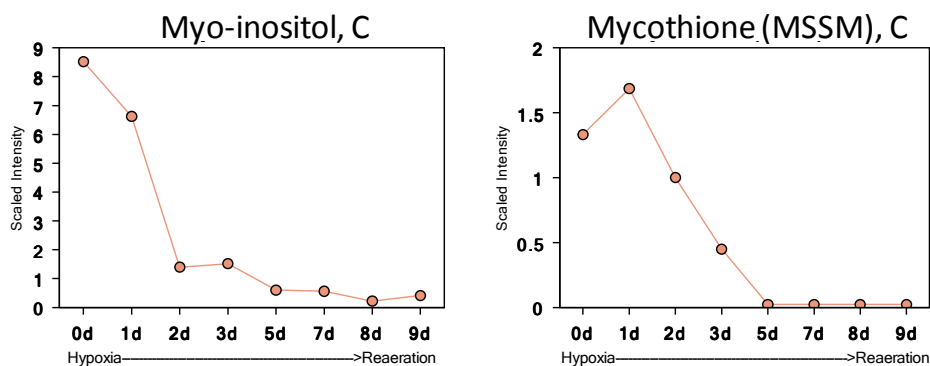


Figure 7. Time-dependent changes in mycothione and a metabolite precursor myo-inositol.

The longitudinal profiles of amino acids for the fermentor model are shown in **Figures 8 and 9**. With the exception of aspartate, which increased in abundance during the cell culture period, and the branched-chain amino acids (valine leucine, and isoleucine) which showed no overall change in abundance, the other amino acids decreased in amount during hypoxia and levels did not increase upon reaeration of the culture. Lysine, serine, glutamine, tyrosine, glutamate, and glycine showed the greatest amount of change during the cell culture period. Levels of each of these amino acids decreased ~3- to 6-fold over the nine days of culture and did not increase upon aeration of the culture. The decrease in the abundance of amino acids could happen because the amino acids are being utilized as a source of energy. Second to carbohydrates, amino acids are utilized by cells as they are metabolized into intermediates of the TCA cycle/glyoxylate shunt. The decrease in amino acids happens rapidly meaning that if they are utilized as a source of energy, they are metabolized quickly after the onset of hypoxia. Alternatively, the amino acids could be utilized for protein production. As the cells enter hypoxia, there may be a demand for protein synthesis which would decrease the overall pool of amino acids.

In addition to the decrease in the intracellular levels of amino acids, all of the amino acids detected in the experiment, with the exceptions of asparagine, glutamine, and lysine showed profound increases in the spent media. The increase in amino acids could be the result of protein degradation and export of amino acids from the cells. This export of amino acids also fits with the decrease in intracellular amounts of several of amino acids. Additionally, the large amount of extracellular amino acids could result from the presence of proteases secreted by the bacteria which, in turn, would cleave extracellular proteins. Asparagine is the nitrogen source for Sauton's media and it is reasonable to think that its level would not change. Glutamine is utilized by cells as a major carbon source as it can be metabolized to glutamate, which is further metabolized to the TCA cycle intermediate α -ketoglutarate. Although there is a loss of a loss of one CO_2 from α -ketoglutarate to succinyl-CoA, the utilization of glutamine as a carbon source would be significant if carbohydrate sources are limiting.

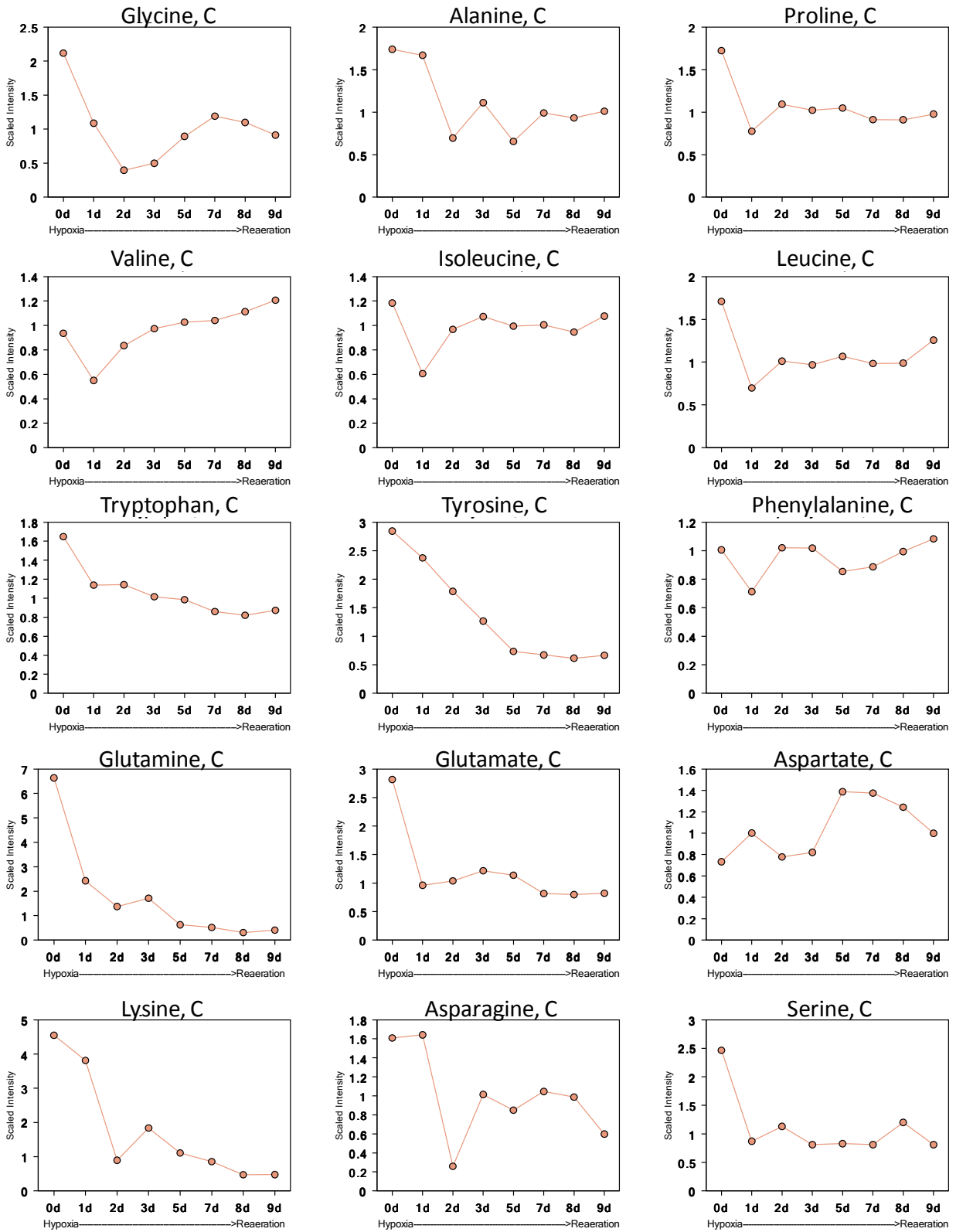


Figure 8. Time-dependent alterations of amino acids in *M. tuberculosis*.

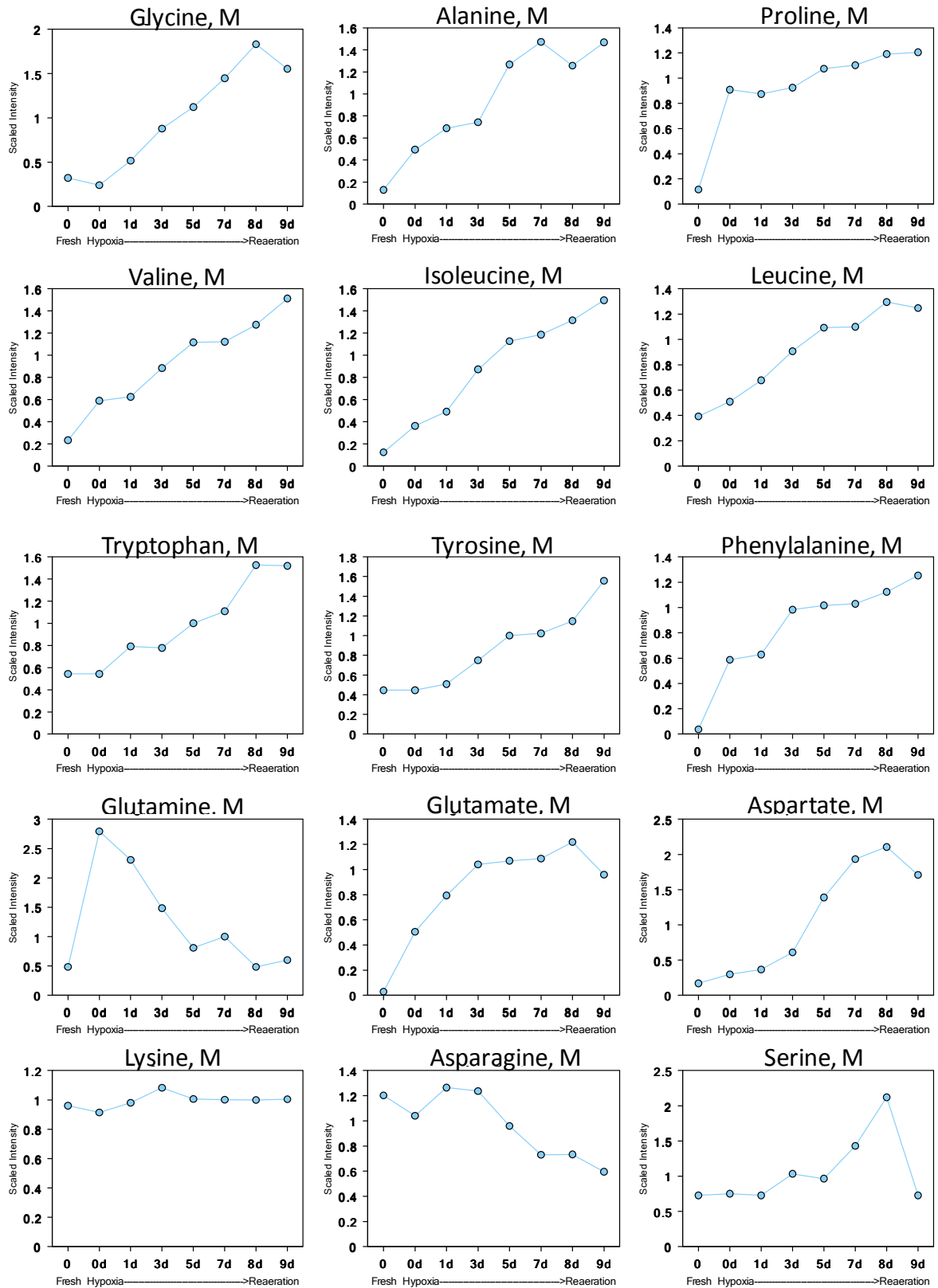


Figure 9. Longitudinal profiles for the amino acids detected in cell culture media.

Gamma-glutamyl amino acids decreased at the onset of hypoxia and continued to decrease throughout the entire incubation period (**Figure 10**). Except for γ -glutamyltryptophan, the same γ -glutamyl amino acids were detected as in project MPIB-0202-09VSB. Amino acids are post-translationally modified with glutamate creating γ -glutamyl amino acids. This modification allows for amino acids to be transported across membranes without the need for specific, energetically expensive amino acid transporters. The dramatic decrease in γ -glutamyl amino acids could suggest that amino acid levels in the cells were not high or that the demand to transport amino acids into the cells was not high. Cells may harvest amino acids by degrading existing proteins and recycling the component amino acids.

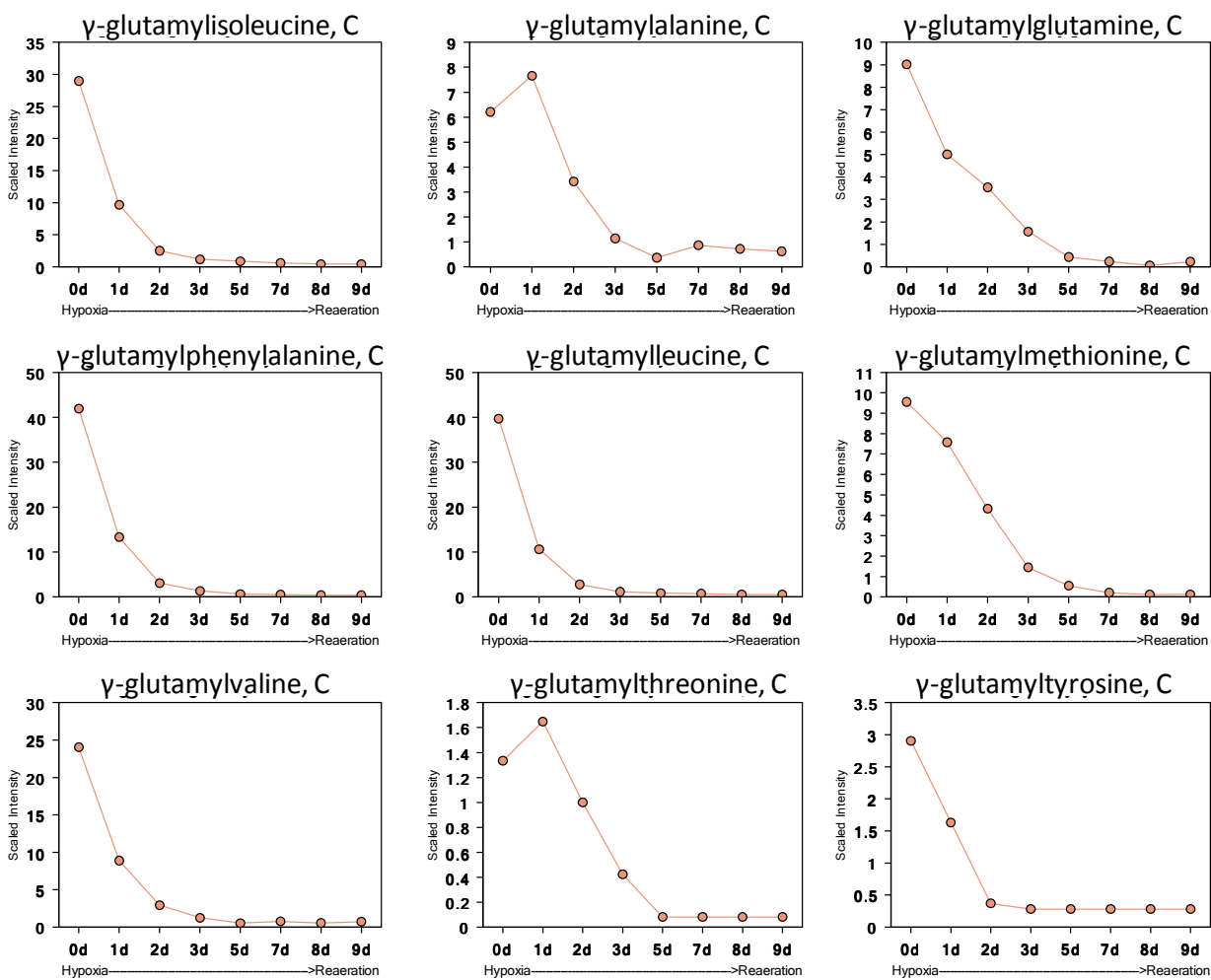


Figure 10. Gamma-glutamyl amino acid profiles.

Path Forward

In order to advance the understanding gained from the results of the current project, the following suggestions are a path forward for future experiments.

- Repeat the fermentor model using multiple replicates for each of the time points. This will decrease any variation that may be due to the low number of replicates per time point.
- Increase the number and duration of post-aeration time points. *M. tuberculosis* is a relatively slow growing pathogen and increasing the time post aeration may result in the identification of additional and more significant changes due to the oxygen status of the cell culture. For example, obtain samples 3 to 5 days post aeration (days 10 to 12 of the culture) to determine the levels of compounds after the bacteria has adjusted to an oxygenated environment.

Supplemental Information

A. Raw data files, statistics and plots

B. Key References:

Timm, J., Post, F.A., Bekker, L.-G., Walther, G.B., Wainwright, H.c., Manganelli, R., Chan, W.-T., Tsenova, L., Gold, B., Smith, I., et al. 2003. Differential expression of iron-, carbon- and oxygen-responsive mycobacterial genes in the lungs of chronically infected mice and tuberculosis patients. *Proc Natl Acad Sci U S A* 100:14321-14326.

Kim, M.-J., Wainwright, H.C., Locketz, M., Bekker, L.-G., Walther, G.B., Dittrich, C., Visser, A., Wang, W., Hsu, F.-F., Wiehart, U., et al. 2010. Caseation of human tuberculosis granulomas correlates with elevated host lipid metabolism. *EMBO Mol Med.* 2:258-274.

Russell, D.G. 2003. Phagosomes, fatty acids and tuberculosis. *Nat Cell Biol* 5:776-778.

Marrero, J., Rhee, K.Y., Schnappinger, D., Pethe, K., and Ehrt, S. 2010. Gluconeogenic carbon flow of tricarboxylic acid cycle intermediates is critical for *Mycobacterium tuberculosis* to establish and maintain infection. *Proc Natl Acad Sci U S A* 107:9819-9824.

Munoz-Elias, E.J., Upton, A.M., Cherian, J., and McKinney, J.D. 2005. Role of the methylcitrate cycle in *Mycobacterium tuberculosis* metabolism, intracellular growth, and virulence. *Mol. Microbiology* 60:1109-1122.

Minnikin, D.E., Kremer, L., Dover, L.G., and Besra, G.S. 2002. The methyl-branched fortifications of *Mycobacterium tuberculosis*. *Chem Biol* 9:545-553.

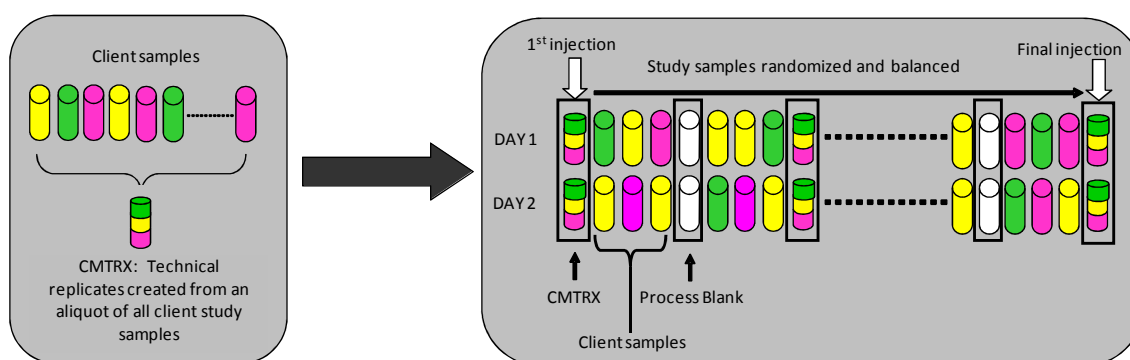
Gould, T.A., van de Langemheen, H., Munoz-Elias, E.J., McKinney, J.D., and Sacchettini, J.C. 2006. Dual role of isocitrate lyase 1 in the glyoxylate and methylcitrate cycles in *Mycobacterium tuberculosis*. *Mol Microbiol* 61:940-947.

Nambi, S., Basu, N., and Visweswariah, S.S. 2010. cAMP-regulated protein lysine acetylases in mycobacteria. *J. Biol. Chem.* 285:24313-24323.

Methods

Sample Preparation

At the time of analysis, samples were thawed and extracts prepared according to Metabolon's standard protocol, which is designed to remove protein, dislodge small molecules bound to protein or physically trapped in the precipitated protein matrix, and recover a wide range of chemically diverse metabolites. As shown in the figure below, a separate aliquot of each experimental plasma sample was taken then pooled for the creation of "Client Matrix" (CMTRX) samples. These CMTRX samples were injected throughout the platform run and served as technical replicates, allowing variability in the quantitation of all consistently detected biochemicals to be determined and overall process variability and platform performance to be monitored. Extracts of all experimental and CMTRX samples were split for analysis on the GC/MS and LC/MS/MS platforms (see APPENDIX A).



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.

Data Collection and Normalization

The CMTRX technical replicate samples were treated independently throughout the process as if they were client study samples. All process samples (CMTRX, GROBs – a mixture of organic components used to assess GC column performance, process blanks, etc.) were spaced evenly among the injections for each day and all client samples were randomly distributed throughout each day's run. Data were collected over one platform run day. Missing values (if any) were assumed to be below the level of detection for that biochemical with the instrumentation used and were imputed with the observed minimum for that particular biochemical.

Process Evaluation and Compound Summary

A number of internal standards were added to each experimental and process standard sample just prior to injection into the mass spectrometers. A measure of the platform variability was determined by calculating the median relative standard deviation (RSD) for these internal standards. The table below shows the median relative standard deviation (RSD) for the internal standards. Because these standards are added to the samples immediately prior to injection into the instrument, this value reflects instrument variation. In addition, the median relative standard deviation (RSD) for the biochemicals that were consistently measured in the CMTRX represents the total variability within the process for the actual experimental samples and the variability in quantitation of the endogenous metabolites within these samples. Results for the CMTRX and internal standards indicated that the platform produced data that met process specifications.

QC statistics for this study

Quality Control Sample (Matrix)	Median RSD	
	Cells	Media
Internal Standards	5%	4%
Endogenous Biochemicals	12%	8%

Data Analysis

There was an n=1 for each time point for fresh media, cells and spent media. Fresh media was run as a technical triplicate in order to produce a robust initial assessment of the media. All data was graphed as a line plot and no statistical analyses were performed due to the single sample per time point.

Appendix A: Metabolon Platform

Sample Accessioning: Each sample received was accessioned into the Metabolon LIMS system and was assigned by the LIMS a unique identifier, which 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 bar-coded and 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: The sample preparation process was carried out using the automated MicroLab STAR® system from Hamilton Company. Recovery standards were added prior to the first step in the extraction process for QC purposes. Sample preparation was conducted using a proprietary series of organic and aqueous extractions to remove the protein fraction while allowing maximum recovery of small molecules. The resulting extract was divided into two fractions; one for analysis by LC and one for analysis by GC. Samples were placed briefly on a TurboVap® (Zymark) to remove the organic solvent. Each sample was then frozen and dried under vacuum. Samples were then prepared for the appropriate instrument, either LC/MS or GC/MS.

QA/QC: For QA/QC purposes, a number of additional samples are included with each day's analysis. Furthermore, a selection of QC compounds is added to every sample, including those under test. These compounds are carefully chosen so as not to interfere with the measurement of the endogenous compounds. Tables 1 and 2 describe the QC samples and compounds. These QC samples are primarily used to evaluate the process control for each study as well as aiding in the data curation.

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 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
DS	Derivatization Standard	Assess variability of derivatization for GC/MS samples.
IS	Internal Standard	Assess variability and performance of instrument.
RS	Recovery Standard	Assess variability and verify performance of extraction and instrumentation.

Liquid chromatography/Mass Spectrometry (LC/MS, LC/MS²): The LC/MS portion of the platform was based on a Waters ACQUITY UPLC and a Thermo-Finnigan LTQ mass spectrometer, which consisted of an electrospray ionization (ESI) source and linear ion-trap (LIT) mass analyzer. The sample extract was split into two aliquots, dried, then reconstituted in acidic or basic LC-compatible solvents, each of which contained 11 or more injection standards at fixed concentrations. 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. Extracts reconstituted in acidic conditions were gradient eluted using water and methanol both containing 0.1% Formic acid, while the basic extracts, which also used water/methanol, contained 6.5mM Ammonium Bicarbonate. The MS analysis alternated between MS and data-dependent MS² scans using dynamic exclusion.

Gas chromatography/Mass Spectrometry (GC/MS): The samples destined for GC/MS analysis were re-dried under vacuum desiccation for a minimum of 24 hours prior to being derivatized under dried nitrogen using bistrimethyl-silyl-trifluoroacetamide (BSTFA). The GC column was 5% phenyl and the temperature ramp is from 40° to 300° C in a 16 minute period. Samples were analyzed on a Thermo-Finnigan Trace DSQ fast-scanning single-quadrupole mass spectrometer using electron impact ionization. The instrument was tuned and calibrated for mass resolution and mass accuracy on a daily basis. The information output from the raw data files was automatically extracted as discussed below.

Accurate Mass Determination and MS/MS fragmentation (LC/MS), (LC/MS/MS): The LC/MS portion of the platform was based on a Waters ACQUITY UPLC and a Thermo-Finnigan LTQ-FT mass spectrometer, which had a linear ion-trap (LIT) front end and a Fourier transform ion cyclotron resonance (FT-ICR) mass spectrometer backend. For ions with counts greater than 2 million, an accurate mass measurement could be performed. Accurate mass measurements could be made on the parent ion as well as fragments. The typical mass error was less than 5 ppm. Ions with less than two million counts require a greater amount of effort to characterize. Fragmentation spectra (MS/MS) were typically generated in data dependent manner, but if necessary, targeted MS/MS could be employed, such as in the case of lower level signals.

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 Quality Assurance: The data extraction of the raw mass spec data files yielded information that could be loaded into a relational database and manipulated without resorting to BLOB manipulation. Once in the database the information was examined and appropriate QC limits were imposed. Peaks were identified using Metabolon's proprietary peak integration software, and component parts were stored in a separate and specifically designed complex data structure.

Compound identification: Compounds were identified by comparison to library entries of purified standards or recurrent unknown entities. Identification of known chemical entities was based on comparison to metabolomic library entries of purified standards. As of this writing, more than 1800 commercially available purified standard compounds had been acquired and registered into LIMS for distribution to both the LC and GC platforms for determination of their analytical characteristics. The combination of chromatographic properties and mass spectra gave an indication of a match to the specific compound or an isobaric entity. Additional entities could be 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.

Normalization: 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 1). For

studies that did not require more than one day of analysis, no normalization is necessary, other than for purposes of data visualization.

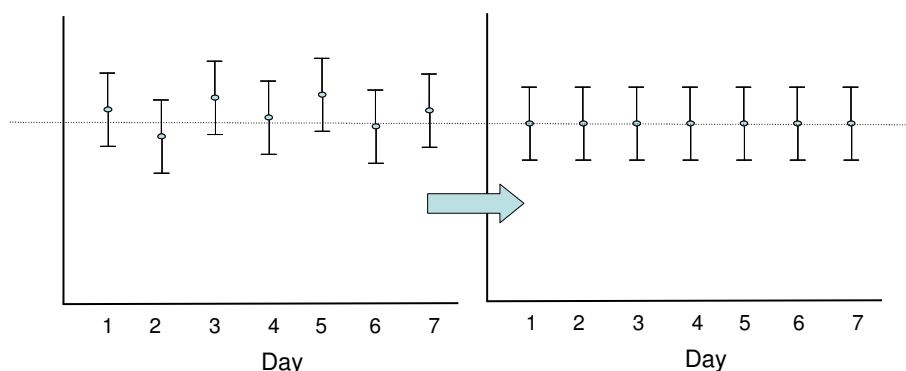


Figure 1. Visualization of Data Normalization

Statistical Calculation: For many studies, two types of statistical analysis are usually performed: (1) significance tests and (2) classification analysis. (1) For pair-wise comparisons we typically perform Welch’s t-tests and/or Wilcoxon’s rank sum tests. For other statistical designs we may perform various ANOVA procedures (e.g., repeated measures ANOVA). (2) For classification we mainly use random forest analyses. Random forests give an estimate of how well we can classify *individuals* in a *new* data set into each group, in contrast to a t-test, which tests whether the unknown means for two populations are different or not. Random forests create a set of classification trees based on continual sampling of the experimental units and compounds. Then each observation is classified based on the majority votes from all the classification trees. Statistical analyses are performed with one or both of the statistical analysis software programs: Array Studio (Omicsoft, Inc) or “R” <http://cran.r-project.org/>.

Appendix B: Statistical Terminology

***t*-tests:** *t*-tests test whether the unknown means for two populations are different or not. The *p*-value gives the amount evidence that the population means are different based on the data (through the *t*-statistic). The smaller the *p*-value, the more evidence we have that the population means are different. Often, a significance level of 0.05 is used. When the *p*-value is less than 0.05, we have enough evidence to conclude that the population means are different (“statistical significance”). The level of 0.05 is the false positive rate. This means that 5% of the time, the *t*-test would incorrectly conclude the population means are different when they are actually the same.

***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. If the data were simply random noise, approximately 5% of the *p*-values would be less than 0.05, 10% of the *p*-values would be less than 0.10, etc. Thus, even if the data were only random noise, we would get approximately 10 “significant” results out of 200 compounds when the false positive rate is 0.05.

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).

To interpret the *q*-value, first sort the data by the *p*-value then choose the cutoff for significance (typically $p \leq 0.05$). 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	<i>p</i> -value	<i>q</i> -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

Random Forest: *t*-tests are used to determine if the population means are different, but do not tell us about individual observations. Random forests (see Leo Breiman’s “Random Forest” from *Machine Learning*, 2001) are used to *classify* individuals. Random forests are based on a consensus of a large number of decision trees. Classification trees split the data into groups based on the compounds that provide the best separation. An example is shown in Figure 1. From Figure 1, we see that if the value of Compound 145 is less than 1.456, then it classified as a control, and otherwise it is classified an individual with the disease. (Decision trees can have more branches that split from previous branches).

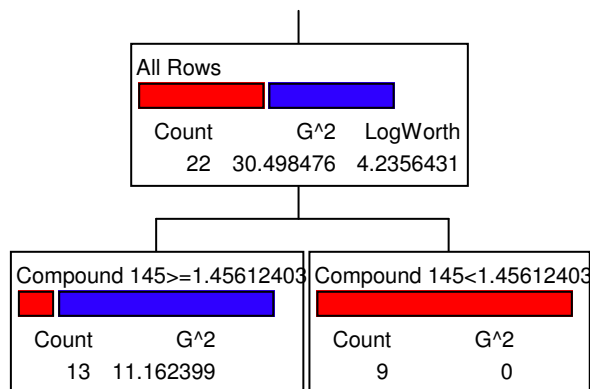


Figure 1. Sample Decision Tree (Red = Control, Blue=Disease)

Classification trees often overfit the data. Overfitting means that the model does well for the existing data, but will poorly predict new data. Even with random noise, one can find a classification tree that will give good separation to the data. Random forests are based on a collection of trees and do not overfit the data. The error rate from the random forest will give you an estimate of how well you can predict new data.

With a random forest, one continually re-samples from the data set. For example, suppose Group A (e.g., Control) and Group B (e.g., Disease) have 10 members each. Then one can sample five from each of these to create a decision tree and then predict the remaining samples from this tree. Thus, predictions are not based on the same samples that were used to create the tree to begin with. The random forest repeats this for a large number of times to create a large numbers of trees. (The random forest also samples the compounds). An example of the process is shown in Table 2.

Table 2. Example of Random Forest Construction

Tree	samples used to construct the tree (in-bag)			
1	A2, A4, A6, A8, A9	B1, B2, B3, B7, B8	create tree	classify the out-of-bag samples (A1, A3, A5, A7, A10, B4, B5, B6, B9, B10)
2	A1, A4, A6, A7, A10	B1, B2, B3, B5, B8	create tree	classify the out-of-bag samples (A2, A3, A5, A8, A9, B4, B6, B7, B9, B10)
3	A1, A6, A7, A8, A9	B1, B2, B5, B7, B9	create tree	classify the out-of-bag samples (A2, A3, A4, A5, A10, B3, B4, B6, B8, B10)
...				
K	A3, A6, A7, A9, A10	B2, B3, B5, B7, B9	create tree	classify the out-of-bag samples (A1, A2, A4, A5, A8, B1, B4, B6, B8, B10)

To obtain the final classification for each observation, compute the proportion of times the observation was classified into each group for each time it was an “out-of-bag” sample. For example, A2 was classified in trees 2, 3, and k shown above (and ones in between not shown). If 70% of the trees classified A2 as an “A,” and 30% classified A2 as “B,” then the final classification of A2 is “A.” Expressed as decimals, 0.7 and 0.3 are referred to as the “votes.” Once the final classifications are determined for each sample, we construct a confusion matrix to show how well the random forest classified the data. The row labels are the actual group labels, and the columns are the predicted labels. An example is shown in Table 3.

Table 3. Example of a confusion matrix

	A	B	error
A	7	3	0.3
B	2	8	0.2
	OOB Error = 25%		

From Table 3 we see that of the 10 observations that belong to Group A that seven are correctly classified, while three are classified as belonging to Group B. Of the 10 observations belonging to Group B, eight are correctly classified, while the other two are incorrectly classified as belonging to Group A. Thus, overall 15 of the 20 observations are correctly classified, which gives an error of $5/20 = 25\%$. Since each observation was classified based on trees from *different* observations, this error rate is a good estimate of well we can predict *new* observations.

Finally, we can see which variables contributed the most to the final classifications. Variable importance is determined by permuting the values of the variables and comparing the original

“out-of-bag” error of the samples to the “out-of-bag” error of the samples using the permuted values. The more important the variable, the larger the decrease in prediction accuracy there will be. The *randomForest* package in *R* (Andy Liaw is the maintainer) will show a plot of the most important compounds.

Z-score: 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.

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

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 (Figure 2).

Table 4 lists the data from a hypothetical study in which 15 different metabolites were measured for two different treatment groups (CONTROL and DRUG group). Each group consisted of six subjects. Table 5 lists the calculated Z-scores for each metabolite and for each sample from Table 4. Figure 3 shows the Z-score graph for this study. Each dot in the graph represents the Z-score of a single subject for a single metabolite. The zero point on the X-axis is the mean of the six CONTROL group samples. The Y-axis represents the different metabolites, with all points for a single metabolite on the same horizontal line.

It is clear from Figure 3 and Table 4 that all CONTROL samples have relatively small Z-scores while the variation in the Z-scores is much larger for the DRUG group samples.

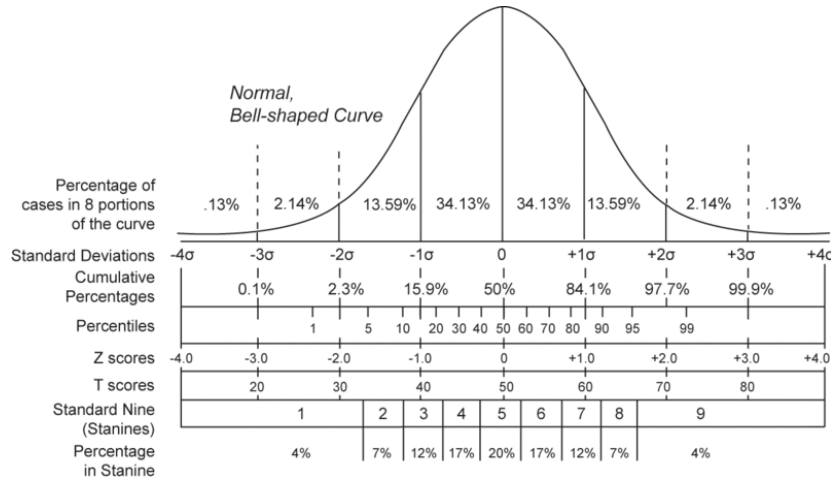


Figure 2. Comparison of the various grading methods in a normal distribution.

Table 4. Ion intensities for the fifteen metabolites measured in each of the twelve study samples and median scaled to 1.

SAMPLE	GROUP	METABOLITE														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
# 01	Contr	0.85	0.98	2.23	1.21	0.41	1.05	1.6	1.49	0.44	0.42	1.1	0.88	1.02	1.13	1.55
# 02	Contr	1.45	0.94	1.79	1.02	0.58	1.38	0.85	1.72	0.22	0.28	1.27	0.71	1.26	1.06	0.82
# 03	Contr	0.94	0.84	0.62	0.61	0.31	0.7	2.24	1.44	0.75	0.75	0.87	1.06	1.07	0.59	0.96
# 04	Contr	1.63	1.04	1.35	0.98	1.04	1.25	0	1.71	0.83	0.87	1	0.87	1.43	0.94	0.58
# 05	Contr	0.9	1	0.91	0.78	1.33	1.14	0.68	1.87	0.66	0.67	0.76	0.71	0.96	1.28	3.33
# 06	Contr	1.31	1.11	1.21	1.13	0.96	1.2	1.01	2.13	0.82	0.5	0.79	0.99	1.51	1.18	1.37
# 07	Drug	1.98	1.01	0.85	1.31	3.51	1.05	0.98	0.56	4.97	3.09	1.49	1.63	2.49	1.13	1.04
# 08	Drug	0.38	3.8	0.09	3.32	0.17	0.24	3.21	0.55	1.17	1.51	0.92	1.01	0.46	0.68	1.04
# 09	Drug	1.15	1	1.09	1.96	2.24	0.95	0.8	0.5	3.51	1.13	1	0.88	0.76	0.72	1.16
# 10	Drug	1.06	1.57	0.46	0.78	1.88	0.78	0.99	0.37	4.3	4.02	1.26	1.55	0.94	0.83	0.93
# 11	Drug	0.31	0.94	0.75	0.66	1.71	0.78	1.9	0.46	3.21	2.49	1.01	1.57	0.98	0.52	0.68
# 12	Drug	0.67	1	1.3	0.62	0.92	0.91	1.39	0.37	2.62	1.68	0.84	1.49	0.97	1.6	0.32

Table 5. Z-scores for each metabolite and study sample calculated from the intensity data in Table 4.

SAMPLE	GROUP	METABOLITE														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
# 01	Contr	-1.01	-0.05	1.5	1.14	-0.9	-0.3	0.69	-0.93	-0.74	-0.73	0.69	0.07	-0.83	0.41	0.12
# 02	Contr	0.82	-0.49	0.75	0.29	-0.48	1.11	-0.28	-0.03	-1.65	-1.37	1.55	-1.12	0.23	0.12	-0.62
# 03	Contr	-0.73	-1.58	-1.25	-1.54	-1.15	-1.8	1.52	-1.13	0.54	0.76	-0.48	1.33	-0.61	-1.8	-0.48
# 04	Contr	1.37	0.6	0	0.11	0.67	0.56	-1.37	-0.07	0.86	1.31	0.18	0	0.97	-0.37	-0.86
# 05	Contr	-0.85	0.16	-0.75	-0.78	1.4	0.09	-0.5	0.57	0.16	0.4	-1.04	-1.12	-1.09	1.02	1.91
# 06	Contr	0.4	1.36	-0.24	0.78	0.47	0.34	-0.07	1.59	0.82	-0.37	-0.89	0.84	1.33	0.61	-0.07
# 07	Drug	2.44	0.27	-0.86	1.59	6.85	-0.3	-0.11	-4.6	17.91	11.38	2.66	5.33	5.63	0.41	-0.4
# 08	Drug	-2.44	30.73	-2.15	10.57	-1.5	-3.77	2.77	-4.64	2.26	4.21	-0.23	0.98	-3.29	-1.43	-0.4
# 09	Drug	-0.09	0.16	-0.45	4.49	3.67	-0.73	-0.34	-4.84	11.9	2.49	0.18	0.07	-1.97	-1.27	-0.28
# 10	Drug	-0.37	6.39	-1.52	-0.78	2.77	-1.46	-0.09	-5.35	15.15	15.59	1.5	4.77	-1.18	-0.82	-0.51
# 11	Drug	-2.65	-0.49	-1.03	-1.32	2.35	-1.46	1.08	-4.99	10.66	8.65	0.23	4.91	-1	-2.09	-0.76
# 12	Drug	-1.56	0.16	-0.09	-1.5	0.37	-0.9	0.42	-5.35	8.23	4.98	-0.63	4.35	-1.05	2.34	-1.12