|  |
| --- |
|  |
| mViewTM **REPORT** |
| ***Saccharomyces cerevisiae* grown under different dilution rates and media conditions**  **CHAL-01-16VW** |
| CLIENT: Chalmers University of Technology  Benjamin Sanchez, PhD  AUTHOR: Brian Ingram, PhD  APPROVAL: Robert Mohney, PhD  DATE: September 21, 2016 |
|  |





Metabolon, Inc. • 617 Davis Drive, Suite 400, Durham, NC 27713 • (919) 572-1711 www.metabolon.com • Contact: [info@metabolon.com](mailto:info@metabolon.com)

Table of Contents

[**Objective** 3](#_Toc448326797)

[**Experimental Procedures** 3](#_Toc448326798)

[**Results and Biological Interpretation** 4](#_Toc448326799)

[Metabolite Summary and Significantly Altered Biochemicals 4](#_Toc448326800)

[Biological Interpretation 8](#_Toc448326801)

[**Conclusions** 12](#_Toc448326802)

[**Study Parameters** 13](#_Toc448326803)

[Data Quality: Instrument and Process Variability 13](#_Toc448326804)

[**Appendix** 14](#_Toc448326805)

[Metabolon Platform 14](#_Toc448326806)

[Statistical Methods and Terminology 19](#_Toc448326808)

# **Objective**

***Purpose of Experiment***

The goal of this study was to compare the metabolic profiles of *Saccharomyces cerevisiae* cells grown chemostatically under carbon-, nitrogen-, and oxygen-limiting conditions.

# **Experimental Procedures**

***Experimental design***

Global metabolic profiles were determined in Saccharomyces cerevisiae cells grown at a steady state under different carbon-, nitrogen-, and oxygen-limiting conditions. The dilution rate was varied for the carbon-limiting conditions and was kept constant at 0.15 (1/h) for the nitrogen-, oxygen-, and oxygen + carbon limited growths. In addition, the organism was also cultivated at a maximum growth rate using batch conditions (batch glucose phase).

|  |  |  |
| --- | --- | --- |
| **Group** | **Description** | **n** |
| C 0.025 | Aerobic steady state growth rate with Carbon Limitation = 0.025 (1/h) | 3 |
| C 0.05 | Aerobic steady state growth rate with Carbon Limitation = 0.05 (1/h) | 3 |
| C 0.1 | Aerobic steady state growth rate with Carbon Limitation = 0.1 (1/h) | 3 |
| C 0.15 | Aerobic steady state growth rate with Carbon Limitation = 0.15 (1/h) | 3 |
| C 0.2 | Aerobic steady state growth rate with Carbon Limitation = 0.2 (1/h) | 3 |
| C 0.25 | Aerobic steady state growth rate with Carbon Limitation = 0.25 (1/h) | 3 |
| C 0.3 | Aerobic steady state growth rate with Carbon Limitation = 0.3 (1/h) | 3 |
| C 0.35 | Aerobic steady state growth rate with Carbon Limitation = 0.35 (1/h) | 3 |
| C 0.4 | Aerobic steady state growth rate with Carbon Limitation = 0.4 (1/h) | 3 |
| N 0.15 | Aerobic steady state growth rate with Nitrogen Limitation = 0.15 (1/h) | 3 |
| O 0.15 | Anaerobic Steady state growth rate = 0.15 (1/h) | 3 |
| Anaerobic C 0.15 | Anaerobic growth rate with Carbon Limitation =0.15 (1/h) | 3 |
| BGP 0.4 | Batch glucose phase in mid-exponential phase | 3 |

# **Results and Biological Interpretation**

## 

## Metabolite Summary and Significantly Altered Biochemicals

The present dataset comprises a total of 418 compounds of known identity (named biochemicals). Following normalization to Bradford protein or DNA concentrations, log transformation and imputation of missing values, if any, with the minimum observed value for each compound, ANOVA contrasts and Welch’s two-sample *t*-test were used to identify biochemicals that differed significantly between experimental groups. A summary of the numbers of biochemicals that achieved statistical significance (*p*≤0.05), as well as those approaching significance (0.05<*p*<0.10), is shown below.

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

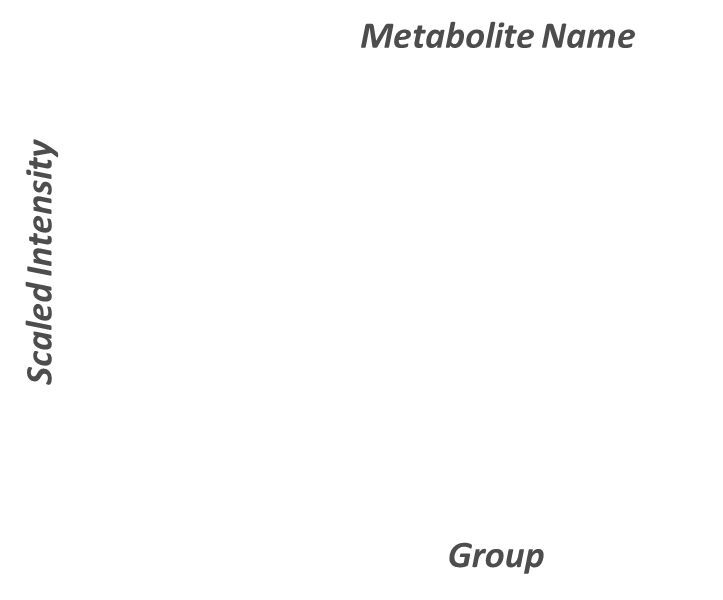
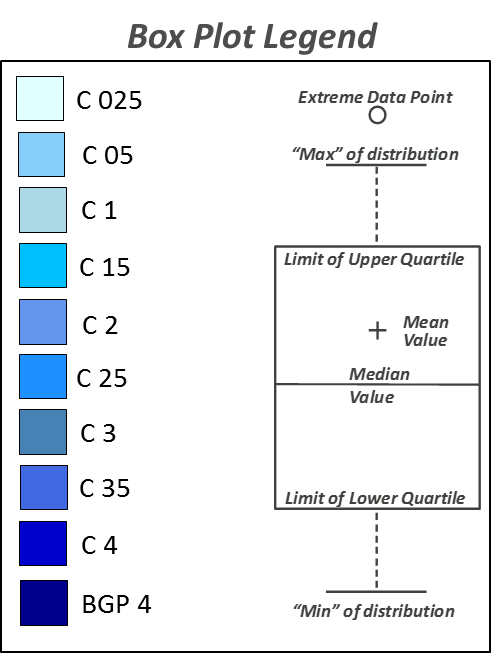
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Statistical Comparisons: Protein Normalized** | | | | | | | | | |
| **ANOVA Contrasts** | **Group Effect** | **C 0.025** | **C 0.05** | **C 0.15** | **C 0.2** | **C 0.25** | **C 0.3** | **C 0.35** | **C 0.4** |
| **C 0.1** | **C 0.1** | **C 0.1** | **C 0.1** | **C 0.1** | **C 0.1** | **C 0.1** | **C 0.1** |
| **Total biochemicals *p*≤0.05** | **317** | **153** | **113** | **12** | **70** | **137** | **171** | **194** | **220** |
| **Biochemicals  (↑↓)** | **- - - -** | **58 | 95** | **76 | 37** | **11 | 1** | **50 | 20** | **62 | 75** | **42 | 129** | **115 | 79** | **141 | 79** |
| **Total biochemicals 0.05<*p*<0.10** | **29** | **33** | **38** | **13** | **30** | **27** | **38** | **33** | **30** |
| **Biochemicals  (↑↓)** | **- - - -** | **14 | 19** | **32 | 6** | **11 | 2** | **23 | 7** | **15 | 12** | **15 | 23** | **17 | 16** | **21 | 9** |

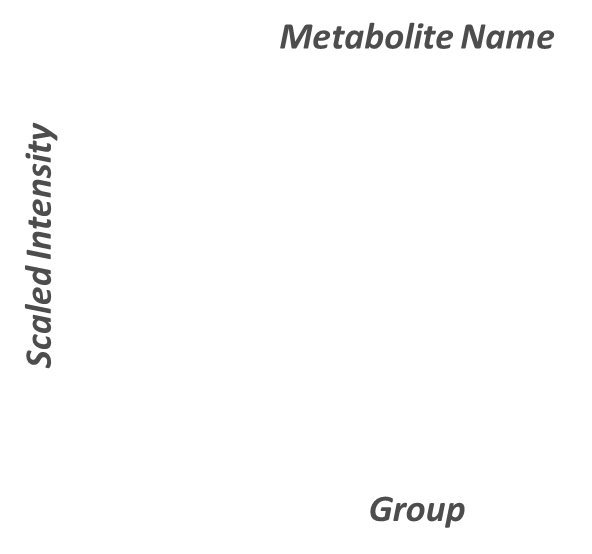
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Statistical Comparisons: DNA Normalized** | | | | | | | | | | |
| **ANOVA Contrasts** | **Group Effect** | **C 0.025** | **C 0.05** | **C 0.15** | **C 0.2** | **C 0.25** | **C 0.3** | **C 0.35** | **C 0.4** |
| **C 0.1** | **C 0.1** | **C 0.1** | **C 0.1** | **C 0.1** | **C 0.1** | **C 0.1** | **C 0.1** |
| **Total biochemicals *p*≤0.05** | **339** | **172** | **90** | **56** | **113** | **185** | **247** | **224** | **232** |
| **Biochemicals  (↑↓)** | **- - - -** | **41 | 131** | **39 | 51** | **1 | 55** | **2 | 111** | **26 | 159** | **7 | 240** | **63 | 161** | **60 | 172** |
| **Total biochemicals 0.05<*p*<0.10** | **12** | **36** | **37** | **39** | **39** | **32** | **30** | **21** | **23** |
| **Biochemicals  (↑↓)** | **- - - -** | **13 | 23** | **22 | 15** | **1 | 38** | **7 | 32** | **3 | 29** | **6 | 24** | **6 | 15** | **10 | 13** |

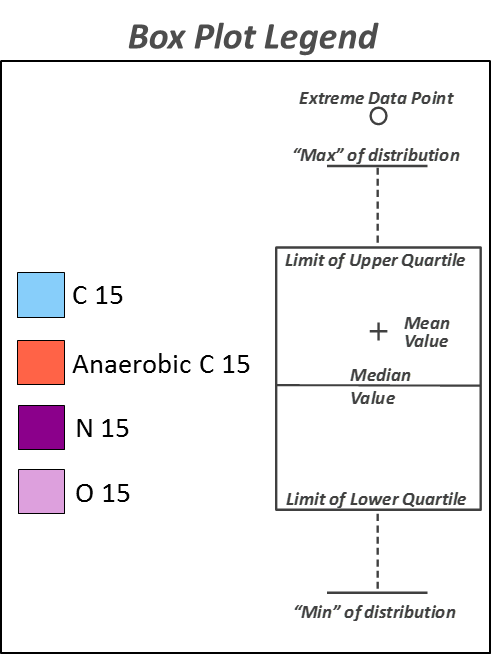
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Statistical Comparisons: Protein Normalized** | | | | | |
| **Welch's Two-Sample *t*-Test** | **N 0.15** | **O 0.15** | **Anaerobic C 0.15** | **BGP 0.4** | **BGP 0.4** |
| **C 0.15** | **C 0.15** | **C 0.15** | **C 0.1** | **C 0.4** |
| **Total biochemicals *p*≤0.05** | **176** | **173** | **236** | **226** | **90** |
| **Biochemicals  (↑↓)** | **62 | 114** | **27 | 146** | **42 | 194** | **163 | 63** | **74 | 16** |
| **Total biochemicals 0.05<*p*<0.10** | **44** | **52** | **43** | **41** | **57** |
| **Biochemicals  (↑↓)** | **12 | 32** | **12 | 40** | **5 | 38** | **27 | 14** | **41 | 16** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Statistical Comparisons: DNA Normalized** | | | | | |
| **Welch's Two-Sample *t*-Test** | **N 0.15** | **O 0.15** | **Anaerobic C 0.15** | **BGP 0.4** | **BGP 0.4** |
| **C 0.15** | **C 0.15** | **C 0.15** | **C 0.1** | **C 0.4** |
| **Total biochemicals *p*≤0.05** | **189** | **177** | **211** | **160** | **158** |
| **Biochemicals  (↑↓)** | **77 | 112** | **33 | 144** | **67 | 144** | **52 | 108** | **132 | 26** |
| **Total biochemicals 0.05<*p*<0.10** | **39** | **48** | **42** | **56** | **40** |
| **Biochemicals  (↑↓)** | **18 | 21** | **14 | 34** | **7 | 35** | **33 | 23** | **29 | 11** |

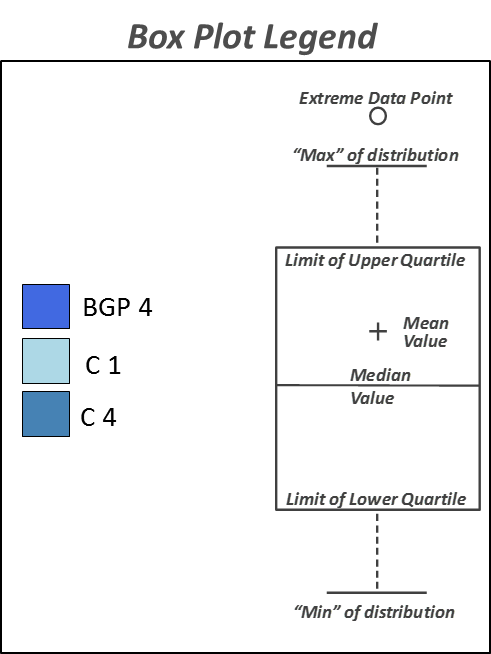
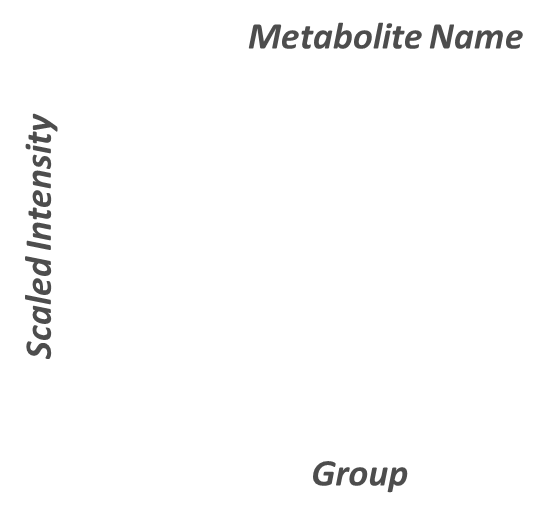
We have also included in the electronic deliverables, a file with data for each biochemical displayed as box plots like that shown in the example figure below.

76372a014ca3a4de.emf





5fbf719d4a76b44c.emf

62bd6aface483b22.emf

## Biological Interpretation

The goal of this study was to compare the metabolic profiles of *Saccharomyces cerevisiae* cells grown chemostatically under carbon-, nitrogen-, and oxygen-limiting conditions. A total of 13 different growth conditions were utilized: 9 aerobic carbon-limiting conditions differing with regards to their dilution rates (0.025, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, and 0.4), 1 aerobic nitrogen-limiting condition (0.15 dilution rate), 1 oxygen-limiting condition (0.15 dilution rate), 1 carbon- and oxygen-limiting condition (0.15 dilution rate), and a batch condition supporting maximum growth (0.4 dilution rate). A total of 3 samples were profiled per experimental condition. The samples were loaded in an equivalent manner across the analytical platforms and the values for each biochemical were normalized based on either their measured protein (Bradford) or DNA concentrations. The provided client data table includes the raw data for each sample and the statistical analysis results for both the protein- and DNA-normalized datasets.

Datasets provided in the mView product can be quite large and contain a great deal of information. To provide an initial focus for further consideration, a few observations are offered below from a cursory view of the data. These are not presented as a comprehensive analysis; the PI, with a much greater knowledge of the experimental system, is encouraged to make a detailed study of the data for additional or alternative interpretations. For the discussion below, please refer to graphical illustrations in the accompanying PowerPoint file, or to the statistical tables, box plots, and other graphics found in the accompanying client data table. For convenience, biochemicals are highlighted in **bold** **text** in the report when they correspond to the data plots or heatmap figures shown in the accompanying Graphics file.

* **Principal component analysis and hierarchical clustering analysis**. Principal component analysis (PCA) and hierarchical clustering are statistical methods that can be used to obtain a high level view of metabolomic datasets. PCA, briefly, is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principle components. Hierarchical clustering (HC), meanwhile, assesses sample similarity by group to build a “hierarchy” of clustered samples. Briefly, populations that differ are expected to cluster separately in these analyses.

In this study, PCA and HC were applied to both the protein and DNA-normalized datasets. As shown on slides 8-15, both methods revealed significant separations based on the particular growth limitations and dilution rates that were employed. The *S. cerevisiae* metabolomeappeared to be quite sensitive therefore to changes in its growth environment.

*Note 1:* As indicated in the PCA plots and HC figures, sample segregation was similar overall in the protein and DNA-normalized datasets. The majority of group-based differences appeared to be conserved therefore in the two datasets. The interpretation that follows focus however on the differences identified in the protein-normalized dataset.

*Note 2:* Moderate levels of heterogeneity were observed in *some* of the experimental groups. This can have a confounding effect on the statistical analysis, particularly in underpowered studies (n=3). For this reason, Metabolon recommends a minimum of 4-6 samples for cell-based studies.

* **Overview of the metabolic differences observed under the different aerobic carbon-limited conditions.** In this study, *S. cerevisiae* was grown under aerobic carbon-limiting conditions using 9 different dilution rates. To assess the effects occurring under each these conditions, fold-change values (along with their statistical significance) were calculated for each of the dilution rates relative to the 0.1 dilution rate. A brief overview of the most significant findings in these comparisons follows.

*Carbohydrate metabolism*. When compared to the carbon-limited culture grown at the 0.1 dilution rate, those grown at higher dilution rates (0.025 and 0.05) exhibited significant increases in their intracellular **glucose** levels. The cultures grown at lower dilution rates (0.2, 0.25, 0.3, 0.35, and 0.4), meanwhile, exhibited significant decreases in their measured glucose levels. These differences, briefly, are consistent, with changes in the rates at which the cultures utilized the glucose that was available to them. In this case, there was some evidence of increased glucose utilization through the glycolytic and pentose phosphate pathways (PPP) in the cultures grown at the higher growth rates (*e.g.,* increases were observed in **fructose 1,6 bisphosphate** (detected here as a hexose diphosphate isobar), **lactate**, and in the PPP intermediate **6-phosphogluconate**). In addition to these changes, the cultures grown at the higher growth rates also exhibited significant decreases in **trehalose** and **trehalose 6-phosphate**. Trehalose, notably, is often regarded as a major reserve carbohydrate for *S. cerevisiae*. Its utilization was likely altered therefore as the growth rates of the organism increased.

*Lipid metabolism*. In this study, the cultures grown at the higher growth rates (*e.g.,* those grown at the 0.25, 0.3, 0.35, and 0.4 dilution rates) tended to exhibited lower levels of several **long-chain fatty acids**. Fatty acids with **3-hydroxy** **modifications**, in contrast, tended to increase within these groups. Given that the latter can be derived during β-oxidation, these changes could relate to changes in β-oxidation rates/efficiency. In addition to these differences, the groups growing at the faster growth rates also exhibited significant increases in metabolites formed via the mevalonate pathway (**3-hydroxy-3-methylglutaryl CoA** and **mevalonate**). This latter set of changes, notably, could be of importance with regards to sterol synthesis (perhaps to support membrane biogenesis at the higher growth rates).

*Amino acid homeostasis.* The aerobic carbon-limited cultures in this study exhibited a number of changes in their amino acid profiles as their growth rates increased. At the highest growth rates (dilution rates of 0.35 and 0.4), for instance, the cells exhibited significant increases in a number of **free amino acids** (*e.g.,* serine, threonine, allo-threonine, aspartate, and lysine). The groups exhibited significant increases, likewise, in a number of **N-acetylated acetylated amino acids** (*e.g.,* acetylasparagine, N-acetylglutamate, N-acetylglutamine, N-acetylphenylalanine, etc.) and **dipeptides** species. These changes, notably, are consistent with increases in protein turnover (and perhaps protein abundance). The N-acetylated amino acids detected here, for instance, are typically derived from proteins that have underwent posttranslational acetylation reactions. Dipeptides, likewise, are short polymers of amino acids that can be liberated during protein turnover.

*TCA cycle activity*. When the cultures grown at the 0.35 and 0.4 dilution rates were compared to the culture grown at the 0.1 dilution rate, they exhibited significant and/or trending (0.05<*p*<0.1) increases in several TCA cycle intermediates including **alpha-ketoglutarate**, **succinate**, **fumarate**, and **malate**. These changes are consistent thus with differences in the rates of carbon entry into the cycle (arising perhaps from the changes noted above in carbohydrate, lipid, and amino acid abundance/utilization). *Note:* The pattern of these metabolites varied widely at the lower growth rates (*i.e.,* those grown at dilution rates 0.025-0.3). The rate of carbon flow into the cycle varied significantly thus with the different dilution rates.

*Nucleotide homeostasis*. When compared to the culture grown at the 0.1 dilution rate, the cells grown at the lower dilution rates (0.25 or lower) exhibited significant alterations in their nucleotide profiles. Metabolites linked to the de novo synthesis of purine and pyrimidine nucleotides (*e.g.,* **AICA ribonucleotide** and **dihydroorotate)**, for example, were significantly higher in the cultures grown at the higher growth rates. These groups also contained lower levels of several **nucleotides**, **nucleosides** and **nucleobases**. These changes may be attributable to changes in the synthesis and degradation of RNA and DNA and/or to alterations in energy metabolism (as certain nucleotides also play a role in energetics).

*Glutathione homeostasis*. When compared to the carbon-limiting culture maintained at the 0.1 dilution rate, those grown at the 0.3 and 0.4 dilution rates exhibited significant increases in oxidized glutathione (GSSG) and cysteine-glutathione disulfide. They exhibited significant decreases, meanwhile, in a number of gamma-glutamyl amino acid species. This latter class of metabolites, notably, are typically generated when glutathione undergoes breakdown and/or recycling reactions. The above changes are consistent thus with changes in glutathione synthesis and utilization at the higher growth rates.

* **Overview of the metabolic differences observed between the batch glucose phase culture and the cultures grown under carbon-limiting conditions.** To gain further insight into the effects of carbon limitation, we compared the carbon-limited cultures grown at the 0.1 and 0.4 dilution rates to the culture grown under the batch glucose phase condition. An overview of the major differences in these comparisons follows.

*BGP growth vs the carbon-limited 0.1 dilution rate growth*. When the culture grown under batch conditions was compared to the carbon-limiting culture grown at the 0.1 dilution rate, 226 statistically significant differences were identified (please see the statistical summary table for the protein-normalized dataset on slides 3-4). Notably, the differences in this comparison were very similar to that which was observed when the high growth rate carbon-limited cultures were compared to the low growth rate carbon-limited cultures (e.g., the 0.4 vs. 0.1 carbon limited comparison); please see the Venn diagram on slide 21. In addition to these common changes, the batch condition also led to robust differences in **polyamine**, **phospholipid**, and **lysolipid** abundance. As expected, **glucose** levels and markers of its utilization were also significantly higher under the BGP condition.

*BGP growth vs the carbon-limited 0.4 dilution rate growth*. As noted above, the metabolic signatures of the BGP carbon-limited growth maintained at the 0.4 dilution rate were highly similar. The most pronounced differences between the two appeared to center on differences relating to glucose utilization and lipid metabolism. **Glucose** levels, for instance, were ~13 fold higher under the batch condition. The glycolytic products **fructose 1,6-bisophosphate**, **dihydroxyacetone** **phosphate**, and **lactate**, likewise, were significantly elevated as well. With regards to lipid metabolism, the batch culture exhibited significant increases in nearly all of the **phospholipids** that were detected and in several **lysolipids**. In addition, the batch culture also exhibited significantly higher levels of both **glycerol** and **glycerol 3-phosphate**. These carbon-based intermediates, notably, are the building blocks for both phospholipid and glycerolipid synthesis. The increase in carbon availability appeared thus to have a significant impact therefore on complex lipid synthesis and abundance.

* **Overview of the metabolic differences observed between the nitrogen-limited and carbon-limited cultures grown at the 0.15 dilution rates.**

*Nitrogen homeostasis*. Given the differences in nitrogen abundance amongst the carbon- and nitrogen-limiting conditions, it was expected that the nitrogen-limited culture may exhibit decreases in metabolites associated with nitrogen balance. There, indeed, was some evidence for this as several amino acids were present at lower levels in the nitrogen-limited culture (*e.g.,* histidine, phenylalanine, tyrosine, and arginine). It exhibited decreases, likewise in several polyamines (*e.g.,* putrescine and spermidine) and in metabolites linked to their synthesis (*e.g.,* ornithine). Metabolites typically derived from protein sources (e.g., dipeptides), likewise, were generally lower as well under the nitrogen-limiting condition.

*Glucose utilization and TCA cycle activity*. As expected, the carbon- and nitrogen-limited cultures exhibited differences with regards to glucose availability and utilization. As shown on slide 23, **glucose** levels were significantly higher in the nitrogen-limited culture as were several markers on its utilization (*e.g.,* **glucose**, **fructose 1,6** **bisphosphate**, **2-phosphoglycerate**, **3-phospholgycerte**, and **lactate**). These observations, briefly, are consistent with there being more carbon available to fuel the glycolytic and fermentative pathways in the nitrogen-limited culture. In addition to these differences, the nitrogen-limited culture also exhibited significant and/or trending increases in several TCA cycle intermediates (*e.g.,* **alpha-ketoglutarate**, **succinate**, **fumarate**, and **malate**).

*Nucleotide homeostasis.* When compared to the carbon-limited culture, the nitrogen-limited culture exhibited significant increases in several pyrimidine biosynthetic intermediates (*e.g.,* dihydroorotate, orotate, etc.). It also exhibited decreases in several pyrimidines and purine nucleotides/nucleosides. These changes, again, are consistent with possible changes in the synthesis and turnover of DNA/RNA.

* **Overview of the metabolic differences observed between the anaerobic, anaerobic and carbon-limited, and carbon-limited cultures grown at the 0.15 dilution rates.**

*Glycolytic and TCA cycle activities*. When compared to the carbon-limited culture at the 0.15 dilution rate, the anaerobic cultures exhibited significant differences with regards to glucose utilization. **Glucose**, for instance, was lower under both of these conditions. The anaerobic culture also exhibited significant increases in **fructose 1,6 diphosphate**, **2-phosphoglycerate**, and **3-phosphoglycerate**. The carbon- and oxygen-limited culture, meanwhile, displayed increases in **fructose 1,6 diphosphate** and **lactate**. The changes are consistent thus with increases in anaerobic glycolytic activity. In parallel with these changes, the above culture also exhibited significant and/or trending decreases in several TCA cycle intermediates including **alpha-ketoglutarate**, **fumarate**, and **malate**. These latter changes likely correlate with decreases in TCA cycle flow and perhaps oxidative phosphorylation rates.

*Nitrogen, lipid, and nucleotide homeostasis*. In this study, both of the anaerobically grown cultures exhibited significant decreases in a number of metabolites linked to nitrogen homeostasis (*e.g.,* freeamino acids, modified amino acids, dipeptides, polyamines etc.). In addition, they also displayed alterations in a number of complex lipid species and nucleotide metabolites.

# **Conclusions**

This global metabolomic profiling study was conducted to compare the metabolic profiles of *Saccharomyces cerevisiae* cells grown under carbon-, nitrogen-, and oxygen-limiting conditions. PCA, hierarchical clustering, and statistical analysis of the profiled samples revealed differences that were directly attributable to the dilution rates and growth limitations that were utilized. The different dilution rates employed under the carbon-limited condition, for instance, had significant effects on metabolic processes relating to glucose, lipid, amino acid, and nucleotide homeostasis. Notably, the metabolic profile obtained at the highest growth rate (dilution factor 0.4) was similar to that which was observed under the batch glucose phase (BGP) condition. The most pronounced differences between the two, in fact, centered on glucose availability (as expected) and utilization (likely to support lipid biogenesis). The nitrogen- and oxygen-limited cultures, meanwhile, differed from the carbon-limited cultures in several manners (*e.g.,* with regards to glucose and nitrogen homeostasis, energy metabolism, etc.).

# **Study Parameters**

## 

## Data Quality: Instrument and Process Variability

|  |  |  |
| --- | --- | --- |
| ***QC Sample*** | ***Measurement*** | ***Median RSD*** |
| Internal Standards | Instrument Variability | 3% |
| Endogenous Biochemicals | Total Process Variability | 7% |

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

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

**QA/QC:** Several types of controls were analyzed in concert with the experimental samples: a pooled matrix sample generated by taking a small volume of each experimental sample (or alternatively, use of a pool of well-characterized human plasma) served as a technical replicate throughout the data set; extracted water samples served as process blanks; and a cocktail of QC standards that were carefully chosen not to interfere with the measurement of endogenous compounds were spiked into every analyzed sample, allowed instrument performance monitoring and aided chromatographic alignment. Tables 1 and 2 describe these QC samples and standards. Instrument variability was determined by calculating the median relative standard deviation (RSD) for the standards that were added to each sample prior to injection into the mass spectrometers. Overall process variability was determined by calculating the median RSD for all endogenous metabolites (i.e., non-instrument standards) present in 100% of the pooled matrix samples. Experimental samples were randomized across the platform run with QC samples spaced evenly among the injections, as outlined in Figure 1.

**Table 1:** **Description of Metabolon QC Samples**

|  |  |  |
| --- | --- | --- |
| **Type** | **Description** | **Purpose** |
| MTRX | Large pool of human plasma maintained by Metabolon that has been characterized extensively. | Assure that all aspects of the Metabolon process are operating within specifications. |
| CMTRX | Pool created by taking a small aliquot from every customer sample. | Assess the effect of a non-plasma matrix on the Metabolon process and distinguish biological variability from process variability. |
| PRCS | Aliquot of ultra-pure water | Process Blank used to assess the contribution to compound signals from the process. |
| SOLV | Aliquot of solvents used in extraction. | Solvent Blank used to segregate contamination sources in the extraction. |

**Table 2:** **Metabolon QC Standards**

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

****

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

**Ultrahigh Performance Liquid Chromatography-Tandem Mass Spectroscopy (UPLC-MS/MS):** All methods utilized a Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer operated at 35,000 mass resolution. The sample extract was dried then reconstituted in solvents compatible to each of the four methods. Each reconstitution solvent contained a series of standards at fixed concentrations to ensure injection and chromatographic consistency. One aliquot was analyzed using acidic positive ion conditions, chromatographically optimized for more hydrophilic compounds. In this method, the extract was gradient eluted from a C18 column (Waters UPLC BEH C18-2.1x100 mm, 1.7 µm) using water and methanol, containing 0.05% perfluoropentanoic acid (PFPA) and 0.1% formic acid (FA). Another aliquot was also analyzed using acidic positive ion conditions, however it was chromatographically optimized for more hydrophobic compounds. In this method, the extract was gradient eluted from the same afore mentioned C18 column using methanol, acetonitrile, water, 0.05% PFPA and 0.01% FA and was operated at an overall higher organic content. Another aliquot was analyzed using basic negative ion optimized conditions using a separate dedicated C18 column. The basic extracts were gradient eluted from the column using methanol and water, however with 6.5mM Ammonium Bicarbonate at pH 8. The fourth aliquot was analyzed via negative ionization following elution from a HILIC column (Waters UPLC BEH Amide 2.1x150 mm, 1.7 µm) using a gradient consisting of water and acetonitrile with 10mM Ammonium Formate, pH 10.8. The MS analysis alternated between MS and data-dependent MSn scans using dynamic exclusion. The scan range varied slighted between methods but covered 70-1000 m/z. Raw data files are archived and extracted as described below.

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

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

**Data Extraction and Compound Identification:** Raw data was extracted, peak-identified and QC processed using Metabolon’s hardware and software. These systems are built on a web-service platform utilizing Microsoft’s .NET technologies, which run on high-performance application servers and fiber-channel storage arrays in clusters to provide active failover and load-balancing. Compounds were identified by comparison to library entries of purified standards or recurrent unknown entities. Metabolon maintains a library based on authenticated standards that contains the retention time/index (RI), mass to charge ratio (*m/z)*, and chromatographic data (including MS/MS spectral data) on all molecules present in the library. Furthermore, biochemical identifications are based on three criteria: retention index within a narrow RI window of the proposed identification, accurate mass match to the library +/- 10 ppm, and the MS/MS forward and reverse scores between the experimental data and authentic standards. The MS/MS scores are based on a comparison of the ions present in the experimental spectrum to the ions present in the library spectrum. While there may be similarities between these molecules based on one of these factors, the use of all three data points can be utilized to distinguish and differentiate biochemicals. More than 3300 commercially available purified standard compounds have been acquired and registered into LIMS for analysis on all platforms for determination of their analytical characteristics. Additional mass spectral entries have been created for structurally unnamed biochemicals, which have been identified by virtue of their recurrent nature (both chromatographic and mass spectral). These compounds have the potential to be identified by future acquisition of a matching purified standard or by classical structural analysis.

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

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

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

## Statistical Methods and Terminology

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

1. **Welch’s two-sample *t*-test**

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

This version of the two-sample *t*-test allows for unequal variances (variance is the square of the standard deviation) and has an *approximate* *t*-distribution with degrees of freedom estimated using Satterthwaite’s approximation. The test statistic is given by *t* , and the degrees of freedom is given by , where , are the sample means, s1, s2, are the sample standard deviations, and n1, n2 are the samples sizes from groups 1 and 2, respectively. We typically use a two-sided test (tests whether the means are different) as opposed to a one-sided test (tests whether one mean is greater than the other).

1. **Matched pairs *t*-test**

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

The matched pairs *t*-test is equivalent to the one-sample *t*-test performed on the differences of the observations taken on each subject (i.e., calculate (x1 – x2) for each subject; test whether the mean difference is zero or not). The test statistic is given by , with *n* – 1 degrees of freedom, where , are the sample means for groups 1 and 2, respectively, sd is the standard deviation of the differences, *n* is the number of *subjects* (so there are 2*n* observations).

1. **One-way ANOVA**

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

An ANOVA uses an F-test which has two parameters – the numerator degrees of freedom and the denominator degrees of freedom. The degrees of freedom in the numerator are equal to *g* – 1, where *g* is the number of groups. If *n* is the total number of observations (*n*1 + *n*2), then, the denominator degrees of freedom is equal to *n* – *g*. The F-statistic is the ratio of the between-groups variance to the within-groups variance, hence the higher the F-statistic the more evidence we have that the means are different.

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

Welch’s two-sample *t*-test

By *t* , and the degrees of freedom is given by

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

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

The contrast from the ANOVA,

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

1. **Two-way ANOVA**

ANOVA stands for analysis of variance. For ANOVA, it is assumed that all populations have the same variances. For a two-way ANOVA, three statistical tests are typically performed: the main effect of each factor and the interaction. Suppose we have two factors A and B, where A represent the genotype and B represent the diet in a mouse study. Suppose each of these factors has two levels (A: wild type, knock out; B: standard diet, high fat diet). For this example, there are 4 combinations (“treatments”): A1B1, A1B2, A2B1, and A2B2. The overall ANOVA F-test gives the p-value for testing whether all four of these means are equal or whether at least one pair is different. However, we are also interested in the effect of the genotype and diet. A main effect is a contrast that tests one factor across the levels of the other factor. Hence the A main effect compares (A1B1 + A1B2)/2 vs. (A2B1 + A2B2)/2, and the B-main effect compares (A1B1 + A2B2)/2 vs. (A1B2 + A2B2)/2. The interaction is a contrast that tests whether the mean difference for one factor depends on the level of the other factor, which is (A1B2 + A2B1)/2 vs. (A1B1 + A2B2)/2.

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







1. **Two-way Repeated Measures ANOVA**

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

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

1. **Correlation**

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

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

1. **Hotelling’s T2 test**

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

The Hotelling statistic is: , where *nx* and *n*y are the numbers of samples in each group, is the mean vector of the variables from group 1, is the mean vector of variables from group 2 and **S** is the pooled estimate of the variance-covariance matrix of the variables. This analysis assumes the underlying variance-covariance matrix is the same for each group. Notice that in the case of uncorrelated variables, this is simply a weighted average of the squared mean differences with weights inversely proportional to the sample variances (i.e., the metabolites less variable within a group are given higher weights).

1. **p- values**

For statistical significance testing, p-values are given. The lower the p-value, the more evidence we have that the null hypothesis (typically that two population means are equal) is not true. If “statistical significance” is declared for p-values less than 0.05, then 5% of the time we incorrectly conclude the means are different, when actually they are the same.

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

1. **q-values**

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

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

Table 1: Example of q-value interpretation



1. **Random Forest**

Random forest is a supervised classification technique based on an ensemble of decision trees (see Breiman L. (2001) Random Forests.  Machine Learning.  45: 5-32; http://link.springer.com/article/10.1023%2FA%3A1010933404324).  For a given decision tree, a random subset of the data with identifying true class information is selected to build the tree (“bootstrap sample” or “training set”), and then the remaining data, the “out-of-bag” (OOB) variables, are passed down the tree to obtain a class prediction for each sample.  This process is repeated thousands of times to produce the forest.  The final classification of each sample is determined by computing the class prediction frequency (“votes”) for the OOB variables over the whole forest.  For example, suppose the random forest consists of 50,000 trees and that 25,000 trees had a prediction for sample 1.  Of these 25,000, suppose 15,000 trees classified the sample as belonging to Group A and the remaining 10,000 classified it as belonging to Group B.  Then the votes are 0.6 for Group A and 0.4 for Group B, and hence the final classification is Group A.  This method is unbiased since the prediction for each sample is based on trees built from a subset of samples that do not include that sample.  When the full forest is grown, the class predictions are compared to the true classes, generating the “OOB error rate” as a measure of prediction accuracy.  Thus, the prediction accuracy is an unbiased estimate of how well one can predict sample class in a new data set.  Random forest has several advantages – it makes no parametric assumptions, variable selection is not needed, it does not overfit, it is invariant to transformation, and it is fairly easy to implement with R.

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

1. **Hierarchical Clustering**

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

1. **Principal Components Analysis (PCA)**

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

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

1. **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 = \frac{x - \mu}{\sigma}

Z-score =

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.